B. J. K. Kleijn

The frequentist theory of Bayesian statistics

May 14, 2019

Springer
Preface

This book grew out of a set of lecture notes that were written for the course in Bayesian statistics, taught at University of Amsterdam in the spring of 2007 and in years thereafter. The course was aimed (initially) at first-year MSc.-students in statistics, mathematics and related fields. The goal was for students to understand the basic properties of Bayesian statistical methods; to be able to apply this knowledge to statistical questions and to know the extent (and limitations) of conclusions based thereon. All material was presented side-by-side with relevant frequentist methods for comparison. Considered were the basic properties of the Bayesian procedure, Bayesian forms of estimation, testing and uncertainty quantification, as well as the choice of the prior by objective and subjective criteria. Presently this material is covered in part I of this book. In addition part I considers the frequentist asymptotic theory of estimation of smooth parameters and efficiency, followed by the Bernstein-von-Mises theorem and the asymptotic identification of credible sets with confidence sets in that context.

It should be stressed that the material presented in part I covers only the most basic Bayesian concepts; further reading is recommended. Various books providing overviews of Bayesian statistics are recommended, depending on the background and interest of the reader: a highly theoretical treatment can be found in Le Cam (1986) [162], which develops a general, mathematical framework for statistics and decision theory, dealing with Bayesian methods as an important area of its application. For a more down-to-earth version of this work, applied only to smooth parametric models, the interested reader is referred to Le Cam and Yang (1990) [7]. A general reference of a more decision-theoretic inclination with a clear focus on Bayesian statistics, is the book by Berger (1985) [19]; a reference of a similar nature is Bernardo and Smith (1993) [25]. Recommendable is also Robert’s “The Bayesian choice” (2001) [194], which offers a very useful explanation on computational aspects of Bayesian statistics. Berger, Bernardo and Smith and Robert devote a great deal of attention to philosophical arguments in favour of the Bayesian approach to statistics, while staying rather brief with regard to mathematical considerations and focusing almost exclusively on parametric models. Finally, Ripley (1996) [195] discusses Bayesian methods with a very pragmatic focus on pattern classification. The
latter reference relates all material with applications in mind but does so based on a
firm statistical and decision-theoretic background.

Part II concerns non-parametric statistics and is built on asymptotic arguments
almost exclusively. On this subject, an early publication in book-form is Ghosh
and Ramamoorthi [103] who give both the theory and examples as they existed
at the time in the field of non-parametric Bayesian statistics. Recently Ghosal and
van der Vaart [102] have published an impressive volume, that provides an overview
of the non-parametric models and methods that have appeared in the intervening
years. The present work has the rather more modest goal of formulating the modern,
asymptotic theory of the posterior for the frequentist with mathematical precision
and in full generality.

An attempt has been made to make part I of this book as self-contained as pos-
sible: in particular, appendix B summarizes the necessary basics of measure the-
ory. Nevertheless the reader is expected to have been exposed to some frequentist
statistics, preferably not just in practical setting but also from a more mathematical
perspective. Throughout part I no familiarity with asymptotic and non-parametric
statistics is assumed on the part of the reader. In part II some such familiarity is
assumed, but the reader is not assumed to be at home with the details of locally
convex spaces, which are summarized in appendix ???. Where possible, definitions,
lemmas and theorems have been formulated such that they cover parametric and
nonparametric models alike.

For corrections to early versions of these notes the author thanks Chris Muris.
For the solutions to many exercises, the author thanks Audrius Jukonis and Mike
Derksen.

B. Kleijn
Amsterdam, February 2018
Contents

Part I  Parametric Bayesian statistics

1  Introduction ................................................................. 3
   1.1  Frequentist statistics ............................................. 3
   1.2  Frequentist estimation ........................................... 8
   1.3  Bayesian statistics ................................................. 11
   1.4  The frequentist analysis of Bayesian methods .............. 14

2  Bayesian basics .......................................................... 17
   2.1  Bayes’s rule, prior and posterior distributions ............. 17
       2.1.1  Bayes’s rule ................................................. 18
       2.1.2  Bayes’s billiard ............................................ 22
       2.1.3  The Bayesian view of the model ......................... 24
       2.1.4  A frequentist’s view of the posterior .................. 27
       2.1.5  From prior to posterior ................................... 29
   2.2  Bayesian point estimators ....................................... 30
       2.2.1  Posterior predictive and posterior mean ............... 30
       2.2.2  Small-ball and formal Bayes estimators ............... 34
       2.2.3  The maximum-a-posteriori estimator ................... 36
   2.3  Confidence sets and credible sets ............................. 38
       2.3.1  Frequentist confidence intervals ....................... 39
       2.3.2  Bayesian credible sets .................................... 41
   2.4  Tests and Bayes factors .......................................... 43
       2.4.1  Neyman-Pearson tests ..................................... 43
       2.4.2  Randomized tests and the Neyman-Pearson lemma ....... 46
       2.4.3  Symmetric and asymptotic testing ....................... 48
       2.4.4  Posterior odds and Bayes factors ....................... 51
   2.5  Decision theory and classification ........................... 54
       2.5.1  Frequentist decision theory ............................. 56
       2.5.2  Bayesian decision theory ................................. 59
       2.5.3  Frequentist versus Bayesian classification ............ 61
## 3 Choice of the prior

- **3.1 Subjective priors**
  - 3.1.1 Motivation for the subjectivist approach
  - 3.1.2 Methods for the construction of subjective priors

- **3.2 Non-informative priors**
  - 3.2.1 Uniform priors
  - 3.2.2 Jeffreys prior and reference priors

- **3.3 Hierarchical priors**
  - 3.3.1 Hyperparameters and hyperpriors
  - 3.3.2 Hierarchical prior construction in an example

- **3.4 Empirical prior choice**
  - 3.4.1 Model selection with empirical methods
  - 3.4.2 Bias and the James-Stein phenomenon

- **3.5 Conjugate families**
  - 3.5.1 Basic definitions and some examples
  - 3.5.2 Exponential families

- **3.6 Dirichlet priors**
  - 3.6.1 Dirichlet distributions

- **3.7 Exercises**

## 4 The Bernstein-von Mises theorem

- **4.1 Efficiency in smooth parametric models**
  - 4.1.1 Optimality in smooth, parametric estimation problems
  - 4.1.2 Regularity and efficiency

- **4.2 Le Cam’s Bernstein-von Mises theorem**
  - 4.2.1 Proof of the Bernstein-von Mises theorem
  - 4.2.2 Three subsidiary lemmas

- **4.3 Model selection with the BIC criterion**

- **4.4 Semi-parametric Bernstein-von Mises theorems**

- **4.5 Exercises**

## 5 Model misspecification

- **5.1 Misspecification in smooth parametric models**
  - 5.1.1 Misspecified maximum likelihood estimation
  - 5.1.2 The misspecified Bernstein-von Mises theorem

- **5.2 Posterior limit distribution**
  - 5.2.1 Asymptotic normality in LAN models
  - 5.2.2 Asymptotic normality in the i.i.d. case
  - 5.2.3 Asymptotic normality of point-estimators

- **5.3 Rate of convergence**
  - 5.3.1 Posterior rate of convergence
  - 5.3.2 Suitable test sequences

- **5.4 Exercises**
Part II Nonparametric Bayesian statistics

6 Non-parametric Bayesian statistics ........................................ 145
   6.1 Asymptotic statistics ................................................. 145
      6.1.1 Consistency, rate and limit distribution .................. 146
   6.2 Posterior concentration and model topology ...................... 147
      6.2.1 Posterior concentration .................................... 148
      6.2.2 Consistency of Bayesian point-estimators ................. 149
   6.3 Bayesian consistency and Doob's theorem ......................... 150
   6.4 Dirichlet priors and posteriors .................................. 152
      6.4.1 Existence of the Dirichlet process ....................... 152
      6.4.2 Support of Dirichlet priors ............................... 154
      6.4.3 Consistency of Dirichlet posteriors ...................... 156
   6.5 Schwartz's frequentist posterior consistency theorem .......... 157
      6.5.1 Proof of Schwartz's theorem ............................. 158
   6.6 Posterior convergence at a rate .................................. 160
      6.6.1 The Ghosal-Ghosh-van derVaart theorem ................... 162
      6.6.2 Proof of the Ghosal-Ghosh-van der Vaart theorem ....... 163
      6.6.3 Entropy numbers and the existence of test sequences .... 165
      6.6.4 Lower bounds on prior mass ............................... 168
   6.7 Frequentist counterexamples ..................................... 169
      6.7.1 Freedman's counterexamples ............................... 169
      6.7.2 Counterexamples: Schwartz and GGV conditions .......... 171
   6.8 Exercises [EMPTY] ................................................ 172

7 Frequentist validity of Bayesian limits ............................. 173
   7.1 Posterior concentration and asymptotic tests .................. 174
      7.1.1 Bayesian test sequences ................................... 174
      7.1.2 Existence of Bayesian test sequences ..................... 176
      7.1.3 Le Cam's inequality ....................................... 179
   7.2 Remote contiguity ................................................ 179
      7.2.1 Definition and criteria for remote contiguity ........... 179
   7.3 Remote contiguity for Bayesian limits ........................... 182
      7.3.1 Remote contiguity, examples in regression ............... 184
   7.4 Posterior concentration .......................................... 188
      7.4.1 Consistency of Bayesian point estimators ............... 191
      7.4.2 Posterior concentration and Hellinger entropy .......... 192
      7.4.3 Finite sample spaces and the tailfree case .............. 193
   7.5 Rates of posterior concentration ................................ 197
      7.5.1 Remote contiguity and the LAN condition ............... 199
   7.6 Consistent hypothesis testing with Bayes factors .............. 200
      7.6.1 Frequentist model selection with posteriors ............ 201
      7.6.2 Goodness-of-fit Bayes factors for random walks .......... 202
   7.7 Confidence sets from credible sets ............................ 206
      7.7.1 Credible/confidence sets in metric spaces .............. 210
7.8 Conclusions .................................................. 211
7.9 Exercises [EMPTY] ........................................ 212

8 Consistent tests and model selection .......................... 213
  8.1 Asymptotic testability ..................................... 213
    8.1.1 Some examples and unexpected answers .......... 214
    8.1.2 Testability over the decades ....................... 215
    8.1.3 The forms that answers take ....................... 217
  8.2 Existence of test sequences ............................. 218
    8.2.1 The Le Cam-Schwartz theorem ...................... 219
  8.3 Uniform testability ...................................... 221
  8.4 Pointwise testability .................................... 223
    8.4.1 Pointwise testability in non-dominated models ... 224
    8.4.2 Pointwise non-testability ......................... 226
    8.4.3 Pointwise testability in dominated models ...... 229
  8.5 Bayesian test sequences .................................. 235
  8.6 Bayesian testing power and model selection for frequentists ... 238
  8.7 Conclusions and discussion ................................ 240
    8.7.1 Model assumptions .................................. 241
    8.7.2 Model selection ..................................... 241
  8.8 Exercises [EMPTY] ......................................... 241

9 Application: non-parametric errors-in-variables regression ... 243
  9.1 Errors-in-variables regression ............................ 243
    9.1.1 Definition of the EIV model ....................... 245
    9.1.2 Posterior concentration theorem .................... 246
  9.2 Rates of posterior convergence in function spaces .... 246
    9.2.1 Lipschitz and smoothness classes .................. 246
    9.2.2 Competing entropy bounds .......................... 247
    9.2.3 Competing lower bounds on prior mass ............ 249
    9.2.4 Various rates of posterior convergence .......... 251
  9.3 Model entropy ............................................ 252
    9.3.1 Nets in parametrizing spaces ...................... 252
    9.3.2 Metric entropy of the errors-in-variables model ... 255
    9.3.3 Proofs of several lemmas .......................... 257
  9.4 Model prior ............................................... 261
    9.4.1 Lemmas ............................................... 265
  9.5 Regression classes ....................................... 267
    9.5.1 Covering numbers of regression classes .......... 268
    9.5.2 Priors on regression classes ...................... 271
  9.6 Exercises [EMPTY] ......................................... 274
10 Application: community detection in the planted bi-section model  
10.1 Communities in random graphs  
10.2 The planted bi-section model  
10.3 Existence of suitable tests  
10.4 Exact recovery and detection with posteriors  
10.4.1 Posterior consistency: exact recovery  
10.4.2 Posterior consistency: detection  
10.5 Uncertainty quantification  
10.5.1 Posterior recovery/detection and confidence sets  
10.5.2 Confidence sets directly from credible sets  
10.6 Exercises [EMPTY]  

A Notation, definitions and conventions  

B Measure theory  
B.1 Sets and sigma-algebras  
B.2 Measures  
B.3 Measurability, random variables and integration  
B.4 Existence of stochastic processes  
B.5 Conditional distributions  

C Topologies on spaces of measures  
C.1 Convergence in spaces of probability measures  
C.2 Inverse limits and locally convex spaces  
C.3 Contiguity  

D Inverse limit measures  
D.1 The Bochner-Kolmogorov existence theorem  
D.2 Topological characterization of existence  
D.3 Inverse limits of positive measures  
D.4 Partitions and inverse limit samplespaces  
D.5 Open and clopen topological bases  
D.6 Inverse limit priors  

References  
References  

Index
Part I

Parametric Bayesian statistics
Chapter 1
Introduction

The goal of statistical inference is to understand, describe and estimate (aspects of) the randomness of measured data. Quite naturally this invites the assumption that the data represents a sample from an unknown but fixed probability distribution.

1.1 Frequentist statistics

Any frequentist inferential procedure relies on three basic ingredients: the data, a model and an estimation procedure. The central assumption in frequentism is that the data has a definite but unknown, underlying distribution to which all inference pertains. The data is a measurement or observation which we denote by $Y$, taking values in a corresponding samplespace.

**Definition 1.1.** The samplespace for an observation $Y$ is a measurable space $\mathcal{Y}$ with $\sigma$-algebra $\mathcal{B}$ (see definition B.2) containing all values that $Y$ can take upon measurement.

Measurements and data can take any form, ranging from categorical data (sometimes referred to as nominal data where the samplespace is simply a (usually finite) set of points or labels with no further mathematical structure), ordinal data (also known as ranked data, where the samplespace is endowed with a total ordering), to interval data (where in addition to having an ordering, the samplespace allows one to compare differences or distances between points), to ratio data (where we have all the structure of the real line). Moreover $Y$ can collect the results of a number of measurements, so that it takes its values in the form of a vector (think of an experiment involving repeated, stochastically independent measurements of the same quantity, leading to a so-called independent and identically distributed (or i.i.d.) sample). The data $Y$ may even be functional data, taking its values in a space of functions or in other infinite-dimensional spaces, for example, in the statistical study of continuous-time time-series.
The samplespace $\mathcal{Y}$ is assumed to be a measurable space to enable the consideration of probability measures on $\mathcal{Y}$, formalizing the uncertainty in measurement of $Y$. As was said in the opening words of this chapter, frequentist statistics hinges on the assumption that there exists a probability measure $P_0 : \mathcal{B} \to [0,1]$ on the samplespace $\mathcal{Y}$ representing the "true distribution of the data":

$$Y \sim P_0$$

(1.1)

Hence from the frequentist perspective, statistics revolves around the central question: "What does the data make clear about $P_0$?", which may be considered in parts by questions like, "From the data, what can we say about the mean of $P_0$?", "Based on the data that we have, how sharp can we formulate hypotheses concerning the value of the variance of $P_0$?", etcetera.

The second ingredient of a statistical procedure is a model, which contains all explanations under consideration of the randomness in $Y$.

**Definition 1.2.** A (frequentist) statistical **model** $\mathcal{P}$ is a collection of probability measures $P : \mathcal{B} \to [0,1]$ on the samplespace $(\mathcal{Y}, \mathcal{B})$. The distributions $P$ are called **model distributions**. For every sample space $(\mathcal{Y}, \mathcal{B})$, the collection of all probability distributions is called the full model (sometimes referred to as the full non-parametric model), denoted $\mathcal{M}_{+1}(\mathcal{Y}, \mathcal{B})$.

The model $\mathcal{P}$ contains the candidate distributions for $Y$ that the statistician finds "reasonable" explanations of the uncertainty he observes (or expects to observe) in $Y$. As such, it constitutes a choice of the statistician analyzing the data rather than a given. From a more mathematical perspective we observe that a model $\mathcal{P}$ on $(\mathcal{Y}, \mathcal{B})$ is a subset of the space $\mathcal{M}(\mathcal{Y}, \mathcal{B})$ of all finite, signed measures $\mu : \mathcal{B} \to \mathbb{R}$ (that is, all countably additive, real-valued set functions) that are of finite total variation. Equipped with the total-variational norm (see appendix B, definition B.6), $\mu \mapsto \|\mu\|$, $\mathcal{M}(\mathcal{Y}, \mathcal{B})$ is a Banach space [77], in which the full model can be characterized by,

$$\mathcal{M}_{+1}(\mathcal{Y}, \mathcal{B}) = \{ P \in \mathcal{M}(\mathcal{Y}, \mathcal{B}) : P \geq 0, P(\mathcal{Y}) = 1 \}.$$

Often, we describe models as families of probability densities rather than distributions.

**Definition 1.3.** If there exists a $\sigma$-finite measure $\mu : \mathcal{B} \to [0,\infty]$ such that for all $P \in \mathcal{P}$, $P \ll \mu$, we say that the model is **dominated** (notation: $\mathcal{P} \ll \mu$).

The Radon-Nikodym theorem (see theorem B.7) guarantees that we may represent a dominated probability measure $P$ in terms of a probability density function $p = dP/d\mu : \mathcal{Y} \to [0,\infty)$ that satisfies $\int_A p(y) d\mu(y) = P(A)$ for all $A \in \mathcal{B}$. For dominated models, it makes sense to adopt a slightly different mathematical perspective: if $\mu$ dominates $\mathcal{P}$, we map $\mathcal{P}$ to the space of all $\mu$-integrable functions $L_1(\mu)$ by means of the Radon-Nikodym mapping.

**Example 1.1.** Suppose that $\mathcal{Y}$ is countable (and let $\mathcal{B}$ be the powerset of $\mathcal{Y}$): then the measure $\mu$ that puts mass one at every point in $\mathcal{Y}$, also known as the **counting
measure on \( \mathcal{Y} \), is \( \sigma \)-finite and dominates every other (finite) measure on \( \mathcal{Y} \). Consequently, any model on \( (\mathcal{Y}, \mathcal{B}) \) can be represented in terms of elements \( p \) in the Banach space \( L_1(\mu) \), more commonly denoted as \( \ell_1 \),

\[
\ell_1 = \{ (f_1, f_2, \ldots) \in [0,1]^\infty : \sum_{i \geq 1} |f_i| < \infty \}.
\]

where it is noted that \( p_i \geq 0 \) and \( \|p\| = \sum_i p_i = 1 \) for all \( P \) in the set \( \Lambda \) of all probability measures on \( (\mathcal{Y}, \mathcal{B}) \).

In case the sample space is not discrete, the full model is not dominated by a \( \sigma \)-finite measure. However, suppose that a \( \sigma \)-finite measure \( \mu \) on the sample space is given. The Radon-Nikodym mapping maps every \( \mu \)-dominated model \( P \) to a subset of,

\[
\mathcal{M}^+_1(\mu) = \{ p \in L_1(\mu) : p \geq 0, \int_{\mathcal{Y}} p(y) d\mu(y) = 1 \}.
\]

**Lemma 1.1.** The mapping between a model \( \mathcal{P} \) dominated by a \( \sigma \)-finite measure \( \mu \) and its \( L_1(\mu) \)-representation is an isometry: for all \( p_1, p_2 \in \mathcal{P} \),

\[
\|P_1 - P_2\| = \frac{1}{2} \int_{\mathcal{Y}} |p_1(y) - p_2(y)| d\mu(y) = \int_{\mathcal{Y}} (p_1(y) - p_2(y))^+ d\mu(y).
\]

Note that a dominating measure is not unique, so there are many \( L_1 \)-representations of \( \mathcal{P} \). The most common way of representing a statistical model is a description in terms of a parameterization.

**Definition 1.4.** A model \( \mathcal{P} \) is parameterized with parameter space \( \Theta \), if there exists a surjective map \( \Theta \rightarrow \mathcal{P} : \theta \mapsto P_\theta \), called the parameterization of \( \mathcal{P} \).

Parametrizations are motivated by the context of the statistical question and the parameter \( \theta \) usually has a clear interpretation when viewed in this context. The formulation of parametric model descriptions is the modelling step of statistics: to the statistician, it transforms the data from a mere list of numbers to an informative (but noisy) representation of an underlying truth.

**Definition 1.5.** A parameterization of a statistical model \( \mathcal{P} \) is said to be identifiable, if the map \( \Theta \rightarrow \mathcal{P} : \theta \mapsto P_\theta \) is injective.

Injectivity of the parameterization means that for all \( \theta_1, \theta_2 \in \Theta, \theta_1 \neq \theta_2 \) implies that \( P_{\theta_1} \neq P_{\theta_2} \). In other words, no two different parameter values \( \theta_1 \) and \( \theta_2 \) give rise to the same distribution. Clearly, in order for \( \theta \in \Theta \) to serve as a useful representation for the candidate distributions \( P_{\theta} \), identifiability is a first requirement. Other common conditions on the map \( \theta \mapsto P_\theta \) are continuity (for example, with respect to the total-variational norm, or through requiring continuity of all maps \( \theta \mapsto P_\theta g \), with \( g \) any bounded measurable function), differentiability (the definition of which may involve technical subtleties in case \( \Theta \) is infinite-dimensional) and other smoothness conditions.
Remark 1.1. Although strictly speaking ambivalent, it is commonplace to refer to both $\mathcal{P}$ and the parameterizing space $\Theta$ as “the model”. This practice is not unreasonable in view of the fact that, in practice, almost all models are parameterized in an identifiable way, so that there exists a bijective correspondence between $\Theta$ and $\mathcal{P}$. Here, reference to the model always concerns $\mathcal{P}$ while $\Theta$ is always called the parameter space.

An assumption often made in frequentist statistics is that the true distribution of the data is a model distribution.

**Definition 1.6.** A model $\mathcal{P}$ is said to be well-specified if it contains the true distribution of the data $P_0$, i.e.

$$P_0 \in \mathcal{P}. \quad (1.2)$$

If (1.2) does not hold, the model is said to be mis-specified.

Clearly if $\mathcal{P}$ is parameterized by $\Theta$, (1.2) implies the existence of a point $\theta_0 \in \Theta$ such that $P_{\theta_0} = P_0$; if, in addition, the model is identifiable, the parameter value $\theta_0$ is unique.

If the full non-parametric model is used, (1.2) holds trivially. However, for smaller models, (1.2) has the status of an assumption on the unknown quantity of interest $P_0$ and may as such be hard to justify. The reason for (the somewhat odd and certainly very contentious) assumption (1.2) lies in the interpretation of statistical conclusions: an estimate of a parameter is of value if that parameter can be attributed to the “true” distribution of the data. If, on the other hand, one assumes that the model is mis-specified, parameter estimates may reflect aspects of the true distribution but cannot be associated with the true distribution of the data directly any more.

The model we use in a statistical procedure constitutes a choice rather than a given: presented with a particular statistical problem, different statisticians may choose to use different models. The only condition is that (1.2) is satisfied, which is why we have to choose the model in a “reasonable way” given the nature of $Y$. When choosing the model, two considerations compete: on the one hand, small models are easy to handle mathematically and statistically and parameters usually have clear interpretations, on the other hand, for large models, assumption (1.2) is more realistic since they have a better chance of containing $P_0$ (or at least approximate it more closely). The amount of data available plays a crucial role: if we have a limited sample, simple models have a better chance of leading to sensible results, while an abundance of data enables more sophisticated forms of statistical analysis. In this respect the most important distinction is made in terms of the dimension of the model.

**Definition 1.7.** A model $\mathcal{P}$ is said to be parametric of dimension $d$, if there exists an identifiable parameterization $\Theta \to \mathcal{P} : \theta \mapsto P_\theta$, where $\Theta \subset \mathbb{R}^d$ with non-empty interior $\Theta \neq \emptyset$.

The requirement regarding the interior of $\Theta$ in definition 1.7 ensures that the dimension $d$ really concerns $\Theta$ and not just the dimension of the space $\mathbb{R}^d$ (in which $\Theta$ could otherwise be a lower-dimensional subset).
1.1 Frequentist statistics

Example 1.2. The normal model for a single, real measurement \( Y \), is the collection of all normal distributions on \( \mathbb{R} \), i.e.

\[
\mathcal{P} = \{ N(\mu, \sigma^2) : (\mu, \sigma) \in \Theta \}
\]

where the parameterizing space \( \Theta \) equals \( \mathbb{R} \times (0, \infty) \). The map \((\mu, \sigma) \mapsto N(\mu, \sigma^2)\) is surjective and injective, i.e. the normal model is a two-dimensional, identifiable parametric model. Moreover, the normal model is dominated by the Lebesgue measure on the samplespace \( \mathbb{R} \) and can hence be described in terms of Lebesgue-densities:

\[
p_{\mu, \sigma}(y) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.
\]

Note that for any fixed \( y \in \mathcal{Y} \), the dependence \( \Theta \rightarrow \mathbb{R} : (\mu, \sigma) \mapsto p_{\mu, \sigma}(y) \) is continuous on all of \( \Theta \). So if \((\mu_n, \sigma_n)\) converges to \((\mu, \sigma)\) in \( \Theta \), then \( p_n(y) := p_{\mu_n, \sigma_n}(y) \) converges to \( p(y) := p_{\mu, \sigma}(y) \). Then total-variational distance between the distributions \( P_n \) and \( P \) (associated with the densities \( p_n \) and \( p \) respectively) satisfies,

\[
\|P_n - P\| = \frac{1}{2} \int_{\mathcal{Y}} |p_n(y) - p(y)| \, d\mu(y) \to 0.
\]

by lemma 1.1 and Scheffé’s lemma (see corollary C.1). Conclude that the parametrization \( \Theta \rightarrow \mathcal{P} : (\mu, \sigma) \mapsto P_{\mu, \sigma} \) is continuous with respect to the total-variational metric on \( \mathcal{P} \).

Definition 1.8. If there is no finite-dimensional \( \Theta \) that parameterizes \( \mathcal{P} \), then \( \mathcal{P} \) is called a non-parametric model.

For instance, the full model \( \mathcal{M}_1^+ (\mathcal{Y}, \mathcal{B}) \) is non-parametric unless the samplespace contains only a finite number of points.

Example 1.3. Let \( \mathcal{Y} \) be a finite set containing \( n \geq 1 \) points \( y_1, y_2, \ldots, y_n \) and let \( \mathcal{B} \) be the power-set \( 2^\mathcal{Y} \) of \( \mathcal{Y} \). Any probability measure \( P : \mathcal{B} \rightarrow [0, 1] \) on \((\mathcal{Y}, \mathcal{B})\) is absolutely continuous with respect to the counting measure on \( \mathcal{Y} \) (see example B.2). The density of \( P \) with respect to the counting measure is a map \( p : \mathcal{Y} \rightarrow \mathbb{R} \) such that \( p \geq 0 \) and

\[
\sum_{i=1}^{n} p(y_i) = 1.
\]

As such, \( P \) can be identified with an element of the so-called simplex \( S_n \) in \( \mathbb{R}^n \), defined as follows

\[
S_n = \left\{ p = (p_1, \ldots, p_n) \in \mathbb{R}^n : p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right\}.
\]

This leads to an identifiable parameterization \( S_n \rightarrow \mathcal{P} : p \mapsto P \) of the full model on \((\mathcal{Y}, \mathcal{B})\), of dimension \( n - 1 \). Note that \( S_n \) has empty interior in \( \mathbb{R}^n \), but can be brought in one-to-one correspondence with a compact set in \( \mathbb{R}^{n-1} \) with non-empty interior by the embedding:
\[
\left\{ (p_1, \ldots, p_{n-1}) \in \mathbb{R}^{n-1} : p_i \geq 0, \sum_{i=1}^{n-1} p_i \leq 1 \right\} \rightarrow S_n : (p_1, \ldots, p_{n-1}) \mapsto (p_1, \ldots, p_{n-1}, 1 - \sum_{i=1}^{n-1} p_i).
\]

1.2 Frequentist estimation

The third ingredient of a frequentist inferential procedure is an estimation method. Clearly not all statistical problems involve an explicit estimation step and of those that do, not all estimate the distribution \( P_0 \) directly. Nevertheless, one may regard the problem of point-estimation in the model \( \mathcal{P} \) as prototypical.

**Definition 1.9.** A point-estimator (or estimator) for \( P_0 \) is a map \( \hat{P} : \mathcal{Y} \rightarrow \mathcal{P} \), representing our “best guess” \( \hat{P} \in \mathcal{P} \) for \( P_0 \) based on the data \( Y \) (and other known quantities).

Note that a point-estimator is a statistic: since a point-estimator must be calculable in practice, it may depend only on information that is known to the statistician after he has performed the measurement realized as \( Y = y \). Also note that a point-estimator is a stochastic quantity: \( \hat{P} = \hat{P}(Y) \) depends on \( Y \) and is hence random. Upon measurement of \( Y \) resulting in a realisation \( Y = y \), the realisation of the estimator is an estimate \( \hat{P}(y) \), a definite point in \( \mathcal{P} \). If the model is parameterized, one may define a point-estimator \( \hat{\theta} : \mathcal{Y} \rightarrow \Theta \) for \( \theta_0 \), from which we obtain \( \hat{P} = \hat{P}(\hat{\theta}) \) as an estimator for \( P_0 \). In that case the continuity requirement we impose on the map \( \theta \mapsto P_\theta \) guarantees that \( \theta \mapsto \theta_0 \) implies \( P_\theta \rightarrow P_{\theta_0} \). If the model is identifiable, estimation of \( \theta_0 \) in \( \Theta \) is equivalent to estimation of \( P_0 \) in \( \mathcal{P} \).

Aside from estimates for the distribution \( P_0 \), one is often interested in estimating only certain aspects of \( P_0 \).

**Example 1.4.** Suppose that a bank tries to assess market risk for an asset: they have the asset on the books for price \( x \) but tomorrow’s market will say that it is worth a price \( X \), distributed according to an unknown \( P_0 \). To assess the risk of holding the position until tomorrow, the absolute return \( X - x \) is of importance. Of course, the bank would prefer to have a reliable estimate for \( P_0 \) (and thus for the distribution of \( X - x \)) but that question is often too hard to answer and reliability cannot be guaranteed. Instead, the bank will resort to a simplification by focussing on the aspect of the distribution \( P_0 \) that they find most important for their risk assessment. A popular notion in this context is a quantity called value-at-risk: given a time-horizon (in this case, tomorrow) and a significance level \( \alpha \in (0, 1) \) (often chosen equal to 0.05 or 0.01), value-at-risk \( q \) is defined as the maximal \( q < 0 \) at which,

\[
P_0(X - x < q) \leq \alpha.
\]

To interpret \( q \), note that losses exceeding value-at-risk occur on only an expected fraction \( \alpha \) of all trading days. In statistical terms, \( q \) is a quantile of \( P_0 \).
1.2 Frequentist estimation

Another example occurs in parametric models: if the dimension $d$ of a parametric model is greater than one, we may choose to estimate only one component of $\theta$ (called the \textit{parameter of interest}) and disregard other components (called \textit{nuisance parameters}). More generally, we may choose to estimate certain properties of $P_0$ (e.g. its expectation, variance) rather than $P_0$ itself and in many cases, direct estimation of the property of interest of $P_0$ is more efficient than estimation through $\hat{P}$.

\textbf{Example 1.5.} Consider a model $\mathcal{P}$ consisting of distributions on $\mathbb{R}$ with finite expectation and define the functional $e : \mathcal{P} \to \mathbb{R}$ by the expectation $e(P) = \mathbb{E}[X]$. Suppose that we are interested in the expectation $e_0 = e(P_0)$ of the true distribution. Obviously, based on an estimator $\hat{P}$ for $P_0$ we may define an estimator, $\hat{e} = \int_{\mathbb{R}} x \, d\hat{P}(x)$ (1.3) to estimate $e_0$. For instance, assume that $X$ is integrable under $P_0$ and $Y = (X_1, \ldots, X_n)$ collects the results of an \textit{i.i.d.} experiment with $X_i \sim P_0$ marginally (for all $1 \leq i \leq n$), then the \textit{empirical expectation} of $X$, defined simply as the \textit{sample-average} of $X$,

$$P_nX = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

provides an estimator for $e_0$. (Note that the sample-average is also of the form (1.3) if we choose as our point-estimator for $P_0$ the empirical distribution $\hat{P} = \mathbb{P}_n$.) The law of large numbers guarantees that $P_nX$ converges to $e_0$ almost-surely as $n \to \infty$, and (if $X$ is quadratically integrable) the \textit{central limit theorem} asserts that this convergence proceeds at rate $n^{-1/2}$ (and that the limit distribution is zero-mean normal with $P_0(X - \mathbb{E}[X])^2$ as its variance). Many parameterizations $\theta \mapsto P_\theta$ are such that (components of) $\theta$ coincide with expectations. Often, other properties of $P_0$ can also be related to expectations: for example, if $X \in \mathbb{R}$, the probabilities $F_0(s) = P_0(X \leq s) = P_0\{X \leq s\}$ ($s \in \mathbb{R}$) can be estimated by,

$$F_n(s) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq s\},$$

\textit{i.e.} as the empirical expectation of the function $x \mapsto 1\{x \leq s\}$. This leads to a step-function with $n$ jumps of size $1/n$ at samplepoints, which estimates the distribution function $F_0$.

Generalizing, any property of $P_0$ that can be expressed in terms of an expectation of a $P_0$-integrable function of $X$, $P_0g$, is estimable by the corresponding empirical expectation, $\mathbb{P}_n g$. (With regard to the estimator $F_n$, the convergence $F_n(s) \to F_0(s)$ does not only hold pointwise but even uniform in $s$, \textit{i.e.} $\sup_{s \in \mathbb{R}} |F_n(s) - F_0(s)| \to 0$ almost-surely, \textit{c.f.} the Glivenko-Cantelli theorem.)

To estimate a probability distribution (or any of its properties or parameters), many different estimators may exist. Therefore, the use of any particular estimator constitutes (another) \textit{choice} made by the statistician analyzing the problem. Whether
such a choice is a good or a bad one depends on optimality criteria, which are either dictated by the particular nature of the problem (see section 2.5 which extends the purely inferential point of view), or based on more generically desirable properties of the estimator (note the use of the rather ambiguous qualification “best guess” in definition 1.9).

Example 1.6. To illustrate what we mean by “desirable properties”, note the following. When estimating $P_0$ one may decide to use an estimator $\hat{P}$ because it has the property that it is close to the true distribution of $Y$ in total variation: there exist small constants $\varepsilon > 0$ and $\alpha > 0$ such that for all $P \in \mathcal{P}$,

$$ P(\|\hat{P}(Y) - P\| < \varepsilon) > 1 - \alpha, $$

i.e. if $Y \sim P$, then $\hat{P}$ lies close to $P$ with high $P$-probability. Note that we formulate this property “for all $P$ in the model”: since $P_0 \in \mathcal{P}$ is unknown, the only way to guarantee that this property holds under $P_0$, is to prove that it holds for all $P \in \mathcal{P}$ (provided that (1.2) holds). By contrast, for Bayesians any claim concerning points $P$ in the model is acceptable if it is true almost-everywhere in $\mathcal{P}$ with respect to the prior measure.

A popular method of estimation that satisfies common optimality criteria in many (but certainly not all, see [166]) problems is maximum-likelihood estimation.

Definition 1.10. Suppose that the model $\mathcal{P}$ is dominated by a $\sigma$-finite measure $\mu$ and parametrized through $\mu$-densities by $\theta \mapsto p_\theta \in L_1(\mu)$. The likelihood principle (see [194] for an elaborate overview and spirited argument in favour) says that all information implied by data $Y$ concerning the parameter $\theta$ is contained in the likelihood-function $\theta \mapsto p_\theta(Y)$ (note that this defines a random function $\theta \mapsto [0, \infty]$).

Accordingly, one can define $\hat{\theta} \in \Theta$ as an estimator for the true parameter value $\theta_0$ by maximization,

$$ p_\hat{\theta}(Y) = \sup_{\theta \in \Theta} p_\theta(Y). $$

So $\hat{\theta}$ is the point in the parameter space for which the likelihood-function evaluated in $Y$, $\Theta \mapsto [0, \infty] : \theta \mapsto p_\theta(Y)$ attains its maximum. This defines the maximum-likelihood estimator (or MLE) $\hat{\theta}$ for $\theta_0$. (Note that the Bayesian posterior we define later can be written in terms of the likelihood function, thus defining another way to estimate on the basis of the likelihood principle.)

Remark 1.2. The MLE $\hat{P} = P_\theta$ does not depend on the dominating measure $\mu$ chosen to define the densities $p_\theta = dP_\theta/d\mu$.

A word of caution is in order: mathematically, the above “definition” of the MLE begs questions of existence and uniqueness: regarding $\theta \mapsto p_\theta(Y)$ as a (random) map on the parameter space, there may not be any point in $\mathcal{P}$ where the likelihood takes on its supremal value (with $P_\theta$-probability one), nor is there any guarantee that such a maximal point is unique (with $P_\theta$-probability one).
1.3 Bayesian statistics

The above is only a very brief and rather abstract overview of the basic framework of frequentist statistics, highlighting the central premise that a true underlying distribution $P_0$ for $Y$ exists. It makes clear, however, that frequentist inference concerns itself primarily with the stochastics of the random variable $Y$ and not with the context in which $Y$ resides. Other than the fact that the model has to be chosen “reasonably” based on the nature of $Y$, frequentist inference does not involve any information regarding the background of the statistical problem in its procedures unless one chooses to use such information explicitly (see, for example, remark 2.6 on penalized maximum-likelihood estimation). In Bayesian statistics the use of background information is an integral part of the procedure unless one chooses to disregard it: by the definition of a prior measure, the statistician may express that he believes in certain points of the model more strongly than others. This thought is elaborated on further in section 1.3 (e.g. example 1.8).

Similarly, results of estimation procedures are sensitive to the context in which they are used: two statistical experiments may give rise to the same model formally, but the estimator used in one experiment may be totally unfit for use in the other experiment.

Example 1.7. For example, if we interested in a statistic that predicts the rise or fall of a certain share-price on the stock market based on its value over the past week, the estimator we use does not have to be a very conservative one: we are interested primarily in its long-term performance and not in the occasional mistaken prediction. However, if we wish to predict the rise or fall of white-bloodcell counts in an HIV-patient based on last week’s counts, overly optimistic predictions can have disastrous consequences.

Although in the above example, data and models are very similar in these statistical problems, the estimator used in the medical application should be much more conservative than the estimator used in the stock-market problem. The purely statistical aspects of both questions are the same, but the context in which inference is expressed calls for different approaches. Such considerations form the motivation for statistical decision theory, as explained further in section 2.5.

1.3 Bayesian statistics

The subject here is an alternative approach to statistical questions known as Bayesian statistics, after Rev. Thomas Bayes, the author of “An essay towards solving a problem in the doctrine of chances” published posthumously in 1763 [13]. Bayes considered a number of probabilistic questions in which data and parameters are treated on equal footing. The Bayesian procedure itself is explained in detail in chapter 2 and further chapters explore its properties. In this section we have the more modest goal of illustrating the conceptual differences with frequentist statistical analysis.

In Bayesian statistics, data and model form two factors of the same space, i.e. no formal distinction is made between measured quantities $Y$ and parameters $\theta$. One
may envisage the process of generating a measurement outcome $Y = y$ as two draws, one draw from $\Theta$ (or $\mathcal{P}$) to select a value of $\theta$ (or distribution $P$) and a subsequent draw from a model distribution $P_\theta$ to arrive at $Y = y$. This perspective may seem rather strange in view of the definitions made in section 1.1, but in [13], Bayes gives examples in which this perspective is perfectly reasonable (see subsection 2.1.2). An element $P_\theta$ of the model is interpreted simply as the distribution of $Y$ given the parameter value $\theta$, i.e. as the conditional distribution of $Y|\theta$. The joint distribution of $(Y, \theta)$ then follows upon specification of the marginal distribution of $\theta$ on $\Theta$, which is called the prior. Based on the joint distribution for the data $Y$ and the parameters $\theta$, straightforward conditioning on $Y$ gives rise to a conditional distribution $\Pi(\cdot|Y)$ for the parameters $\theta|Y$ called the posterior distribution on the model $\Theta$. Hence, given the model, the data and a prior distribution, the Bayesian procedure leads to a posterior distribution that incorporates the information provided by the data. All statistical questions are then answered using the posterior. For example, what a frequentist would call point-estimation of the underlying distribution with the posterior mean,

$$P^{\Pi|Y}(A) = \int P_\theta(A) d\Pi(\theta|Y),$$

(for all measurable $A$), is called prediction by Bayesians, who refer to $P^{\Pi|Y}$ as the posterior predictive distribution.

Often in applications, the nature of the data and the background of the problem suggest that certain values of $\theta$ are more “likely” than others, even before any measurements are done. The model $\mathcal{P}$ describes possible probabilistic explanations of the data and, in a sense, the statistician believes more strongly in certain explanations than in others. This is illustrated by the following example, which is due to L. Savage (1961) [197].

**Example 1.8.** Consider the following three statistical experiments:

1. A lady who drinks milk in her tea claims to be able to tell which was poured first, the tea or the milk. In ten trials, she determines correctly whether it was tea or milk that entered the cups first.
2. A music expert claims to be able to tell whether a page of music was written by Haydn or by Mozart. In ten trials conducted, he correctly determines the composer every time.
3. A drunken friend says that he can predict the outcome of a fair coin-flip. In ten trials, he is right every time.

Let us analyze these three experiments in a frequentist fashion, e.g. we assume that the trials are independent and possess a definite Bernoulli distribution, c.f. (1.1). In all three experiments, $\theta_0 \in \Theta = [0, 1]$ is the per-trial probability that the person gives the right answer. We test their respective claims posing the hypotheses:

$$H_0 : \theta_0 = \frac{1}{2}, \quad H_1 : \theta_0 > \frac{1}{2}.$$  

The total number of successes out of ten trials is a sufficient statistic for $\theta$ and we use it as our test-statistic, noting that its distribution is binomial with $n = 10$, $\theta = \theta_0$
under $H_0$. Given the data $Y$ with realization $y$ of ten correct answers, applicable in all three examples, we reject $H_0$ at $p$-value $2^{-10} \approx 0.1\%$. So there is strong evidence to support the claims made in all three cases. Note that there is no difference in the frequentist analyses: formally, all three cases are treated exactly the same.

Yet intuitively (and also in every-day practice), one would be inclined to treat the three claims on different footing: in the second experiment, we have no reason to doubt the expert’s claim, whereas in the third case, the friend’s condition makes his claim less than plausible. In the first experiment, the validity of the lady’s claim is hard to guess beforehand. The outcome of the experiments would be as expected in the second case and remarkable in the first. In the third case, one would either consider the friend extremely lucky, or begin to doubt the fairness of the coin being flipped.

The above example convincingly makes the point that in our intuitive approach to statistical issues, we include all knowledge we have, even resorting to strongly biased estimators if the model does not permit a non-biased way to incorporate it. The Bayesian approach to statistics allows us to choose priors that reflect this subjectivity: from the outset, we attach more prior mass to parameter-values that we deem more likely, or that we believe in more strongly. In the above example, we would choose a prior that concentrates more mass at high values of $\theta$ in the second case and at low values in the third case. In the first case, the absence of prior knowledge would lead us to remain objective, attaching equal prior weights to high and low values of $\theta$. Although the frequentist’s testing procedure can be adapted to reflect subjectivity, the Bayesian procedure incorporates it rather more naturally through the choice of a prior.

Subjectivist Bayesians view the above as an advantage; objectivist Bayesians and frequentists view it as a disadvantage. Subjectivist Bayesians argue that personal beliefs are an essential part of statistical reasoning, deserving of an explicit role in the formalism and interpretation of results. Objectivist Bayesians and frequentists reject this thought because scientific reasoning should be devoid of any personal beliefs or interpretation. So the above freedom in the choice of the prior is also the Achilles’s heel of Bayesian statistics: fervent frequentists and objectivist Bayesians take the point of view that the choice of prior is an undesirable source of ambiguity, rather than a welcome way to incorporate “expert knowledge” as in example 1.8. After all, if the subjectivist Bayesian does not like the outcome of his analysis, he can just go back and change the prior to obtain a different outcome. Similarly, if two subjectivist Bayesians analyze the same data they may reach completely different conclusions, depending on the extent to which their respective priors differ.

To a certain extent such ambiguity is also present in frequentist statistics, since frequentists have the freedom to choose biased point-estimators. For example, the use of either a maximum-likelihood or penalized maximum-likelihood estimator leads to differences, the size of which depends on the relative sizes of likelihood and penalty. Indeed, through the maximum-a-posteriori Bayesian point-estimator (see definition 2.12), one can demonstrate that the log-prior-density can be viewed as a penalty term in a penalized maximum-likelihood procedure, c.f. remark 2.6. Yet the natural way in which subjectivity is expressed in the Bayesian setting is more
explicit. Hence the frequentist or objectivist Bayesian sees in this a clear sign that Bayesian statistics lacks universal value unless one imposes that the prior should not express any bias (see section 3.2).

A second difference in philosophy between frequentist and Bayesian statisticians arises as a result of the fact that the Bayesian procedure does not require that we presume the existence of a “true, underlying distribution” \( P_0 \) of \( Y \) (compare with (1.1)). The subjectivist Bayesian views the model with (prior or posterior) distribution as his own, subjective explanation of the uncertainty in the data. For that reason, subjectivists prefer to talk about their (prior or posterior) “belief” concerning parameter values rather than implying objective validity of their assertions. On the one hand, such a point of view makes intrinsic ambiguities surrounding statistical procedures explicit; on the other hand, one may wonder about the relevance of strictly personal belief in a scientific tradition that emphasizes universality of reported results.

The philosophical debate between Bayesians and frequentist has raged with varying intensity for decades, but remains undecided to this date. In practice, the choice for a Bayesian or frequentist estimation procedure is usually not motivated by philosophical considerations, but by far more practical issues, such as ease of computation and implementation, common custom in the relevant field of application, specific expertise of the researcher or other forms of simple convenience. More recent developments [12] suggest that the philosophical debate will be put to rest in favour of more practical considerations as well. In later chapters it is demonstrated how Bayesian and frequentist statistical limits are related in the large-sample asymptotic regime.

### 1.4 The frequentist analysis of Bayesian methods

Since this point has the potential to cause great confusion, we emphasize the following: this text presents Bayesian statistics from a hybrid perspective, \( i.e. \) we consider Bayesian techniques but analyze them in frequentist setting and with frequentist methods.

We take the frequentist point of view with regard to the data, \( e.g. \) assumption (1.1); we distinguish between samplespace and model and we do not adhere to subjectivist interpretations of results (although their perspective is discussed in the main text). On the other hand, we endow the model with a prior probability measure and calculate the posterior distribution, \( i.e. \) we use concepts and definitions from Bayesian statistics. This enables us to assess Bayesian methods on equal footing with frequentist statistical methods and extends the range of interesting questions. Moreover, it resolves the inherent ambiguity haunting the subjectivist interpretation of statistical results. Note, however, that the derivation of expression (2.12) for the posterior, for example, is the result of subjectivist Bayesian assumptions on data and model. Since these assumptions are at odds with the frequentist perspective, we shall take (2.12) as a \textit{definition} rather than a derived form (see subsection 2.1.4).
1.4 The frequentist analysis of Bayesian methods

Much of the material covered in this book does not depend on any particular philosophical point of view, especially when the subject matter is purely mathematical. Nevertheless, it is important to realize when philosophical issues may come into play and there will be points where this is the case. In particular when discussing asymptotic properties of Bayesian procedures, adoption of assumption (1.1) is instrumental, basically because discussing convergence requires a limit-point.

Exercises

1.1. Let $Y \in \mathcal{Y}$ be a random variable with unknown distribution $P_0$. Let $\mathcal{P}$ be a model for $Y$, dominated by a $\sigma$-finite measure $\mu$ and parametrized by $\Theta \rightarrow P : \theta \mapsto P_\theta$. Assume that the maximum-likelihood estimator $\hat{\theta}$ (see definition 1.10) is well-defined, $P_0$-almost-surely. Show that if $\nu$ is a $\sigma$-finite measure dominating $\mu$ and we calculate the likelihood using $\nu$-densities, then the associated MLE is equal to $\hat{\theta}$. Conclude that the MLE does not depend on the dominating measure used, c.f. remark 1.2.

1.2. Prove lemma 1.1.

1.3. Let $\mathcal{Y} = \mathbb{R}$ with $\sigma$-algebra $\mathcal{B}$. Show that if $\mathcal{B}$ is the usual Borel $\sigma$-algebra, then $\mathcal{M}_1^+(\mathcal{Y}, \mathcal{B})$ is not dominated. Also show, that if $\mathcal{B}$ is generated by the collection of all half-open intervals $(x, y]$, where $x, y \in \mathbb{Z}, x < y$, then $\mathcal{M}_1^+(\mathcal{Y}, \mathcal{B})$ is dominated.

1.4. Although customarily the model is defined first and estimators follow, it is possible to reverse the order: suppose that we have a certain fixed estimator in mind, how should we choose the model in order for the fixed estimator to perform?

More explicitly, consider a data vector $Y = (X_1, \ldots, X_n)$ that forms an independent and identically distributed sample from a unknown distribution $P_0$ on $\mathbb{R}$. We are interested in estimation of the quantity $\psi = P_0g(X)$, assumed to be finite, where $g : \mathbb{R} \rightarrow \mathbb{R}$ is a given measurable function defined on the sample space for $X$. Examples: if $g$ is the identity map, then $\psi$ is the expectation of $X$; if $g = (X - P_0X)^2$, then $\psi$ is the variance of $X$; if $g = 1\{X \leq x\}$ for some $x \in \mathbb{R}$, then $\psi = P_0(x)$, the value of the distribution function associated with $P_0$ at $x$. In such cases, estimation of $\psi$ by the sample mean $\bar{\psi}_n := \frac{1}{n} \sum X_i$ appears sensible for large $n$.

a. Based on the limit $n \rightarrow \infty$, characterize the largest model $\mathcal{P}$ in which estimation of $\psi$ by $\bar{\psi}_n$ makes sense. What can be said of $\mathcal{P}$ if $g$ is a bounded function?

b. Show that $\bar{\psi}_n$ is consistent: $\lim_n \bar{\psi}_n = \psi$, $P_0$-almost-surely.

To analyze the behaviour of $\bar{\psi}_n$ in some more detail, consider the following question.

c. Restrict the model $\mathcal{P}$ further to demonstrate that $\bar{\psi}_n$ converges to $\psi$ at rate $n^{-1/2}$, i.e. for every sequence $M_n \rightarrow \infty$,

$$P_0(\sqrt{n}(\bar{\psi}_n - \psi) > M_n) \rightarrow 0,$$

as $n \rightarrow \infty$. 

d. Compare the property under c. above with example 1.6 and state in words how the quality of $\hat{\psi}_n$ as an estimator for $\psi$ improves as $n \to \infty$.

1.5. In the three experiments of example 1.8, describe a test for hypotheses $H_0$ and $H_1$ at level $\alpha \in (0,1)$, for example the likelihood ratio test. Calculate the $p$-value of the realization of 10 successes and 0 failures (in 10 Bernoulli trials according to $H_0$).
Chapter 2
Bayesian basics

In this chapter, we consider the basic definitions and properties of Bayesian statistical and decision-theoretic methods. We derive the posterior distribution from data, model and prior and we discuss how the posterior should be viewed if one assumes the frequentist point of view of section 2.1. In section 2.2 we consider point estimators derived from the posterior and in section 2.3 we discuss confidence sets and credible sets. Section 2.4 discusses the Neyman-Pearson theory of hypothesis testing, as well as a brief introduction to the Le Cam’s theory of asymptotically optimal test sequences and, of course, posterior odds and Bayes factors. Section 2.5 concludes the chapter with a discussion of minimax risk and Bayes risk, and the differences that occur in decision theory. Throughout the chapter the explicit goal is to consider frequentist methods side-by-side with the Bayesian procedures, for comparison and reference.

2.1 Bayes’s rule, prior and posterior distributions

In this section, we introduce the basic definitions and procedures in Bayesian statistics. Formalizing the Bayesian procedure can be done in several ways. We start this section with considerations that are traditionally qualified as being of a “subjectivist” nature: in subsection 2.1.1 we derive the relation between data, model and prior on the one hand, and the posterior on the other based on Bayes’s Rule without reference to the frequentist’s “true distribution of the data”. We take due care with regard to the support of the prior (see subsection 2.1.3) and consider a prototypical example usually referred to as Bayes’s Billiard in subsection 2.1.2. In subsection 2.1.4 we revert to the “frequentist” point of view through an assumption relating the “true distribution of the data” to the prior predictive distribution (see definition 2.4).
2.1.1 Bayes’s rule

The Bayesian framework does not just view the data $Y$ as a random variable but casts the parameter in that form as well. The parameter space $\Theta$ is assumed to be a measurable space, with $\sigma$-algebra $\mathcal{G}$ and, rather than just taking on fixed values $\theta$ as in the frequentist case, the parameter is represented by a random variable $\vartheta$ taking values in $\Theta$. We assume that on the product-space $\mathcal{Y} \times \Theta$ (with product $\sigma$-algebra $\sigma(\mathcal{B} \times \mathcal{G})$) we have a probability measure,$$
abla^* : \sigma(\mathcal{B} \times \mathcal{G}) \to [0,1]. \quad (2.1)
$$

The probability measure $\nabla^*$ provides a joint probability distribution for $(Y, \vartheta)$, where $Y$ is the observation and $\vartheta$ (the random variable associated with) the parameter of the model.

Implicitly the choice for the measure $\nabla^*$ defines the model in Bayesian context, by the possibility to condition the distribution of $Y$ on fixed values $\vartheta = \theta$ in $\Theta$. The conditional distribution for $Y|\vartheta$ (see appendix B.5) describes the distribution of the observation $Y$ given the parameter $\vartheta$. As such, the distributions for $Y|\vartheta = \theta$ can be identified with the elements $P_\theta$ of what was referred to as a parametrized model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ in chapter 1.

**Definition 2.1.** The distribution of the data $Y$ conditional on the parameter $\vartheta$ (c.f. definition B.15) is a regular conditional distribution,

$$
\Pi_{Y|\vartheta} : \mathcal{B} \times \Theta \to [0,1], \quad (2.2)
$$

which describes the model distributions. (see definition B.16). Since conditional probabilities are defined almost-surely with respect to the marginal (see definition B.15), the Bayesian notion of a model is represented only up to null-sets of the marginal distribution of $\vartheta$, referred to in Bayesian context as the prior for the parameter $\vartheta$.

**Definition 2.2.** The marginal distribution $\Pi : \mathcal{G} \to [0,1]$ for $\vartheta$ is called the prior.

The prior is interpreted in the subjectivist’s philosophy as the “degree of belief” attached to subsets of the model a priori, that is, before any observation has been made or incorporated in the calculation. It is important to note that $\nabla^*$ is usually constructed by choice of a prior measure $\Pi$ for $\vartheta$ and model distributions $\theta \mapsto P_\theta$,

$$
\Pi^*(A \times B) = \int_B \Pi(A|\theta = \theta) d\Pi(\theta) = \int_B P_\theta(A) d\Pi(\theta),
$$

for all $A \in \mathcal{B}$ and $B \in \mathcal{G}$ (where $\theta \mapsto P_\theta(A)$ is assumed to be measurable for all $A \in \mathcal{B}$). Central to the Bayesian framework is the conditional distribution for $\vartheta$ given $Y$, called the posterior. The transition from prior to posterior represents the way in which “prior belief” is turned into “posterior belief” (concerning the parameter) based on the data. The posterior is interpreted as a data-amended version of the prior,
that is to say, the subjectivist’s original “degree of belief” corrected by observation of $Y$ through conditioning. Below we define the posterior in conjunction with the marginal distribution for the data, the so-called prior predictive distribution.

**Definition 2.3.** The conditional distribution $\Pi_{\theta|Y} : \mathcal{G} \times \mathcal{Y} \to [0, 1]$ for $\theta|Y$ is called the posterior distribution.

The definition of the posterior is almost-sure with respect to the marginal data distribution $P$ (see definition B.14 and the concluding remarks of subsection B.5).

**Definition 2.4.** The marginal distribution $P_{\Pi} : \mathcal{B} \to [0, 1]$ for $Y$ is called the prior predictive distribution. If, in the above, one replaces the prior by the posterior, the resulting distribution for $Y$ is referred to as the posterior predictive distribution.

In the subjectivist philosophy, the prior predictive distribution describes a subjectivist’s expectations concerning the observation $Y$ based only on model and prior, i.e. before involving the data or realizations thereof. Given model and prior, the prior predictive distribution is of mixture form.

**Lemma 2.1.** The prior predictive $P_{\Pi}$ can be expressed in terms of model distributions $P_{\theta}$ and the prior, taking the form,

$$P_{\Pi}(Y \in A) = \int_{\Theta} P_{\theta}(A) d\Pi(\theta),$$

(2.3)

for all $A \in \mathcal{B}$.

The Bayesian symmetry between observation and parameter invites an identity expressing its essence. Bayes’s Rule relates model distributions, prior, posterior and prior predictive distribution through

$$\int_{A} \Pi(B|Y = y) dP_{\Pi}(y) = \int_{B} P_{\theta}(A) d\Pi(\theta),$$

(2.4)

for all $A \in \mathcal{B}$ and $B \in \mathcal{G}$ (see proposition B.2). The following theorem restates this fact in terms of the concepts we have introduced above, in a property which is sometimes referred to as a disintegration of the joint measure on model times samplespace: (2.4) should be viewed as a double-sided version of definition B.15.

**Theorem 2.1.** Posterior, prior predictive, model distributions and prior are related through Bayes’s Rule,

$$\int_{A} \Pi(B|Y = y) dP_{\Pi}(y) = \int_{B} P_{\theta}(A) d\Pi(\theta),$$

(2.4)

for all $A \in \mathcal{B}$ and $B \in \mathcal{G}$.

**Proof.** Equality (2.4) follows since both sides are equal to $\Pi^*(A \times B)$, c.f. definition B.15.

Note that, given model and prior, property (2.4) characterizes the posterior, up to re-definition on null sets of the prior predictive distribution $P_{\Pi}$. Consequently, we may turn this theorem around and use property (2.4) as the defining property of the posterior.
Definition 2.5. Given model and prior, any map $\pi : \mathcal{G} \times \mathcal{Y} \to [0, 1]$ such that $y \mapsto \pi(B, y)$ is measurable for all $B \in \mathcal{G}$ and such that $\pi$ satisfies,

$$\int_A \pi(B, y) dP^\Pi(y) = \int_B P_\theta(A) d\Pi(\theta),$$

(2.5)

for all $A \in \mathcal{B}$ and $B \in \mathcal{G}$, is called a version of the posterior.

Unfortunately property (2.5) does not imply that $\pi$ is a regular conditional probability, so we are left with an equivalence in which property 2 of definition B.16 remains as a condition.

Proposition 2.1. A map $\pi : \mathcal{G} \times \mathcal{Y} \to [0, 1]$ is a regular version of the posterior iff $\pi$ is $\mathcal{G}$-measurable for all $B \in \mathcal{G}$ and satisfies (2.5), and $B \mapsto \pi(B, y)$ is a (probability) measure on $\mathcal{G}$ for $P^\Pi$-almost-all $y \in \mathcal{Y}$.

By contrast, the following construction does imply regularity. Assuming that the model $\mathcal{P}$ is dominated, there exists an expression for the posterior in terms of model densities. Since most statistical models are defined as families of densities (e.g. Lebesgue-densities on $\mathbb{R}$ or $\mathbb{R}^n$) this accessible form of the posterior is used often.

Theorem 2.2. Assume that the model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is dominated by a $\sigma$-finite measure $\mu$ on $(\mathcal{Y}, \mathcal{B})$ with densities $p_\theta = dP_\theta/d\mu$. Then the posterior can be expressed as,

$$\Pi(\theta \in B | Y) = \frac{\int_B p_\theta(Y) d\Pi(\theta)}{\int_\Theta p_\theta(Y) d\Pi(\theta)},$$

(2.6)

for all $B \in \mathcal{G}$. This version of the posterior is regular.

Proof. Since the model is dominated, the prior predictive distribution has a density with respect to $\mu$, because for every $A \in \mathcal{B}$,

$$P^\Pi(A) = \int_\Theta P_\theta(A) d\Pi(\theta) = \int_A \int_\Theta p_\theta(y) d\mu(y) d\Pi(\theta)$$

$$= \int_A \left( \int_\Theta p_\theta(y) d\Pi(\theta) \right) d\mu(y).$$

in accordance with the Fubini and Radon-Nikodym theorems. The prior predictive density $p^\Pi : \mathcal{Y} \to \mathbb{R}$ is therefore equal to the denominator on the r.h.s. of (2.6). Let $A \in \mathcal{B}$ and $B \in \mathcal{G}$ be given. Substituting (2.6) into the l.h.s. of (2.4), we obtain,

$$\int_A \Pi(B | Y = y) dP^\Pi(y) = \int_A \left( \int_B p_\theta(Y) d\Pi(\theta) / \int_\Theta p_\theta(Y) d\Pi(\theta) \right) dP^\Pi(y)$$

$$= \int_A \int_B p_\theta(y) d\Pi(\theta) d\mu(y) = \int_B P_\theta(A) d\Pi(\theta).$$

According to theorem 2.1, (2.6) is a version of the posterior and property 3 of definition B.16 is satisfied. Property 1 of definition B.16 follows from Fubini’s theorem.
2.1 Bayes’s rule, prior and posterior distributions

(which guarantees measurability of the r.h.s. of (2.6)). Since $P^\Pi(p^\Pi > 0) = 1$, the denominator in (2.6) is non-zero $P^\Pi$-almost-surely and the posterior is well-defined (as a map $\mathcal{G} \to [0, 1]$), $P^\Pi$-almost-surely. In addition, for all $y$ such that $p^\Pi(y) > 0$ and any sequence $(B_n)$ of disjoint, $\mathcal{G}$-measurable sets,

$$
\Pi \left( \theta \in \bigcup_{n \geq 1} B_n \bigg| Y = y \right) = (p^\Pi(y))^{-1} \int_{\bigcup_n B_n} p_{\theta}(y) d\Pi(\theta)
$$

$$
= (p^\Pi(y))^{-1} \sum_{n \geq 1} 1_{\theta \in B_n} p_{\theta}(y) d\Pi(\theta)
$$

$$
= \sum_{n \geq 1} (p^\Pi(y))^{-1} \int_{B_n} p_{\theta}(y) d\Pi(\theta) = \sum_{n \geq 1} \Pi(\theta \in B_n | Y = y),
$$

by monotone convergence. We have established that on an event of $P^\Pi$-measure one, this version of the posterior is well-defined and $\sigma$-additive, so that also property 2 of definition B.16 holds. Conclude that (2.6) is a regular version of the posterior.

In the rest of part I and most of part II, we shall not concern ourselves with regularity of posteriors any more: in all parametric and most non-parametric settings explored here and in the literature, the model is dominated or it is a Polish space (see theorem B.9), either of which implies existence of regular posteriors. But in part II we shall also encounter topological circumstances (from rather compelling theoretical arguments based on certain weak model topologies) and questions regarding regularity will resurface there.

To demonstrate that it is easy to define a model (with prior) that theorem B.9 does not cover, consider the following example.

**Example 2.1.** Suppose that the samplespace is $\mathbb{R}$ and the model $\mathcal{P}$ consists of all measures of the form (see example B.3):

$$
P = \sum_{j=1}^{m} \alpha_j \delta_{x_j}, \tag{2.7}
$$

for some $m \geq 1$, with $\alpha_1, \ldots, \alpha_m$ satisfying $\alpha_j \geq 0$, $\sum_{j=1}^{m} \alpha_j = 1$ and $x_1, \ldots, x_m \in \mathbb{R}$. A suitable prior for this model exists (if one is willing to allow $m = \infty$): distributions drawn from a so-called Dirichlet process prior (see section 6.4) are of the form (2.7) with probability one. There is no $\sigma$-finite dominating measure for this model (not even if we restrict to measures of the form (2.7) with $m = 1$) and the model can not be represented by a family of densities, c.f. definition 1.3. Definition (2.6) cannot be used in this case. We have to resort to definition 2.3 in order to make sense of the posterior distribution and existence of a version of the posterior that displays regularity is a concern in this case.

This model $\mathcal{P}$ can also be used as a parametrizing space for a so-called mixture model $\mathcal{P}'$ of distributions on $\mathbb{R}$. For a fixed probability distribution $F$ with Lebesgue density $f : \mathbb{R} \to \mathbb{R}$ and any probability distribution $P$ on $\mathbb{R}$, define the convolution $f_P$ as follows,
\[ f_P(x) = \int f(x-y) dP(y), \]

for (Lebesgue-almost-all) \( x \in \mathbb{R} \). Note that \( f_P \) is a Lebesgue probability density on \( \mathbb{R} \) (due to Fubini’s theorem), describing the distribution of the random variable \( Z = X + Y \), for some \((X,Y)\) that are independent and marginally, \( X \sim P \) and \( Y \sim F \). If we let \( P \) be from the model \( \mathcal{P} \) above, convolution defines a map from \( \mathcal{P} \) to a new model \( \mathcal{P}' \) of densities, \( \mathcal{P} \to \mathcal{P}' : P \mapsto f_P \), of the form,

\[ f_P(x) = \sum_{j=1}^{m} \alpha_j f(x_j - x), \]

where \( P \) is as in (2.7), a convex combination of \( m \) clusters in which the intra-cluster variability is described by the density \( f \): the model describes observation of a randomly selected cluster location \( X = x_j \) with random (e.g. noisy) displacement \( Y \sim F \).

### 2.1.2 Bayes’s billiard

To many who have been introduced to statistics from the frequentist point of view, treating the parameter \( \theta \) for the model as a random variable \( \vartheta \) seems somewhat unnatural because the frequentist role for the parameter is entirely different from that of the data. The following example demonstrates that in certain situations the Bayesian point of view is not unnatural at all.

**Example 2.2.** In the posthumous publication of “An essay towards solving a problem in the doctrine of chances” in 1763 [13], Thomas Bayes included an example of a situation in which the above, subjectivist perspective arises quite naturally. It involves a number of red balls and one white ball placed on a table and has become known in the literature as Bayes’s billiard.

We consider the following experiment: unseen by the statistician, someone places \( n \) red balls and one white ball on a billiard table of length 1. The statistician will be reported the number \( K \) of red balls that is closer to the cushion than the white ball (\( K \) plays the role of the data in this example) and is asked to give a distribution reflecting his beliefs concerning the position of the white ball \( X \) (\( X \) plays the role of the parameter) based on \( K \). Calling the distance between the white ball and the bottom cushion of the table \( X \) and the distances between the red balls and the bottom cushion \( Y_i \), \((i = 1, \ldots, n)\), it is known to the statistician that their joint distribution is:

\[ (X; Y_1, \ldots, Y_n) \sim U[0, 1]^{n+1}, \]

*i.e.* all balls are placed independently and uniformly. This distribution gives rise both to the model (for \( K \)) and to the prior (for \( X \)). Prior knowledge concerning \( X \) (i.e. without knowing the observed value \( K = k \)) offers little information: the best that can be said is that \( X \sim U[0, 1] \), *i.e.* the prior is uniform. The question is how
this distribution for $X$ changes when we incorporate the observation $K = k$, that is, when we use the observation to arrive at posterior beliefs.

Since for every $i$, $Y_i$ and $X$ are independent c.f. (2.8), we have model distributions that give rise to,

$$P(Y_i \leq X | X = x) = P(Y_i \leq x) = x,$$

for any $x \in [0, 1]$. So for each of the red balls, determining whether it lies closer to the cushion than the white ball amounts to a Bernoulli experiment with parameter $x$. Since in addition the positions $Y_1, \ldots, Y_n$ are independent, counting the number $K$ of red balls closer to the cushion than the white ball amounts to counting “successes” in a sequence of independent Bernoulli experiments. We conclude that $K$ has a binomial distribution $\text{Bin}(n; x)$, i.e.,

$$P(K = k | X = x) = \frac{n!}{k!(n-k)!} x^k (1-x)^{n-k}. $$

It is possible to obtain the density for the distribution of $X$ conditional on $K = k$ from the above display using Bayes’s Rule:

$$p(x | K = k) = P(K = k | X = x) \frac{p(x)}{P(K = k)}, \quad (2.9)$$

![Fig. 2.1 Posterior densities for the position $X$ of the white ball, given the number $k$ of red balls closer to the cushion of the billiard (out of a total of $n = 6$ red balls). For the lower values of $k$, the white ball is close to the cushion with high probability, since otherwise more red balls would probably lie closer to the cushion. This is reflected by the posterior density for $X | K = 1$, for example, by the fact that it concentrates much of its mass close to $x = 0.$]
but in order to use it, we need the two marginal densities \( p(x) \) (the prior density) and \( P(K = k) \) (the prior predictive density) in the fraction. From (2.8) it is known that \( p(x) = 1 \) and \( P(K = k) \) can be obtained by integrating,

\[
P(K = k) = \int_0^1 P(K = k | X = x) p(x) \, dx.
\]

Substituting in (2.9), we find:

\[
p(x | K = k) = \frac{P(K = k | X = x) p(x)}{\int_0^1 P(K = k | X = x) p(x) \, dx} = B(n, k) x^k (1 - x)^{n-k},
\]

where \( B(n, k) \) is a normalization factor. The \( x \)-dependence of the density in the above display reveals that \( X | K = k \) is distributed according to a Beta-distribution, \( B(k+1, n-k+1) \), so that the normalization factor \( B(n, k) \) must equal \( B(n, k) = \Gamma(n+2)/\Gamma(k+1)\Gamma(n-k+1) \).

This provides the statistician with distributions reflecting his beliefs concerning the position of the white ball for all possible values \( k \) for the observation \( K \). Through conditioning on \( K = k \), the prior distribution of \( X \) is changed into the posterior for \( X \): if a relatively small number of red balls is closer to the cushion than the white ball (i.e. in case \( k \) is small compared to \( n \)), then the white ball is probably close to the cushion; if \( k \) is relatively large, the white ball is probably far from the cushion (see figure 2.1). The illustration on the cover of appears in [13] and is Bayes’s own version of his Billiard, complete with Beta-density drawn along the bottom edge.

### 2.1.3 The Bayesian view of the model

Based on the definitions of subsection 2.1.1 a remark is in order with regard to the notion of the model in Bayesian statistics: if, for a subset \( \mathcal{P}_1 \subset \mathcal{P} \), the prior assigns mass zero, then for all practical purposes \( \mathcal{P}_1 \) does not play a role since omission of \( \mathcal{P}_1 \) from \( \mathcal{P} \) does not influence the posterior. As long as the model is parametric, i.e. \( \Theta \subset \mathbb{R}^d \), we can always use priors that dominate the Lebesgue measure, ensuring that any \( \mathcal{P}_1 \) of prior measure zero has Lebesgue measure zero in \( \Theta \) and can therefore be thought of as negligibly small. However, in non-parametric models null-sets of the prior and posterior may be much larger than expected intuitively.

**Example 2.3.** Taking the above argument to the extreme, consider a normal location model \( \mathcal{P} = \{ N(\theta, 1) : \theta \in \mathbb{R} \} \) with a prior \( \Pi = \delta_{\theta_1} \) (see example B.3), for some \( \theta_1 \in \mathbb{R} \), defined on the Borel \( \sigma \)-algebra \( \mathcal{B} \). Then the model is dominated by the Lebesgue measure and the posterior takes the form:

\[
\Pi(\theta \in A | Y) = \int_A p_{\theta}(Y) d\Pi(\theta) / \int_{\Theta} p_{\theta}(Y) d\Pi(\theta) = \frac{p_{\theta_1}(Y)}{p_{\theta_1}(Y)} \Pi(A) = \Pi(A).
\]
for any \( A \in \mathcal{B} \). In other words, the posterior equals the prior, concentrating all its mass in the point \( \theta_1 \). Even though we started out with a model that suggests estimation of location, effectively the model consists of only one point due to the degeneracy of the prior. In subjectivist terms, the prior belief is fully biased towards \( \theta_1 \), leaving no room for amendment by the data when we condition to obtain the posterior.

This example raises the question which part of the model proper \( \mathcal{P} \) plays a role in the Bayesian approach. From a topological perspective it is helpful to make the following definition.

**Definition 2.6.** In addition to \((\Theta, \mathcal{G}, \Pi)\) being a probability space, let \((\Theta, \mathcal{T})\) be a topological space and assume that \( \mathcal{G} \) contains the Borel \( \sigma \)-algebra \( \mathcal{B} \) corresponding to the topology \( \mathcal{T} \). The support \( \text{supp}(\Pi) \) of the prior \( \Pi \) is defined as the smallest closed set \( S \) such that \( \Pi(\Theta \setminus S) = 0 \).

It is tempting to equate the support of a prior to the set described by the following intersection.

\[
S = \bigcap \{ G \in \mathcal{G} : G \text{ closed}, \; \Pi(\Theta \setminus G) = 0 \}.
\]  
(2.10)

Perhaps somewhat surprisingly, the validity of this identification is hard to establish: for any \((\Theta, \mathcal{G})\) as in definition 2.6, \( S \) is measurable, in fact, \( S \) is (an intersection of closed sets so \( S \) is) closed. Since the Borel \( \sigma \)-algebra is generated by the open sets (see definition B.1 and the example that follows), \( S \in \mathcal{B} \subseteq \mathcal{G} \). To show that \( \Pi(\Theta \setminus S) = 0 \) requires extra conditions on the space \( \Theta \); the following lemma covers a large (but not exhaustive) class of models.

**Lemma 2.2.** In addition to \((\Theta, \mathcal{G}, \Pi)\) being a probability space, let \((\Theta, \mathcal{T})\) be a topological space and assume that \( \mathcal{G} \) contains the Borel \( \sigma \)-algebra \( \mathcal{B} \) corresponding to the topology \( \mathcal{T} \). If \( \mathcal{T} \) is second countable, \( S = \text{supp}(\Pi) \).

**Proof.** Consider the complement \( V = \Theta \setminus S \). We can write,

\[
V = \bigcup \{ U \in \mathcal{G} : U \text{ open}, \; \Pi(U) = 0 \}.
\]

The set \( V \) is open and contains every open subset of \( \Pi \)-measure zero. Because the topology is second countable, \( V \) can be written as a countable union of open sets \( \{ U_k : k \geq 1 \} \) of \( \Pi \)-measure zero. Therefore, \( \Pi(V) = \Pi(\bigcup_{k \geq 1} U_k) \leq \sum_{k \geq 1} \Pi(U_k) = 0 \) and we conclude that \( \Pi(S) = 1 \).

This implies, for example, that the support of \( \Pi \) is of the form (2.10) if \((\Pi, \mathcal{T})\) is a separable metrizable space (the interested reader should describe the countable basis for \( \mathcal{T} \) in that case).

**Example 2.4.** In example 2.3, the model \( \mathcal{P} \) consists of all normal distributions of the form \( N(\theta, 1) \), \( \theta \in \mathbb{R} \), but the support of the prior \( \text{supp}(\Pi) \) equals the singleton \( \{ N(\theta_1, 1) \} \subseteq \mathcal{P} \).
Note that the support of the prior is defined based on a topology, the Borel $\sigma$-algebra of which must belong to the domain of the prior measure. In parametric models this assumption is rarely problematic but in non-parametric models, finding such a prior may be difficult and the support may be an ill-defined concept. Therefore we may choose to take a less precise but more generally applicable perspective: the model is viewed as the support of the prior $\Pi$, but only up to $\Pi$-null-sets (c.f. the $\Pi$-almost-sure nature of the identification (2.2)). That means that we may add to or remove from the model at will, as long as we make sure that the changes have prior measure equal to zero: the model itself is a $\Pi$-almost-sure concept.

To demonstrate how the notion of the support of a probability measure can collide with the above remark concerning prior null-sets in non-parametric models, the following example.

**Example 2.5.** Consider again the full model on a countable samplespace $\mathcal{X} = \{i : i \geq 1\}$, as in example 1.1, and represent it as the $\ell_1$-subset, 

$$\Lambda = \{p \in \ell_1 : p_i \geq 1, \Sigma_{i \geq 1} p_i = 1\}.$$ 

Also define the subsets $\Lambda_k \subset \Lambda, (k \geq 1)$, $\Lambda_k = \{p \in \Lambda : p_i = 0, i \geq k\}, \Lambda_0 = \cup_k \{\Lambda_k : k \geq 1\}$ and $\mathcal{N} = \{p \in \Lambda : p_i > 0, i \geq 1\}$. Note that $\mathcal{N}$ can be thought of as describing a generic point in $\Lambda$ (and this is made rigorous when one remarks that $\mathcal{N}$ is residual in $\Lambda$ in the Baire sense [134]). For all $k \geq 1$, place priors $\Pi_k$ of full support on the finite-dimensional simplices that the $\Lambda_k$ describe (and embed in $\Lambda$). Define a prior $\Pi$ on $\Lambda$ based on a sequence $(\lambda_k)$ such that $\lambda_k > 0$ for all $k \geq 1$ and $\sum_k \lambda_k = 1$, 

$$\Pi(A) = \sum_{k \geq 1} \lambda_k \Pi_k(A).$$ 

It is noted that the normed space $\ell_1$ is separable, so the support $S$ of $\Pi$ is well-defined and coincides with (2.10). To find $S$, let $\varepsilon > 0$ and $p \in \Lambda$ be given. There exists a $k \geq 1$ such that $\sum_{i > k} p_i < \frac{1}{2} \varepsilon$. Therefore, there exists a $q \in \Lambda_k$ and an $\ell_1$-neighbourhood $U$ of $q$ in $\Lambda_k$ such that for all $q' \in U, \|p - q'\| < \varepsilon$. Therefore, 

$$\Pi(\{q \in \Lambda : \|p - q\| < \varepsilon\}) \geq \lambda_k \Pi_k(U) > 0.$$ 

Conclude that $S = \Lambda$, that is, $\Pi$ is of full support. Nevertheless, $\mathcal{N} \cap \Lambda_k = \emptyset$ for all $k \geq 1$, so $\Pi(\mathcal{N}) = 0$. The set $\mathcal{N}$ is a null-set of $\Pi$: any Bayesian analysis with this prior involves support $\Lambda$ but $\Lambda \setminus \mathcal{N}$ is called the model with equal validity from the Bayesian perspective.

The conceptual confusion about where a prior places its mass in a topological sense lies at the heart of the first examples of Bayesian non-parametric asymptotic inconsistency [93].
2.1.4 A frequentist’s view of the posterior

So far, we have not discussed the details of the data $Y$, we have treated $Y$ completely abstractly. In this section we consider, firstly, the relation between the frequentist distribution of $Y$ (the “true” $P_0$) and the Bayesian distribution of $Y$ (the marginal $P^\Pi$). Secondly, we consider samples of independent, repeated measurements of a random variable $X$. We shall see that the Bayesian way to describe data and statistical experiments is in contradiction with the frequentist assumption. The section concludes with a sort of *recipe* listing the steps involved in the Bayesian analysis of a data set. These steps are illustrated with a very simple parametric example.

The derivation of the posterior in subsection 2.1.1 does not refer to the any “true, underlying distribution of the data” but it does involve a marginal distribution for $Y$, the prior predictive distribution of definition 2.4. If one adopts the frequentist framework to analyze Bayesian tools like the posterior, a discrepancy arises since $P_0$ and $P^\Pi$ are two distributions for the data $Y$ that are not equal (for a striking instance of the discrepancy, see remark 2.1 below). To the frequentist, $P^\Pi$ is a side-product of the Bayesian construction that has no realistic interpretation. There is, however, a clear technical issue: all definitions and derivations in subsection 2.1.1 are almost-sure with respect to the prior predictive distribution. To ensure that all of this continues to make sense after we adopt assumption (1.1) we require that $P^\Pi$ dominates $P_0$:

$$P_0 \ll P^\Pi. \quad (2.11)$$

In that case, null-sets of $P^\Pi$ are also null-sets of $P_0$, so that all $P^\Pi$-almost-sure statements and definitions are also $P_0$-almost-sure. In particular, expression (2.6) for the posterior in a dominated model satisfies the regularity condition not only $P^\Pi$- but also $P_0$-almost-surely, if we assume (2.11). We shall adopt the frequentist philosophy to analyse Bayesian tools, i.e. we assume (1.1) and (2.11).

In many experiments or observations, the data consists of a sample of $n$ repeated, stochastically independent measurements of the same quantity (an i.i.d. sample). To accommodate this or other situations where the data is gathered sequentially, we assume that we observe data $X^n$ taking values in measurable spaces $(X_n, B_n)$ for all $n \geq 1$, and we consider parametrized models $\Theta \to P_n : \theta \mapsto P_{\theta,n}$. The frequentist assumes that there is some sequence of probability measures $(P_{0,n})$ such that $X^n \sim P_{0,n}$ for all $n \geq 1$, and often, that there exists a $\theta_0 \in \Theta$, such that, $P_{0,n} = P_{\theta_0,n}$ for all $n \geq 1$. In the case of i.i.d. data from a measurable space $(X, \mathcal{B})$, $X^n = (X_1, \ldots, X_n) \in X^n$ with $\Theta$ some collection $\mathcal{P}$ of probability measures $P$ on $(X, \mathcal{B})$ and parametrization $\mathcal{P} \to P_n : \theta \mapsto P_{\theta,n}$. Assuming a well-specified model $\mathcal{P}$ implies the existence of some $P_0 \in \mathcal{P}$ such that $P_{0,n} = P_0^n$ for all $n \geq 1$.

For Bayesians $(\Theta, \mathcal{G}, \Pi)$ is a measurable space and $\Theta \to \mathcal{P}_n : \theta \mapsto P_{\theta,n}(B)$ must be measurable for all $B \in \mathcal{B}_n$. But because Bayesians do not entertain the concept of a ‘true’ distribution of the data, they express assumptions concerning the data only through model distributions. Particularly for the i.i.d. assumption, the Bayesian assumes *conditional independence* of the observations, given $\vartheta = \theta$:
for all \((A_1, \ldots, A_n) \in \mathcal{A}^n\) and \(\Pi\)-almost all \(\theta\). Similarly we see that the prior predictive distribution for \(i.i.d.\) data takes the form:

\[
P^{\Pi}_n(X_1 \in A_1, \ldots, X_n \in A_n) = \int_\Theta \prod_{i=1}^n P_\theta(A_i) d\Pi(\theta).
\]

The posterior is now a solution for Bayes’s Rule in the following form,

\[
\int_A \Pi(B|X_1 = x_1, \ldots, X_n = x_n) dP^{\Pi}_n(x_1, \ldots, x_n) = \int_B \prod_{i=1}^n P_\theta(A_i) d\Pi(\theta),
\]

where \(A = A_1 \times \ldots \times A_n\), \(B \in \mathcal{G}\). Assuming that the model \(\mathcal{P} = \{P_\theta : \theta \in \Theta\}\) for the marginal distributions is dominated by a \(\sigma\)-finite measure \(\mu\) on \(\mathcal{F}\), the above can also be expressed in terms of \(\mu\)-densities \(p_\theta = dP_\theta/d\mu\). Using theorem 2.2 we obtain the following expression for the posterior distribution:

\[
\Pi(\theta \in B|X_1, X_2, \ldots, X_n) = \int_B \prod_{i=1}^n p_\theta(X_i) d\Pi(\theta) / \int_\Theta \prod_{i=1}^n p_\theta(X_i) d\Pi(\theta),
\]  

(2.12)

for any \(B \in \mathcal{G}\). Since \(P_0(p_0(X) > 0) = 1\), the assumption that \((X_1, \ldots, X_n) \sim P^n_0\) allows us to rewrite this expression with likelihood ratios,

\[
\Pi(\theta \in B|X_1, X_2, \ldots, X_n) = \int_B \prod_{i=1}^n \frac{p_\theta(X_i)}{p_0(X_i)} d\Pi(\theta) / \int_\Theta \prod_{i=1}^n \frac{p_\theta(X_i)}{p_0(X_i)} d\Pi(\theta),
\]  

(2.13)

\(P^n_0\)-almost-surely. In a dominated model, the Radon-Nikodym derivative (see theorem B.7) of the posterior with respect to the prior is the likelihood function, normalized to be a probability density function:

\[
\frac{d\Pi(\cdot|X_1, \ldots, X_n)}{d\Pi}(\theta) = \prod_{i=1}^n p_\theta(X_i) / \int_\Theta \prod_{i=1}^n p_\theta(X_i) d\Pi(\theta),
\]  

(2.14)

\(P^{\Pi}_n\)-almost-surely, and under (2.11), also \(P^n_0\)-almost-surely. The latter fact explains why such strong relations exist between Bayesian and maximum-likelihood methods. Indeed, the proportionality of the posterior density and the likelihood provides a useful qualitative picture of the posterior as a measure that concentrates on regions in the model where the likelihood is relatively high. This may serve as a direct, Fisherian motivation for the use of Bayesian methods in a frequentist context, c.f. section 1.4.

Remark 2.1. Note that the prior predictive distribution for \(i.i.d.\) data is itself not a product distribution but a mixture of product distributions! This illustrates the discrepancy between \(P_0\) and \(P^{\Pi}\) quite clearly: while the true distribution of the data describes an \(i.i.d.\) random vector, the prior predictive distribution describes a ran-
dom vector that is just \textit{exchangeable} (in accordance with De Finetti’s theorem (see theorem A.2.2)).

\textbf{Remark 2.2.} For the frequentist to use Bayesian tools, \textit{e.g.} a posterior calculated using (2.12), he has to assume condition (2.11). In the context of \textit{i.i.d.} samples, that requirement takes the form,

\[ P_0^n \ll P_n^{\Pi}, \text{ (for all } n \geq 1). \]

\subsection*{2.1.5 From prior to posterior}

To summarize, the Bayesian procedure consists of the following steps,

(i) Based on the background of the data \( Y \), the statistician chooses a model \( \mathcal{P} \) of “reasonable” candidate distributions for the data (usually with some parameterization \( \Theta \rightarrow \mathcal{P} : \theta \mapsto P_\theta \)).

(ii) A prior measure \( \Pi \) on \( \mathcal{P} \) is chosen, reflecting “belief” concerning these candidates, see chapter 3, (usually as a probability measure on \( \Theta \)).

(iii) Based on definition 2.3, on expression (2.6) or in the case of an \textit{i.i.d.} sample, on (2.12), we calculate the posterior as a function of the data \( Y \).

(iv) We observe a realization of the data \( Y = y \) and use it to calculate a realisation of the posterior.

The statistician may then infer properties of the parameter \( \theta \) from the posterior \( \Pi(\cdot | Y = y) \). One important point: when reporting the results of any statistical procedure, one is obliged to reveal all relevant details concerning the procedure followed and the data. So when making inference on \( \theta \), the statistician should report on the nature of the sample used and his choice of model, and in the Bayesian case, should always report his choice of prior as well, with a clear motivation.

\textbf{Example 2.6.} To illustrate the above “recipe” with a concrete example, consider the one-dimensional parametric model \( \mathcal{P} \) consisting of exponential distributions:

\[ \mathcal{P} = \{ \text{Exp}(\theta) : \theta \in \Theta = (0, \infty) \}. \]

Lebesgue measure dominates the model and densities take the form \( p_\theta(x) = \theta \exp(-\theta x) \), for \( x \geq 0 \). Assume that the data consists of \( n \) observations, (conditionally) independent and identically distributed. As a prior on the model, we take another exponential distribution with density \( \pi(\theta) = \exp(-\theta) \) (for \( \theta \in \Theta \)). The posterior density relative to Lebesgue measure on \( \Theta \) takes the form,

\[ d\Pi|X_1, \ldots, X_n = C(X_1, \ldots, X_n) \left( \prod_{i=1}^n \theta e^{-\theta X_i} 1_{\{X_i \geq 0\}} \right) e^{-\theta} d\theta \]

where \( C(X_1, \ldots, X_n) \) denotes the (data-dependent) normalization factor that makes the posterior a probability measure. We calculate,
\[ d\Pi(\theta|X_1, \ldots, X_n) = C(X_1, \ldots, X_n) \theta^n e^{-\theta(1+\sum_i X_i)} 1_{\{X_1 \geq 0\}} d\theta \]

(where \( X_{(1)} = \min X_i \)). Since,

\[ \int_0^\infty n! \theta^n e^{-\alpha \theta} d\theta = \frac{n!}{\alpha^{n+1}}. \]

we see that \( C(X_1, \ldots, X_n) \) must be equal to \( (1 + \sum_i X_i)^{n+1}/n! \). So for any measurable \( A \subset \Theta \), the posterior probability is given by:

\[ \Pi(\theta \in A|X_1, \ldots, X_n) = \frac{1}{n!} \left( 1 + \sum_{i=1}^n X_i \right)^{n+1} 1_{\{X_{(1)} \geq 0\}} \int_A \theta^n e^{-\theta(1+\sum_{i=1}^n X_i)} d\theta. \]

Note that the posterior density collapses to zero (and no longer describes a probability distribution!) if \( X_i < 0 \) for some \( 1 \leq i \leq n \). As Bayesians, we insist that the data must be compatible with the model, we require that \( \Pi^* (X_i \geq 0) = P^\Pi(X_i \geq 0) = 1. \)

As frequentists we involve the underlying distribution \( P_0 \), requiring that \( P_0(X \geq 0) = 1 \) so that the posterior is well-defined \( P_n^\Pi \)-almost-surely. More generally, \( P_n^\Pi \) dominates Lebesgue measure, so \( P_0 \ll P_n^\Pi \) as long as \( P_0 \) has a density with respect to Lebesgue measure.

### 2.2 Bayesian point estimators

When considering questions of statistical estimation, the outcome of a frequentist procedure is of a different nature than the outcome of a Bayesian procedure: a point-estimator (the frequentist outcome) gives a point in the model whereas the posterior (the Bayesian outcome) is a distribution on the model. A first question, then, concerns the manner in which to compare the two. We assume the frequentist philosophy to analyse Bayesian methodology (c.f. subsection 2.1.4) and, in this section, we derive point-estimators from the posterior distribution in various ways: we consider the posterior predictive distribution, as well as the parametric posterior mean and the maximum-a-posteriori estimator. In later sections we approach the matter from the opposite perspective: every point-estimator has a sampling distribution, which can be compared with the posterior because both are distributions on the model or the parameter space. This is the view that gives rise to the Bernstein-von Mises theorem of chapter 4.

#### 2.2.1 Posterior predictive and posterior mean

We think of a Bayesian point-estimator as a point in the model around which posterior mass is accumulated most, a point around which the posterior distribution is concentrated in some way. As such, any reasonable Bayesian point-estimator should
represent the “location” of the posterior distribution. However there is no unique def-
inition for the “location” of a distribution and, accordingly, there are many dif-
ferent ways to define Bayesian point-estimators.

Remark 2.3. Arguably, there are distributions for which even the existence of a “lo-
cation” is questionable. For instance, consider the convex combination of point-
masses $P = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{+1}$ on $(\mathbb{R}, \mathcal{B})$. Reasonable definitions of location, like the
mean and the median of $P$, all assign as the location of $P$ the point $0 \in \mathbb{R}$. Yet small
neighbourhoods of $0$ do not receive any $P$-mass, so $0$ can hardly be viewed as a point
around which $P$ concentrates its mass. The problem is not of a mathematical nature,
it is conceptual: when we think of the “location” of a distribution we normally think
of unimodal distributions which have unambiguous “locations”. However, it is com-
mon practice to formulate the notion for all distributions by the same definitions.

One quantity that is often used to represent a distribution’s location is its expec-
tation. This motivates the first and most Bayesian definition of a posterior-based
point-estimator: the posterior predictive distribution.

Definition 2.7. Consider a statistical problem involving data $Y$ taking values in a
samplespace $(\mathcal{Y}, \mathcal{B})$ and a model $(\mathcal{P}, \mathcal{G})$ with prior $\Pi$. Assume that all the maps
$\mathcal{P} \to [0, 1]: P \mapsto P(B)$, $(B \in \mathcal{B})$ are measurable with respect to $\mathcal{G}$ and that the
posterior $\Pi(\cdot|Y)$ is a regular conditional distribution. The posterior predictive dis-
tribution is a data-dependent set-function $\hat{P}: \mathcal{B} \to [0, 1]$, defined by,

$$\hat{P}(B) = \int_{\mathcal{P}} P(B) d\Pi(P | Y),$$

(2.15)

for every event $B \in \mathcal{B}$.

Proposition 2.2. The posterior predictive distribution $\hat{P}: \mathcal{B} \to [0, 1]$ is a probability
measure, almost surely.

Proof. (Assuming that the posterior is a regular conditional distribution) $\hat{P}$ is de-
efined $P^\Pi$-almost-surely as a map $\mathcal{B} \to [0, 1]$. Let $F \in \mathcal{B}$ denote the event that $\hat{P}$
is well-defined and let $y \in F$ be given. Clearly, for all $B \in \mathcal{B}$, $0 \leq \hat{P}(B) \leq 1$. Let $(B_i)_{i \geq 1} \subset \mathcal{B}$ be any sequence of disjoint events. Since $(P, i) \mapsto P(B_i)$ is non-negative
and measurable, Fubini’s theorem applies in the third equality below:

$$\hat{P}\left(\bigcup_{i \geq 1} B_i\right) = \int_{\mathcal{P}} P\left(\bigcup_{i \geq 1} B_i\right) d\Pi(P | Y = y) = \int_{\mathcal{P}} \sum_{i \geq 1} P(B_i) d\Pi(P | Y = y)$$

$$= \sum_{i \geq 1} \int_{\mathcal{P}} P(B_i) d\Pi(P | Y = y) = \sum_{i \geq 1} \hat{P}(B_i),$$

(or by monotone convergence), which proves $\sigma$-additivity of $\hat{P}$ for all $y \in F$, that is,
$P^\Pi$-almost-surely.

Remark 2.4. The qualification “almost surely” in the formulation of proposition 2.2
has distinct explanations for Bayesians and for frequentists: for the Bayesian, the
data $Y$ is marginally distributed according to the prior predictive distribution, so it is with respect to null sets of $P^\Pi$ that “almost surely” is to be interpreted in that case. By contrast, the frequentist assumes that $Y \sim P_0$, so he is forced to adopt assumption (2.11) and the interpretation of “almost surely” refers to null sets of $P_0$ in that case. This dual purpose of the phrase “almost surely” re-occurs in many places below.

Although we refer to $\hat{P}$ as a point-estimator (see definition 1.9), strictly speaking, we have not shown that $\hat{P} \in \mathcal{P}$ and indeed, generically, $\hat{P} \notin \mathcal{P}$ as the following proposition shows.

**Proposition 2.3.** Let $(\mathcal{Y}, \mathcal{B})$ be a Polish sample space with model $\mathcal{P}$. Endow $\mathcal{P}$ with the total-variational norm as a metric, assume that the model is pre-compact in the corresponding metric topology and let $\Pi$ be a Borel prior on $\mathcal{P}$. Then $\hat{P}$ lies in the closed convex hull of $\mathcal{P}$, almost-surely.

**Proof.** Let $\varepsilon > 0$ be given. Let $\hat{\mathcal{P}}$ denote the completion of $\mathcal{P}$. Since $\hat{\mathcal{P}}$ is a compact metric space, $\hat{\mathcal{P}}$ is Polish. The posterior extends trivially from a Borel probability measure on the subspace $\mathcal{P}$ to such a measure on $\hat{\mathcal{P}}$, so the posterior has a almost-surely unique regular version (see theorem B.9). Therefore, (2.15) is a well-defined probability measure, on an event $F \in \mathcal{B}$ of $P^\Pi$-probability one. Since $\mathcal{P}$ is pre-compact by assumption, for every $\varepsilon > 0$, there exists an $N \geq 1$ and $\{P_1, \ldots, P_N\} \subset \mathcal{P}$ such that the total-variation balls $B_1 = \{P' \in \mathcal{P} : \|P' - P_1\| < \varepsilon\}$ form a finite cover of $\mathcal{P}$, i.e. $\mathcal{P} = \bigcup_{i=1}^N B_i$. Through definition of $C_{i+1} = B_{i+1} \setminus C_i$ (for all $i \geq 1$, with $C_1 = B_1$) this cover is turned into a finite measurable partition $\{C_1, \ldots, C_N\}$ of $\mathcal{P}$. For $y \in F$ and $1 \leq i \leq N$, define $\lambda_i = \Pi(C_i | Y = y)$ and note that,

$$\|\hat{P} - \sum_{i=1}^N \lambda_i P_i\| = \sup_{B \in \mathcal{B}} \left| \sum_{i=1}^N \int_{C_i} (P(B) - P_i(B)) d\Pi(P | Y = y) \right|$$

$$\leq \sup_{B \in \mathcal{B}} \left| \sum_{i=1}^N \left( \int_{C_i} P(B) d\Pi(P | Y = y) - \int_{C_i} P_i(B) d\Pi(P | Y = y) \right) \right|$$

$$\leq \sum_{i=1}^N \int_{C_i} \sup_{B \in \mathcal{B}} |P(B) - P_i(B)| d\Pi(P | Y = y) \leq \varepsilon.$$

We have shown that there exist elements in the convex hull $\text{co}(\mathcal{P})$ of $\mathcal{P}$ that are arbitrarily close to $\hat{P}$ in total variation and conclude that $\hat{P}$ lies in its closure $\text{cl}(\mathcal{P})$.

Two remarks regarding the pre-compactness requirement in proposition 2.3: firstly, if the model is (such that it has as its completion) a Polish space, any prior is a Radon measure (on the completion), so for any $\varepsilon$ there is a compact subset (of the completion) $\mathcal{P}'$ such that $\Pi(\mathcal{P}') > 1 - \varepsilon$. Conditional on $\mathcal{P}'$ the (pre-)compactness condition is satisfied and the resulting approximate posterior predictive distribution $\hat{P}'$ lies in the (compact, hence closed) convex hull of $\mathcal{P}'$. Secondly, in part II, we shall see that there are uniform topologies weaker than the total-variational topology, in which all models are pre-compact: with a regular posterior, the above proof
continues to hold so the posterior predictive distribution always lies in the closed convex hull of the model (albeit with a wider notion of what closure means).

In many practical situations, the model \( \mathcal{P} \) is parametric with measurable parameterization \( \Theta \to \mathcal{P} : \theta \mapsto P_\theta \). In case \( \Theta \) is convex, a different form of “averaging over the model” applies.

**Definition 2.8.** Let the model \( \mathcal{P} \) be parametric with measurable parameterization \( \Theta \to \mathcal{P} : \theta \mapsto P_\theta \). Assume that \( \Theta \) is a Borel measurable, convex subset \( \Theta \subseteq \mathbb{R}^d \). Let \( \Pi \) be a Borel prior defined on \( \Theta \) with a posterior \( \Pi(\cdot|Y) \). If \( \vartheta \) is integrable with respect to the posterior, the posterior mean is defined

\[
\hat{\theta}_1(Y) = \int_{\Theta} \theta \, d\Pi(\theta|Y) \in \Theta,
\]

almost-surely.

In the above definition the posterior is has an almost-surely unique regular version because the completion of \( \Theta \) is a Polish subspace of \( \mathbb{R}^d \). For the same reason, approximate

**Example 2.7.** In example 2.6 the posterior takes the form:

\[
\Pi(\vartheta \in A|X_1, \ldots, X_n) = \frac{1}{n!} \left( 1 + \sum_{i=1}^n X_i \right)^{n+1} \int_A \theta^n e^{-\theta(1+\sum_{i=1}^n X_i)} \, d\theta.
\]

Assuming that \( P_0(X \geq 0) = 1 \), we omit the indicator for \( X_{(1)} \geq 0 \) and write the posterior mean of definition 2.8 as follows:

\[
\hat{\theta}_1(Y) = \int_{\Theta} \theta \, d\Pi(\theta|Y) = \frac{1}{n!} \left( 1 + \sum_{i=1}^n X_i \right)^{n+1} \int_0^\infty \theta^n e^{-\theta(1+\sum_{i=1}^n X_i)} \, d\theta
\]

\[
= \frac{1}{n!} \left( 1 + \sum_{i=1}^n X_i \right)^{-1} \int_0^\infty \zeta^{n+1} e^{-\zeta} \, d\zeta = (n+1) \left( 1 + \sum_{i=1}^n X_i \right)^{-1},
\]

where we have used that \( \int_0^\infty \zeta^{n+1} e^{-\zeta} \, d\zeta = \Gamma(n+2) = (n+1)! \).

In definition 2.8 convexity of \( \Theta \) is a condition (rather than an afterthought, as with definition 2.7): if \( \Theta \) is not convex there is no guarantee that \( \hat{\theta}_1 \in \Theta \), in which case \( P_{\hat{\theta}_1} \) is not defined since \( \hat{\theta}_1 \) does not lie in the domain of the parameterization. It is tempting to think that there is no difference between the posterior predictive and the posterior mean if the model is parametric. This is not the case, however, as demonstrated by the following (counter)example.

**Example 2.8.** Consider a normal location model in two dimensions for an observation \( Y \), where the location \( \mu \in \mathbb{R}^2 \) lies on the unit circle and the covariance \( \Sigma \) is fixed and known:

\[
\mathcal{P} = \{ P_\theta = N(\mu(\theta), \Sigma) : \mu(\theta) = (\cos \theta, \sin \theta), \theta \in [0,2\pi) \}.
\]
This is an identifiable, one-dimensional parametric model with convex parameterizing space \( \Theta = [0, 2\pi] \). Assume that \( \Xi \) is the uniform distribution on \( \Theta \) (\( \Xi \) plays the role of the posterior; it does not matter what shape the posterior really has, all we need is a counterexample). We define the corresponding measure \( \Xi' \) on \( \mathcal{P} \) by applying \( \Xi \) to the pre-image of the parameterization. By rotational symmetry of \( \Xi \) and Fubini’s theorem, the expectation of \( Y \) under \( \hat{P} \) is

\[
\int Y \, dP = \int_{\mathcal{P}} PY \, d\Xi'(P) = \int_\Theta P_\theta Y \, d\Xi(\theta) = \frac{1}{2\pi} \int_0^{2\pi} \mu(\theta) \, d\theta = (0, 0).
\]

Note that none of the distributions in \( \mathcal{P} \) has the origin as its expectation. We can also calculate the expectation of \( Y \) under \( P_{\hat{\theta}} \) in this situation:

\[
\hat{\theta}_1(Y) = \int_\Theta \theta \, d\Xi(\theta) = \frac{1}{2\pi} \int_0^{2\pi} \theta \, d\theta = \pi,
\]

which leads to \( P_{\hat{\theta}}Y = P_{\pi}Y = (-1, 0) \). Clearly, the posterior predictive distribution does not equal the model distribution corresponding to the posterior mean. In fact, the above illustrates that the posterior predictive distribution does not belong to the model necessarily: \( \hat{P} \not\in \mathcal{P} \).

The fact that the expectations of \( \hat{P} \) and \( P_{\hat{\theta}} \) in example 2.8 differ makes it clear that \( \hat{P} \neq P_{\hat{\theta}} \) unless special circumstances apply: if we consider a parameterization \( \theta \mapsto P_\theta \) from a (convex) parameterizing space \( \Theta \) (like example 2.1) with posterior measure \( \Pi(d\theta) \) onto a space of probability measures \( \mathcal{P} \) (with induced posterior \( \Pi(dP) \)), it makes a difference whether we consider definition (2.15), or calculate \( P_{\hat{\theta}} \).

### 2.2.2 Small-ball and formal Bayes estimators

Since there are multiple ways of defining the location of a distribution, there are multiple ways of obtaining point-estimators from the posterior distribution. A straightforward alternative for the posterior averages of the previous subsection, is given in the following definition which requires that the model is one-dimensional.

**Definition 2.9.** Let \( \Theta \) be a closed, non-empty subset of \( \mathbb{R} \) and let \( \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \) be a parametric model with Borel prior \( \Pi \) on \( \Theta \) and posterior \( \Pi(\cdot | Y) \). The **posterior median** is defined by,

\[
\tilde{\theta}(Y) = \inf\{ s \in \Theta : \Pi(\theta \leq s | Y) \geq 1/2 \},
\]

almost-surely. Thus the posterior median represents the smallest value for \( \theta \) such that the posterior mass to its left is greater than or equal to 1/2. This definition simplifies drastically in case the posterior has a continuous, (strictly) monotone distribution function: in
2.2 Bayesian point estimators

that case the above definition reduces to the perhaps more familiar definition as the
(unique) point $\hat{\theta} \in \Theta$ where $\Pi(\theta \leq \hat{\theta} | Y) = 1/2$. In some situations, the posterior
median offers an advantage over the posterior mean since its definition does not depend
on integrability requirements and because of robustness against perturbation
of the tails of the posterior.

Another alternative is decision-theoretic in essence (see section 2.5), that is, one
takes the perspective in which an assessment of loss is inherent. Suppose that we
consider estimation in a metric model $(\mathcal{P}, d)$ and we quantify errors in estimation
as follows: if the true distribution of the data is $P_0$ and we estimate that it is $P$,
then we incur a loss (to be specified further by the context of the problem) that is
a monotone increasing function $\ell: [0, \infty) \rightarrow [0, \infty)$ of the distance $d(P_0, P)$. If we
assume that the posterior concentrates its mass around $P_0$ (as well as possible) then
the estimator that minimizes the expected loss relative to the posterior optimizes the
so-called Bayesian risk function,

$$r(\Pi, \theta') = \int_{\Theta} P_\theta \ell(d(\theta, \theta')) d\Pi(\theta),$$

for given prior $\Pi$ and estimator $\theta'$.

Definition 2.10. Let $\mathcal{P}$ be a model with metric $d: \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$ and a Borel prior $\Pi$
on $\mathcal{P}$. Any monotone increasing function $\ell: [0, \infty) \rightarrow [0, \infty)$ is called a loss-function
if it is such that $\ell(0) = 0$. (Often $\ell$ is assumed to be convex or (semi-)continuous.)
Provided it exists, the formal Bayes estimator is a minimizer $\hat{P}$ of the function,

$$\mathcal{P} \rightarrow \mathbb{R} : P \mapsto \int_{\mathcal{P}} \ell(d(Q, P)) d\Pi(Q | Y),$$

over the model $\mathcal{P}$, defined almost-surely.

Note that definition 2.10 retains its form when expressed in terms of a formal Bayes
estimator $\hat{\theta}$ for a parameter $\theta \in \Theta$.

Theorem 2.3. Let the model $\mathcal{P}$ be dominated and parametrized by a metric space
$(\Theta, d)$. If the formal Bayes estimator $\hat{\theta}$ is well-defined, it minimizes the Bayesian
risk function:

$$r(\Pi, \hat{\theta}) = \inf_{\theta' \in \Theta} r(\Pi, \theta').$$

Proof. Rewrite the Bayesian risk function for the formal Bayes estimator:

$$r(\Pi, \hat{\theta}) = \int_\Theta P_\theta \ell(d(\theta, \hat{\theta})) d\Pi(\theta) = \int_\Theta \int_{\mathcal{P}} \ell(d(\theta, \hat{\theta}(y))) dP_\theta(y) d\Pi(\theta)$$

$$= \int_\Theta \int_{\mathcal{P}} \ell(d(\theta, \hat{\theta}(y))) p_\theta(y) d\Pi(\theta) d\mu(y)$$

$$= \int_\Theta \left( \int_{\mathcal{P}} p_\theta(y) d\Pi(\theta) \right) \int_{\Theta} \ell(d(\theta, \hat{\theta}(y))) d\Pi(\theta | Y = y) d\mu(y).$$

where we use the Radon-Nikodym theorem (see theorem B.7), Fubini’s theorem
(see theorem B.6) and the definition of the posterior, c.f. (2.12). Using the prior
predictive distribution \((2.3)\), we rewrite the Bayesian risk function further:

\[
  r(\Pi, \tilde{\theta}) = \int_{\Theta} \int_{\Omega} \ell(d(\theta, \tilde{\theta}(y))) d\Pi(\theta | Y = y) dP^\Pi(y). \tag{2.18}
\]

By assumption, the formal Bayes estimator \(\tilde{\theta}\) exists. Since \(\tilde{\theta}\) satisfies

\[
  \int_{\Theta} \ell(d(\theta, \tilde{\theta}(y))) d\Pi(\theta | Y = y) = \inf_{\theta' \in \Theta} \int_{\Theta} \ell(d(\theta, \theta')) d\Pi(\theta | Y = y)
\]

for \(P^\Pi\)-almost all \(y \in \mathcal{Y}\), we obtain

\[
  r(\Pi, \tilde{\theta}) = \int_{\mathcal{Y}} \inf_{\theta' \in \Theta} \int_{\Theta} \ell(d(\theta, \theta')) d\Pi(\theta | Y = y) dP^\Pi(y) \leq \inf_{\theta' \in \Theta} \int_{\mathcal{Y}} \ell(d(\theta, \theta')) d\Pi(\theta | Y = y) dP^\Pi(y) = \inf_{\theta' \in \Theta} r(\Pi, \theta').
\]

One estimator of this type is defined in the following intuitively reasonable way.

**Definition 2.11.** Let the data \(Y\) with model \(\mathcal{P}\), metric \(d\) and prior \(\Pi\) be given. Suppose that the \(\sigma\)-algebra on which \(\Pi\) is defined contains the Borel \(\sigma\)-algebra. For given \(\epsilon > 0\), the small-ball estimator is defined to be the maximizer \(\tilde{P}\) of the function

\[
  P \mapsto \Pi(B_d(P, \epsilon) | Y),
\]

over the model, where \(B_d(P, \epsilon)\) is the \(d\)-ball in \(\mathcal{P}\) of radius \(\epsilon\) centred on \(P\). Provided that such a maximizer exists and is unique, it is defined almost-surely.

Note that this is simply the formal Bayes estimator for the loss function \(\ell(d) = 1\{d \geq \epsilon\}\). Existence of a small-ball estimator \(\tilde{P}\) therefore implies optimality in the sense that,

\[
  \Pi(d(P, \tilde{P}) \geq \epsilon | Y) = \inf_{Q \in \mathcal{P}} \Pi(d(P, Q) \geq \epsilon | Y).
\]

**Remark 2.5.** Similarly to definition 2.11, for a fixed value \(p\) such that \(1/2 < p < 1\), we may define a Bayesian point estimator as the centre point of the smallest \(d\)-ball with posterior mass greater than or equal to \(p\) (if it exists and is unique).

Suitable conditions for the existence of small-ball estimators form the subject of exercise 2.12.

### 2.2.3 The maximum-a-posteriori estimator

If the posterior is dominated by a \(\sigma\)-finite measure \(\nu\), the posterior density with respect to \(\nu\) can be used as a basis for defining Bayesian point estimators.

**Definition 2.12.** Let \(\mathcal{P}\) be a model with prior \(\Pi\). Assume that the posterior is absolutely continuous with respect to a \(\sigma\)-finite measure \(\nu\) on \(\mathcal{P}\), with \(\nu\)-density
2.2 Bayesian point estimators

\( \theta \mapsto \pi(\theta | Y) \). The maximum-a-posteriori estimator (or MAP-estimator, or posterior mode) \( \hat{\theta}_2 \) for \( \theta \) is defined as the point in the model where the posterior density takes on its maximal value:

\[
\pi(\hat{\theta}_2 | Y) = \sup_{\theta \in \Theta} \pi(\theta | Y).
\] (2.20)

Provided that such a point exists and is unique, the MAP-estimator is defined almost-surely.

The MAP-estimator has a serious weak point: a different choice of dominating measure \( \nu \) leads to a different MAP estimator! A MAP-estimator is therefore not fully specified unless we indicate which dominating measure was used to define the posterior density. It is with respect to this dominating measure that we define our estimator, so a motivation for the dominating measure used is inherently necessary. Often the Lebesgue measure is used without further comment, or objective measures (see section 3.2) are used. Another option is to use the prior measure as the dominating measure, in which case the MAP estimator equals the maximum-likelihood estimator.

**Remark 2.6.** There is an interesting connection between (Bayesian) MAP-estimation and (frequentist) maximum-likelihood estimation. Referring to formula (2.12) we see that in an i.i.d. experiment with parametric model, the MAP-estimator maximizes:

\[
\Theta \to \mathbb{R} : \theta \mapsto \prod_{i=1}^{n} p_{\theta}(X_i) \pi(\theta),
\]

where it is assumed that the model is dominated and that the prior has a density \( \pi \) with respect to the Lebesgue measure \( \nu \). If the prior had been uniform, the last factor would have dropped out and maximization of the posterior density is maximization of the likelihood. Therefore, differences between ML and MAP estimators are entirely due to non-uniformity of the prior. Subjectivist interpretation aside, prior non-uniformity has an interpretation in the frequentist setting as well, through what is called penalized maximum likelihood estimation (see, for example, van de Geer (2000) [98]): Bayes’s rule applied to the posterior density \( \pi_n(\theta | X_1, \ldots, X_n) \) gives:

\[
\log \pi_n(\theta | X_1, \ldots, X_n) = \log \prod_{i=1}^{n} p_{\theta}(X_i) + \log \pi(\theta) + D(X_1, \ldots, X_n),
\]

where \( D \) is a (\( \theta \)-independent) normalization constant. The first term equals the log-likelihood and the logarithm of the prior plays the role of a penalty term when maximizing over \( \theta \). Hence, maximizing the posterior density over the model \( \Theta \) can be identified with maximization of a penalized likelihood over \( \Theta \). So defining a penalized MLE \( \hat{\theta}_n \) with the logarithm of the prior density \( \theta \mapsto \log \pi(\theta) \) in the role of the penalty, the MAP-estimator coincides with \( \hat{\theta}_n \). The above offers a direct connection between Bayesian and frequentist methods of point-estimation. As such, it provides an frequentist interpretation of the prior as a penalty in the ML procedure.
All Bayesian point estimators defined above as maximizers or minimizers over the model suffer from the usual existence and uniqueness issues associated with extrema. However, there are straightforward methods to overcome such issues. We illustrate using the MAP-estimator. Questions concerning the existence and uniqueness of MAP-estimators should be compared to those of the existence and uniqueness of \(M\)-estimators in frequentist statistics. Although it is hard to formulate conditions of a general nature to guarantee that the MAP-estimator exists, often one can use the following lemma to guarantee existence.

**Lemma 2.3.** Consider a parameterized model \(\Theta \to P : \theta \mapsto P_\theta\); If the parameter space \(\Theta\) is compact and the posterior density \(\theta \mapsto \pi(\theta|Y)\) is upper-semicontinuous, then the MAP-estimator exists almost surely.

To prove uniqueness one has to be aware of various possible problems among which, for instance, identifiability of the model (see section 1.1, in particular definition 1.5).

**Example 2.9.** Assuming that \(P_0(X \geq 0) = 1\), the posterior density in example 2.6 has the form:

\[
\pi(\theta|X_1, \ldots, X_n) = \frac{1}{n!} \left(1 + \sum_{i=1}^{n} X_i\right)^{n+1} \theta^n e^{-\theta(1+\sum_{i=1}^{n} X_i)},
\]

\(P_0^n\)-almost-surely, where \(\theta > 0\). Setting the \(\theta\)-derivative to zero, we find that the MAP-estimator is given by:

\[
\hat{\theta}_2(Y) = n \left(1 + \sum_{i=1}^{n} X_i\right)^{-1}.
\]

The MAP-estimator is similar to the maximum likelihood estimator (equal to \(n(\Sigma_iX_i)^{-1}\)) and the posterior mean (equal to \((n+1)(1+\Sigma_iX_i)^{-1}\), see 2.17)). Although it is possible technically that these three estimators differ substantially, in many (e.g. unimodal or convergent) cases the maximum of the posterior density lies in the bulk that determines the posterior mean as well, and MAP and posterior mean are close. If, in addition, the influence of the prior is relatively small because the likelihood function peaks very sharply at its maximum, the maximum-likelihood estimator is expected to be close too. Note that differences between these three estimators become negligible in the limit \(n \to \infty\).

### 2.3 Confidence sets and credible sets

Besides point-estimation, frequentist statistics has several other inferential techniques at its disposal. The two most prominent are the analysis of confidence intervals and the testing of statistical hypotheses. In the next section, we consider frequentist testing of hypotheses; in this section, we discuss frequentist confidence sets and their Bayesian counterparts, called credible sets.
2.3 Confidence sets and credible sets

2.3.1 Frequentist confidence intervals

Assume that we have a model $\mathcal{P}$ parametrized by an identifiable parameter $\theta$ in a parameter set $\Theta$, assuming that the true distribution of the data $Y \sim P_0$ belongs to the model, that is, $P_0 = P_{\theta_0}$ for some $\theta_0 \in \Theta$. The inferential goal is to use the data $Y$ to define a model subset $C(Y)$ that contains $\theta_0$ with ‘high’ probability. ‘High’ probability requires quantification in terms of a level $\alpha$, called the confidence level.

**Definition 2.13.** Let $\Theta \rightarrow \mathcal{P} : \theta \mapsto P_{\theta}$ be an identifiable parametrization and assume that $Y \sim P_{\theta_0}$ for some $\theta_0 \in \Theta$. Choose a confidence level $\alpha \in (0, 1)$. Let $C_{\alpha}(Y)$ be a subset of $\Theta$ dependent only on the data $Y$. Then $C_{\alpha}$ is a confidence set for $\theta$ of confidence level $\alpha$, if it solves the equation,

$$P_{\theta_0}(\theta_0 \in C_{\alpha}) \geq 1 - \alpha.$$  (2.21)

The condition that $C_{\alpha}$ be dependent on the data $Y$ is meant to express the requirement that $C_{\alpha}$ can be calculated once the data has been observed. The confidence set may also depend on other quantities that are known to the statistician, so $C_{\alpha}$ is a statistic (see definition 1.9). The dependence of $C_{\alpha}$ on the data $Y$ makes $C_{\alpha}$ a random subset of the model. (Compare this to point estimators which are random points in the model.) The measurability of the event $\{Y : \theta_0 \in C_{\alpha}(Y)\}$ is a non-trivial technical matter that has to be addressed at the model-specific level.

Clearly confidence sets are not unique and small confidence sets are more informative than large ones: if, for some confidence level $\alpha$, we have two different procedures of finding confidence sets, leading to sets $C_{\alpha}$ and $D_{\alpha}$ of confidence level $\alpha$ respectively, and $C_{\alpha} \subset D_{\alpha}$, $P_{\theta}$-almost-surely for all $\theta$, then $C_{\alpha}$ is preferred over $D_{\alpha}$.

**Example 2.10.** Let $X^n = (X_1, \ldots, X_n)$ be an i.i.d. sample from a normal distribution $N(\mu_0, \sigma^2)$, with known variance $\sigma^2 > 0$ and unknown $\mu_0 \in \mathbb{R}$. As is well-known, the sample average is normally distributed,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N(\mu_0, \sigma^2/n),$$

with a variance $\sigma^2_n = \sigma^2/n$. We rephrase as follows:

$$P_{\theta_0}(\frac{\bar{X} - \mu_0}{\sigma_n} \leq x) = \Phi(x),$$

for all $x \in \mathbb{R}$, where $\Phi$ denotes the distribution function of the standard normal distribution. Consequently,

$$P_{\theta_0}(\frac{\bar{X} - \sigma x}{\sqrt{n}} \leq \mu_0 \leq \frac{\bar{X} + \sigma x}{\sqrt{n}}) = \Phi(x) - \Phi(-x).$$

Fixing some confidence level $\alpha > 0$, we solve for $x_{\alpha/2}$ in the equation $\Phi(x_{\alpha/2}) - \Phi(-x_{\alpha/2}) = 1 - \alpha$ to arrive at the conclusion that the interval,
\[ C_\alpha = \left[ \overline{X}_n - \frac{\sigma \sqrt{\alpha / 2}}{\sqrt{n}}, \overline{X}_n + \frac{\sigma \sqrt{\alpha / 2}}{\sqrt{n}} \right] \]

is a level-\(\alpha\) confidence set for the parameter \(\mu\).

Like in the example 2.10, confidence intervals for a parameter \(\theta\) are often derived from estimators for \(\theta\): in the case of example 2.10 the sample average estimates \(\mu\) and it is the distribution of the sample average around \(\mu\) that determines the confidence interval.

**Definition 2.14.** Let \(\Theta \rightarrow \mathcal{P} : \theta \mapsto P_\theta\) be an identifiable model and assume that \(Y \sim P_{\theta_0}\) for some \(\theta_0 \in \Theta\). Let \(\hat{\theta} : \mathcal{Y} \rightarrow \Theta\) be an estimator for \(\theta\). The distribution of \(\hat{\theta}\) over \(\Theta\) is called the *sampling distribution* of the estimator.

Note that the randomness in \(\hat{\theta}\) occurs because it is a function of the data, which is distributed according to \(P_{\theta_0}\) itself. Therefore the sampling distribution of \(\hat{\theta}\) depends on the parameter \(\theta_0\); in fact, it is exactly this dependence that allows us to draw statistical conclusions. In example 2.10, the location of the sampling distribution is \(\mu_0\), so it may be assumed that the sample average \(\bar{X}_n\) lies close to its expectation, the true mean \(\mu_0\). Exactly how close is quantified by the variance of the sampling distribution of \(\bar{X}_n\). A confidence set is an inferential instrument to express how much uncertainty remains concerning the true value of a parameter after estimation.

In case we have a very large sample, the analysis is sometimes more conveniently done based on limits, in particular the weak limit in the central limit theorem. To that end, we also define sequences of confidence sets that reach the required confidence level only in the limit \(n \to \infty\).

**Definition 2.15.** For every \(n \geq 1\), let \(X^n\) be data taking values in sample spaces \((\mathcal{Y}_n, \mathcal{B}_n)\), with a model \(\mathcal{P}_n\) and identifiable parametrization \(\Theta \rightarrow \mathcal{P}_n : \theta \mapsto P_{\theta,n}\). Assume that for some \(\theta_0 \in \Theta\), \(X^n \sim P_{\theta_0,n}\) for all \(n \geq 1\), and choose a confidence level \(\alpha \in (0,1)\). A sequence \((C_{\alpha,n})\) of sets such that \(C_{\alpha,n}\) depends only on \(X^n\), that solves,

\[ \liminf_{n \to \infty} P_{\theta_0,n}(\theta_0 \in C_{\alpha,n}(X^n)) \geq 1 - \alpha, \quad (2.23) \]

is called an *asymptotic confidence set* of level \(\alpha\). In case \(\alpha = 0\), \(\lim_n P_{\theta_0,n}(\theta_0 \in C_{0,n}) = 1\) and we say that the confidence sets \(C_{0,n}\) are (*asymptotically*) *consistent*.

As it turns out, example 2.10 applies in an approximate form in all asymptotic cases where the central limit theorem applies, as the following example demonstrates.

**Example 2.11.** Let \(\Theta \rightarrow \mathcal{P} : \theta \mapsto P_\theta\) be an identifiable, parametric model for measurements \(X\) and assume that \(X^n = (X_1, \ldots, X_n)\) is an *i.i.d.* sample from a distribution \(P_{\theta_0}\) for some \(\theta_0 \in \Theta\). Suppose that there exists a measurable function \(f : \mathbb{R} \rightarrow \mathbb{R}\), such that \(P_{\theta_0} f(X) = \theta\) and \(P_{\theta_0} f(X)^2 < \infty\) for all \(\theta \in \Theta\). Moreover, we assume that for some known constant \(S > 0\), \(\sigma^2(\theta) = P_{\theta}(f(X) - \theta)^2 \leq S^2\), for all \(\theta \in \Theta\). Consider the sample-average \(\bar{X}_n = \overline{X}_n f(X)\). According to the central limit theorem, estimators that are sample averages for such \(f\) have sampling distributions that converge weakly to normal distributions. Choose a confidence level \(\alpha \in (0,1)\) and note that,
2.3 Confidence sets and credible sets

\[ P_{\hat{\theta}_n} \left( \frac{\sigma(\theta_0) x_{\alpha/2}}{\sqrt{n}} < \hat{\theta}_n - \theta_0 \leq \frac{\sigma(\theta_0) x_{\alpha/2}}{\sqrt{n}} \right) \to 1 - \alpha, \quad (2.24) \]

as \( n \to \infty \). Define \( C_{\alpha,n} \) by

\[ C_{\alpha,n} = \left[ \hat{\theta}_n - \frac{S x_{\alpha/2}}{\sqrt{n}}, \hat{\theta}_n + \frac{S x_{\alpha/2}}{\sqrt{n}} \right]. \]

Then \( P_{\hat{\theta}_n}(\theta_0 \in C_{\alpha,n}) \to 1 - \alpha \), so for any \( \alpha' > \alpha \), \( P_{\hat{\theta}_n}(\theta_0 \in C_{\alpha,n}) \geq 1 - \alpha' \) if \( n \) is large enough. Note that if we had not used \( S \) but \( \sigma(\theta_0) \) instead, the \( \theta_0 \)-dependence of \( \sigma(\theta_0) \) would violate the requirement that \( C_{\alpha,n} \) be a statistic: since the true value \( \theta_0 \) of \( \theta \) is unknown, so is \( \sigma(\theta_0) \). Substituting the (known) upper-bound \( S \) for \( \sigma(\theta_0) \) enlarges the \( \sigma(\theta_0) \)-interval that follows from (2.24), while eliminating the \( \theta_0 \)-dependence. In a practical situation one would not assume that there is some upper bound \( S > 0 \), but substitute \( \sigma(\theta_0) \) by an estimator \( \hat{\sigma}_n \) (a practice known as studentization, after the Student \( t \)-distribution one obtains upon plugging in \( \hat{\sigma} \) for finite \( n \geq 1 \) with \( X \) that are marginally normal; refer to the case of example 2.10 if \( \sigma^2 \) had not been known.) Since the asymptotics of the studentized version are equal to those of the version based on \( \sigma(\theta_0) \), studentization does not change the conclusions we based on (2.24), that is,

\[ C_{\alpha',n} = \left[ \hat{\theta}_n - \frac{\hat{\sigma}_n x_{\alpha/2}}{\sqrt{n}}, \hat{\theta}_n + \frac{\hat{\sigma}_n x_{\alpha/2}}{\sqrt{n}} \right], \]

are asymptotic confidence intervals of any level \( \alpha' > \alpha \).

2.3.2 Bayesian credible sets

The Bayesian analogs of confidence sets are called credible sets and are derived from the posterior distribution. The rationale behind the definition of credible sets is exactly the same one that motivated confidence sets: we look for a subset \( D \) of the model that is as small as possible while receiving a certain minimal probability. Presently, however, the word ‘probability’ is not based on the sampling distribution of an estimator, but in line with the Bayesian notion, that is, ‘probability’ according to the posterior distribution.

**Definition 2.16.** Let \((\Theta, \mathcal{G})\) be a measurable space parameterizing an identifiable model \( \Theta \to \mathcal{G} : \theta \mapsto P_\theta \) for data \( Y \), with prior \( \Pi : \mathcal{G} \to [0, 1] \). Choose a level \( \alpha \in (0, 1) \). Let \( D_\alpha \in \mathcal{G} \) be a subset of \( \Theta \). Then \( D_\alpha \) is a level-\( \alpha \) credible set for \( \theta \), if it solves the equation,

\[ \Pi \left( \theta \in D_\alpha \mid Y \right) \geq 1 - \alpha, \quad (2.25) \]

almost-surely.

To find credible sets in examples one starts by calculating the posterior distribution from the prior and the data and, based on that, derives a subset \( D_\alpha(Y) \) such that
(2.25) is satisfied. From a frequentist perspective, credible sets are statistics since they are defined based entirely on the posterior (which is a statistic itself). A credible set is sometimes referred to as a credible region, or, if \(D\) is an interval in a one-dimensional parameter space, a credible interval. We can extend this definition to the

**Definition 2.17.** For every \(n \geq 1\), let \(X^n\) be data taking values in sample spaces \((\mathcal{X}_n, \mathcal{B}_n)\), with a model \(\mathcal{P}_n\) and identifiable parametrization \(\Theta \rightarrow \mathcal{P}_n: \theta \mapsto P_{\theta,n}\). Choose a sequence of credible levels \((\alpha_n)\), \(\alpha \in (0, 1)\), \(\alpha_n \downarrow 0\). A sequence of measurable subsets \((D_{\alpha,n})\) that solves,

\[
\Pi(\theta \in D_{\alpha,n} \mid X^n) = o(\alpha_n),
\]

almost-surely, is called a sequence of asymptotic credible sets of levels \((\alpha_n)\).

**Remark 2.7.** In smooth, parametric models for i.i.d. data there is a close, asymptotic relation between Bayesian credible sets and frequentist confidence intervals centred on the maximum-likelihood estimator: the Bernstein-von Mises theorem implies that level-\(\alpha\) credible sets coincide with abovementioned level-\(\alpha\) confidence sets asymptotically! In situations where it is hard to calculate the ML estimator or to construct the corresponding confidence interval explicitly, it is sometimes relatively easy to obtain credible sets (based on a simulated sample from the posterior, as obtained from the MCMC procedure, for example). In such cases, one can calculate credible sets and conveniently interpret them as confidence intervals centred on the MLE, due to the Bernstein-von Mises theorem.

Definition 2.16 suffices to capture the concept of a credible set, but offers too much freedom in the choice of \(D\): given a level \(\alpha > 0\), many sets will satisfy (2.25), just like confidence sets can be chosen in many different ways. Note that, also here, we prefer smaller sets over large ones: if, for some level \(\alpha\), two different level-\(\alpha\) credible sets \(F_{\alpha}\) and \(G_{\alpha}\) are given, both satisfying (2.25) and \(F_{\alpha} \subset G_{\alpha}\) then \(F_{\alpha}\) is preferred over \(G_{\alpha}\). If the posterior is dominated with density \(\theta \mapsto \pi(\theta | Y)\), we can be more specific. We define, for every \(k \geq 0\), the level-sets,

\[
D(k) = \{ \theta \in \Theta : \pi(\theta | Y) \geq k \},
\]

and consider the following.

**Definition 2.18.** Let \((\Theta, \mathcal{G})\) a measurable space parameterizing a model \(\Theta \rightarrow \mathcal{G}: \theta \mapsto P_\theta\) for data \(Y \in \mathcal{Y}\), with prior \(\Pi : \mathcal{G} \rightarrow [0, 1]\). Assume that the posterior is dominated by a \(\sigma\)-finite measure \(\mu\) on \((\Theta, \mathcal{G})\), with density \(\pi(\cdot | Y) : \Theta \rightarrow \mathbb{R}\). Choose \(\alpha \in (0, 1)\). A level-\(\alpha\) HPD-credible set (from highest posterior density) for \(\theta\) is the subset \(D_{\alpha} = D(k_{\alpha})\), where \(k_{\alpha}\) equals:

\[
k_{\alpha} = \sup \{ k \geq 0 : \Pi(\theta \in D(k) | Y) \geq 1 - \alpha \}.
\]

In other words, \(D_{\alpha}\) is the smallest level-set of the posterior density that receives posterior mass greater than or equal to \(1 - \alpha\). Note that HPD-credible sets depend
on the choice of dominating measure: if we had chosen to use a different measure \( \mu \), HPD-credible sets would have changed as well!

## 2.4 Tests and Bayes factors

Having discussed confidence sets and credible sets in the previous section, we now turn to the related subject of hypothesis testing. We start with a discussion of the Neyman-Pearson framework and the famous lemma concerning the optimal test for testing one distribution versus another. Next we consider tests of uniform testing power and minimax optimality, as well as asymptotic testing. In the last subsection we consider posterior odds and Bayes factors, as well as Bayesian test functions.

### 2.4.1 Neyman-Pearson tests

Assume that we have data \( Y \) and a parameterized model \( \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \) such that \( Y \sim P_{\theta_0} \) for some \( \theta_0 \in \Theta \). For simplicity, we assume that \( \Theta \subset \mathbb{R} \) whenever the dimension of \( \Theta \) is of importance. In statistical testing the hypotheses are mutually exclusive speculations concerning the distribution of the data. Since the model contains all distributions the frequentist is willing to consider as candidates for \( P_0 \), the hypotheses are formulated in terms of a partition of the model (or its parameterization) into two disjoint subsets.

**Definition 2.19.** Testing of hypotheses proceeds through choice of a model subset \( \Theta_0 \) corresponding to the so-called **null hypothesis** \( H_0 \) and its complement \( \Theta_1 = \Theta \setminus \Theta_0 \), called the **alternative hypothesis** \( H_1 \). We distinguish between so-called **simple** hypotheses which consist of a single point in \( \Theta \) and **composite** hypotheses which consist of bigger subsets in \( \Theta \).

In a setting where \( H_0 \) and \( H_1 \) are treated on equal footing, the statistician is asked to come up with a method to decide between these hypotheses (or to conclude that the data does not have enough evidence to decide either way).

**Example 2.12.** As an example consider an \( i.i.d. \) real-valued sample \( X^n = (X_1, \ldots, X_n) \) from a distribution \( P_0 \) of known form \( Q \), but unknown location \( \theta \). Then the model is a location family: \( P_\theta(A) = Q(A - \theta) \) for all measurable \( A \) and \( \theta \in \Theta \subset \mathbb{R} \). Let \( \theta' \in \Theta \) be a chosen point in the model and consider the (simple) null-hypothesis \( H_0 \) and alternative \( H_1 \) concerning the true value \( \theta_0 \) of the parameter:

\[
H_0 : \quad \theta_0 = \theta', \quad H_1 : \quad \theta_0 \in \Theta \setminus \{ \theta' \}, \quad (2.28)
\]

The procedure now requires that we use the data \( X^n \) to investigate the hypotheses further. A good starting point is a location estimator \( \hat{\theta}_n \) (or test statistic), and more particularly, its sampling distribution under \( P^n_0 \). Fix \( n \geq 1 \) and some \( \alpha \in (0, 1) \) and
suppose that for a certain, known value \( k_n \in \mathbb{R} \),

\[
P_{\theta'}^n(\theta' - k_n \leq \hat{\theta}_n \leq \theta' + k_n) \geq 1 - \alpha.
\]

Then it is so unlikely to find \( \hat{\theta} \) further than \( k_n \) away from \( \theta' \) under the assumption that \( H_0 \) is true, that we would like to reject \( H_0 \) if the data \( X^n \) is such that \( |\hat{\theta}_n(X^n) - \theta'| > k_n \). Of course this is too crude because we have not compared with probabilities to observe differences larger than \( k_n \) under alternative values of \( \theta \).

A warning is in order because of the following conceptual confusion: the null and alternative hypotheses do not play symmetric roles in the Neyman-Pearson procedure. The goal of Neyman-Pearson hypothesis testing is to find out whether or not the data contains “enough” evidence to reject the null hypothesis as a likely explanation when compared to explanations offered by the alternative. By no means does Neyman-Pearson testing equate to the symmetric testing problem, in which \( H_0 \) and \( H_1 \) play interchangeable roles and we make a choice for one or the other based on the data. (By contrast, in subsections 2.4.3 and 2.4.4 we do consider symmetric testing.)

As indicated above, one usually departs from a test statistic \( T(Y) \in \mathbb{R}^d \), displaying different behaviour depending on whether the data \( Y \) is distributed according to distributions in \( H_0 \) or \( H_1 \). One also chooses a so-called critical set \( K \subset \mathbb{R}^d \) such that \( P_\theta(T \in K) \) is “small” for all \( \theta \in \Theta_0 \) and \( P_\theta(T \notin K) \) is “small” for all \( \theta \in \Theta_1 \). What we mean by “small” probabilities in this context is a choice, the so-called significance level \( \alpha \in (0,1) \).

**Definition 2.20.** Let \( \Theta \rightarrow \mathcal{F} : \theta \rightarrow P_\theta \) be an identifiable, parametrized model for a sample \( Y \). Formulate two hypotheses \( H_0 \) and \( H_1 \) by introducing a two-set partition \( \{\Theta_0, \Theta_1\} \) of the model \( \Theta \):

\[
H_0 : \quad \theta_0 \in \Theta_0, \quad H_1 : \quad \theta_0 \in \Theta_1.
\]

We say that a test for these hypotheses based on a test-statistic \( T \) with critical set \( K \) is of level \( \alpha \in (0,1) \) if the power function \( \pi : \Theta \rightarrow [0,1] \), defined by

\[
\pi(\theta) = P_\theta(T(Y) \in K),
\]

is uniformly small over \( \Theta_0 \):

\[
\sup_{\theta \in \Theta_0} \pi(\theta) \leq \alpha. \tag{2.29}
\]

From the above definition we arrive at the conclusion that if \( Y = y \) and \( T(y) \in K \), hypothesis \( H_0 \) is improbable enough to be rejected, since \( H_0 \) forms an “unlikely” explanation of observed data (at said significance level). The degree of “unlikeliness” can be quantified in terms of the so-called p-value, which is the lowest significance level at which the realised value of the test statistic \( T(y) \) would have led us to reject \( H_0 \).

Of course there is the possibility that our decision is wrong and \( H_0 \) is actually true but \( T(y) \in K \) nevertheless, so that our rejection of the null hypothesis is unwar-
ranted. This is called a **type-I error**; a **type-II error** is made when we do not reject $H_0$ while $H_0$ is not true. The significance level $\alpha$ represents a fixed upper bound for the probability of a type-I error, c.f. (2.29). If we have several possible tests, all of which satisfy the Type-I error bound (2.29), the Neyman-Pearson approach calls for minimization of the Type-II error probability: of all the pairs $(T, K)$ satisfying (2.29), one prefers a pair that minimizes $P_\theta(T(Y) \notin K)$, (ideally) for all $\theta \in \Theta_1$. However, generically such uniformly most-powerful tests do not exist due to the possibility that different $(T, K)$ pairs satisfying (2.29) are most powerful over distinct subsets of the alternative.

We consider the Neyman-Pearson approach to testing in some more detail in the following example in the context of normally distributed data.

**Example 2.13.** Denote by $P_\mu$ the normal distribution $N(\mu, \sigma^2)$ with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Let $Y_n = (X_1, \ldots, X_n)$ be an i.i.d. sample from a normal distribution $P_0 = P_{\mu_0}$, assuming that the variance $\sigma^2 > 0$ is known, and unknown $\mu_0 \in \mathbb{R}$. As we have seen in (2.22), the sample average is normally distributed, $\hat{\mu}_n = \frac{1}{n} \sum X_i \sim N(\mu_0, \sigma^2/n)$. For any $\mu$, we have,

$$P_\mu\left(\sqrt{n}(\hat{\mu}_n - \mu) \leq x\right) = \Phi(x),$$

for all $x \in \mathbb{R}$, where $\Phi$ denotes the distribution function of the standard normal distribution. Re-write to obtain,

$$P_\mu\left(\frac{x \sigma}{\sqrt{n}} < \hat{\mu}_n - \mu \leq \frac{x \sigma}{\sqrt{n}}\right) = \Phi(x) - \Phi(-x),$$

for $x > 0$. By choosing some $\bar{\mu}$ and a significance level $\alpha$, we formulate null- and alternative hypotheses as in (2.28),

$$H_0: \quad \mu_0 = \bar{\mu}, \quad H_1: \quad \mu_0 \neq \bar{\mu},$$

and note that under the null-hypothesis,

$$P_\mu\left(\sqrt{n}(\hat{\mu}_n - \mu_0) \leq x_{\alpha/2}\right) = 1 - \alpha,$$

if we choose the quantiles $x_{\alpha/2}$ like in example 2.10. Hence the null-hypothesis makes it improbable to observe $|\hat{\mu}_n - \bar{\mu}| > n^{-1/2}\sigma x_{\alpha/2}$, which gives rise to the following definition of the critical set $K_\alpha$,

$$K_\alpha = \mathbb{R} \setminus \left[\bar{\mu} - \frac{x_{\alpha/2}\sigma}{\sqrt{n}}, \bar{\mu} + \frac{x_{\alpha/2}\sigma}{\sqrt{n}}\right],$$

which enables us to formulate our decision on the null hypothesis, that is, to reject $H_0$ or not:

(i) if the realised $\hat{\mu}_n \in K_\alpha$, we reject $H_0$ at significance level $\alpha$, and,
(ii) if the realised \( \hat{\mu}_n \not\in K_\alpha \), we do not see enough evidence in the data to reject \( H_0 \) at significance level \( \alpha \).

Beware of a common pitfall regarding the interpretation: under case (ii), we do not draw the conclusion that \( H_0 \) is accepted. The data does not provide enough evidence to reject the null hypothesis but that does not imply that we accept it.

Note the behaviour of the procedure with varying sample-size: keeping the significance level fixed, the width of the critical sets \( K_\alpha \) is of order \( O(n^{-1/2}) \), so smaller and smaller critical sets can be used as more information concerning the distribution \( P_0 \) (read, data) comes available. Conversely, if we keep the critical set fixed, the probability for a Type-I error decreases (exponentially) with growing sample-size. Analogous to example 2.11, it is common practice to use asymptotic tests because sampling distributions are rarely available in closed form: in cases where the sample-size is deemed to be large enough, asymptotic approximations to sampling distributions can be used (e.g. the central limit theorem when estimators are based on sample averages).

### 2.4.2 Randomized tests and the Neyman-Pearson lemma

According to the Neyman-Pearson approach, tests are only considered if they satisfy (2.29) and are optimal if, in addition, they maximize testing power over the alternative uniformly. Optimality in this sense is sometimes not achievable, so one wonders if, mathematically, anything can be said at all. The Neyman-Pearson lemma answers this question in the affirmative for problems where the null- and alternative hypotheses each contain a single distribution.

To formulate the Neyman-Pearson lemma, however, we have to generalize the testing procedure slightly: as it turns out the existence of an optimal test can only be guaranteed if we allow for a randomization of our decision.

**Definition 2.21.** Let \( \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \) be a parametrized model for data \( Y \). Formulate two hypotheses \( H_0 \) and \( H_1 \) for \( \theta \) based on the two-set partition \( \{ \Theta_0, \Theta_1 \} \) of the model \( \mathcal{P} \):

\[
H_0 : \quad \theta_0 \in \Theta_0, \quad H_1 : \quad \theta_0 \in \Theta_1.
\]

A test function \( \phi \) is a map \( \phi : \mathcal{Y} \to [0, 1] \) used to formulate the following procedure called a randomized test: given a realisation \( Y = y \) we reject \( H_0 \) with probability \( \phi(Y) \). The power function associated with the test function \( \phi \) is given by \( \pi : \Theta \to [0, 1] : \theta \mapsto P_\theta \phi \).

Note that if we use the test function \( \phi(Y) = 1\{ T(Y) \in K \} \), the randomized test reduces to the original (non-random) procedure of rejecting \( H_0 \) if \( T(Y) \in K \). Clearly the probability for Type-I error when using the randomized procedure equals \( \pi(\theta) \) (for \( \theta \in \Theta_0 \)) and the probability for Type-II error equals \( 1 - \pi(\theta) \) (for \( \theta \in \Theta_1 \)).

Now return to the question regarding optimal Neyman-Pearson tests: we saw that the existence of an optimal test cannot be guaranteed as long as we look at com-
2.4 Tests and Bayes factors

Positive hypotheses. However, if both null and alternative hypothesis are simple a
(randomized) optimal test exists by the famed Neyman-Pearson lemma [154].

Lemma 2.4. Suppose the model is \( \mathcal{P} = \{P_{\theta_0}, P_{\theta_1}\} \) and write \( p_{\theta_0} : \mathcal{Y} \mapsto \mathbb{R} \) and \( p_{\theta_1} : \mathcal{Y} \mapsto \mathbb{R} \) for the densities of \( P_{\theta_0} \) and \( P_{\theta_1} \) relative to some \( \sigma \)-finite measure \( \mu \). Choose a significance level \( \alpha \in (0, 1) \) and consider a test of the form,

\[
\phi(y) = \begin{cases} 1 & \text{if } p_{\theta_1}(y) > cp_{\theta_0}(y) \\ \gamma(x) & \text{if } p_{\theta_1}(y) = cp_{\theta_0}(y) \\ 0 & \text{if } p_{\theta_1}(y) < cp_{\theta_0}(y) \end{cases},
\]

(2.30)

where the measurable function \( \gamma : \mathcal{Y} \to [0, 1] \) and the constant \( c \in [0, \infty] \) form a solution to the equation:

\[
P_{\theta_0}\phi = \alpha.
\]

The following two assertions concern the hypotheses,

\[
H_0 : \theta = \theta_0, \quad H_1 : \theta = \theta_1.
\]

(i.) If a test of the form (2.30) has significance level \( \alpha \) then it is most powerful among all tests of level \( \alpha \).

(ii.) If a test \( \phi' \) is most powerful, then \( \phi' \) is of the form (2.30) for some \( \gamma(x) \) and some \( c \), almost-surely with respect to both \( P_{\theta_0} \) and \( P_{\theta_1} \).

Proof. See Lehmann and Cassela (2005) [154].

The lemma is often used in conjunction with some condition on the model (or its likelihood function) to extend this point-vs-point version to composite hypotheses which are more interesting from a practical point of view.

Example 2.14. Suppose that we consider a random variable \( X \) drawn from a normal distribution \( N(\theta, 1) \) where \( \theta \in \Theta = \{-1, 1\} \). Fixing a significance level \( \alpha \in (0, 1) \), we consider a test of the form (2.30) for the hypotheses,

\[
H_0 : \theta = -1, \quad H_1 : \theta = +1.
\]

(2.31)

A simple calculation shows that \( p_{+1}(X)/p_{-1}(X) = e^{2X} \), so, with \( \Phi \) denoting the distribution function for the standard normal distribution,

\[
P_{-1}\Phi(X) = P_{-1}(p_{+1}(X) > cp_{-1}(X)) = P_{-1}(e^{2X} > c)
\]

\[=
P_{-1}(X > \frac{1}{2}\log c) = 1 - \Phi\left(\frac{1}{2}\log c + 1\right),
\]

where we have used that \( X \) is distributed continuously (so that the middle term in (2.30) does not play a role and any \( \gamma \) will do), and that \( P_{-1}(X \leq x) = \Phi(x + 1) \). So to find \( c \), we solve \( 1 - \Phi\left(\frac{1}{2}\log c + 1\right) = \alpha \), so that the Neyman-Pearson procedure for testing the hypotheses (2.31) has the form,

(i) if \( X > x_{1-\alpha} - 1 \), we reject \( H_0 \), and,
(ii) if \( X \leq x_{1-\alpha} - 1 \), we do not see enough evidence in the data to reject \( H_0 \), at significance level \( \alpha \).

### 2.4.3 Symmetric and asymptotic testing

Two points remain, the first being an asymptotic perspective on testing: just like we often study limits of sequences of estimators and conditions for their optimality, we are interested also in sequences of tests.

**Definition 2.22.** For data \( X^n \) taking values in measurable spaces \( (\mathcal{X}_n, \mathcal{B}_n) \), let \( \mathcal{P}_n \) be models parametrized identifiably on \( \Theta \) by \( \Theta \to \mathcal{P}_n : \theta \mapsto P_{\theta,n} \) for every \( n \geq 1 \). Assume that the true distribution of the data lies in the model, i.e. \( X^n \sim P_{\theta_0,n} \) for some \( \theta_0 \in \Theta \). Formulate two hypotheses \( H_0 \) and \( H_1 \) by introducing a two-set partition \( \{ \Theta_0, \Theta_1 \} \) of the parameter space \( \Theta \):

\[
H_0 : \theta_0 \in \Theta_0, \quad H_1 : \theta_0 \in \Theta_1.
\]

A test sequence \( (\phi_n)_{n \geq 1} \) consists of statistics \( \phi_n : \mathcal{X}_n \to [0,1] \), for all \( n \geq 1 \): having observed \( X^n \), we reject \( H_0 \) with probability \( \phi_n(X^n) \). The power sequence of the test sequence \( (\phi_n) \) is formed by the maps \( \pi_n : \Theta \to [0,1] \) defined by:

\[
\pi_n(\theta) = P_{\theta,n}(\phi_n),
\]

representing the \( P_{\theta,n} \)-probability of rejecting \( H_0 \).

The quality of the test sequence depends on the behaviour of the power sequence on \( \Theta_0 \) and \( \Theta_1 \). If, like in definition 2.20, we are interested exclusively in rejection of the null hypothesis, we could set a significance level \( \alpha \) to select only those test sequences that satisfy \( \sup_{\theta \in \Theta_0} \pi_n(P_{\theta}) \leq \alpha \). Subsequently, we prefer test sequences that have high power on the alternative in the limit \( n \to \infty \). It seems more natural to symmetrize the roles of null- and alternative hypotheses. In that case, the procedure simplifies: we reject \( H_0 \) and accept \( H_1 \) (resp. accept \( H_0 \) and reject \( H_1 \)) randomly with probability \( \phi_n(X^n) \) (resp. \( 1 - \phi_n(X^n) \)). We may then assess type-I and type-II error probabilities in the limit.

**Definition 2.23.** We say that a test sequence \( (\phi_n) \) for hypotheses \( \Theta_0, \Theta_1 \) is consistent, if type-I and type-II error probabilities go to zero,

\[
\pi_n(\theta) = P_{\theta,n}(\phi_n) \to 0, \quad 1 - \pi_n(\eta) = P_{\eta,n}(1 - \phi_n) \to 0,
\]

for all \( \theta \in \Theta_0 \) and \( \eta \in \Theta_1 \).

**Example 2.15.** Suppose that we are given two sequences \( (P_n) \) and \( (Q_n) \) of distributions for data \( X^n \) taking values in measurable spaces \( (\mathcal{X}_n, \mathcal{B}_n) \) for all \( n \geq 1 \). We hypothesise that either \( X^n \sim P_n \) or \( X^n \sim Q_n \) and wish to determine statistically which is
true. This is the setting of the Neyman-Pearson lemma, so a test based on likelihood ratios $dP_n/dQ_n$ seems reasonable (define $\mu_n = P_n + Q_n$ and write $p_n = dP_n/d\mu_n$ and $q_n = dQ_n/d\mu_n$ for the Radon-Nikodym derivatives):

$$
\phi_n(X^n) = 1\{p_n(X^n) < q_n(X^n)\}.
$$

Then,

$$
P_n\phi_n + Q_n(1 - \phi_n)
= \int_{\mathcal{X}^n} (p_n(x^n)1\{p_n(x^n) < q_n(x^n)\} + q_n(x^n)1\{p_n(x^n) \geq q_n(x^n)\}) d\mu_n(x^n)
\leq \int_{\mathcal{X}^n} \sqrt{p_n(x^n)q_n(x^n)} d\mu_n(x^n) = 1 - \frac{1}{2} \int_{\mathcal{X}^n} (\sqrt{p_n(x^n)} - \sqrt{q_n(x^n)})^2 d\mu_n(x^n)
\leq 1 - H^2(P_n, Q_n).
$$

So if the Hellinger distances $H(P_n, Q_n) \to 1$, a consistent test for $(P_n)$ versus $(Q_n)$ exists (namely the likelihood ratio test). This conclusion is very general and underlines the fundamental statistical role that the Hellinger metric plays. In case $X^n = (X_1, \ldots, X_n) \in \mathcal{X}^n$ represents an i.i.d. sample and the sequences are product measures $P_n = P^x$, $Q_n = Q^x$ for two distinct distributions $P, Q$ on $\mathcal{X}$, we arrive at,

$$
P^x\phi_n + Q^x(1 - \phi_n) \leq e^{-nH^2(P, Q)}.
$$

Depending on the subsets $\Theta_0$ and $\Theta_1$, there is a question whether a consistent test sequence for the pair exists or not. The answer in the case of i.i.d. sampling, which is given in part II, characterizes those $\Theta_0, \Theta_1$ that can be tested consistently, as precisely those subsets that can be written as countable unions of ‘closed’ sets (where the relevant model topology requires further discussion).

Given two test sequences, we may compare them through the limits of their type-I and type-II errors.

**Definition 2.24.** Let $(\phi_n)$ and $(\psi_n)$ be two test sequences for $\Theta_0$ versus $\Theta_1$. Let $\theta \in \Theta_0$ (resp. $\eta \in \Theta_1$) be given. We say that $(\phi_n)$ is asymptotically more powerful than $(\psi_n)$ at $\theta$ (resp. $\eta$), if,

$$
\lim_{n \to \infty} P_{\eta,n}\phi_n \leq \lim_{n \to \infty} P_{\theta,n}\psi_n, \quad \text{(resp. } \lim_{n \to \infty} P_{\eta,n}\phi_n \geq \lim_{n \to \infty} P_{\theta,n}\psi_n\text{)},
$$

then $(\phi_n)$ is said to be asymptotically more powerful than $(\psi_n)$ at $\theta$ (resp. $\eta$). If (2.32) holds for all points in $\Theta$, the test sequence $(\phi_n)$ is said to be uniformly asymptotically more powerful than $(\psi_n)$. If one can show that this holds for all test sequences $(\psi_n)$, then $(\phi_n)$ is said to be uniformly asymptotically most powerful.

This ordering of test sequences is not complete: it is quite possible that $(\phi_n)$ is asymptotically more powerful than $(\psi_n)$ on a subset of $\Theta$, whereas on its complement in $\Theta$, $(\psi_n)$ is asymptotically more powerful. As a result, the existence of uniformly asymptotically most powerful test sequences is problematic and no gen-
eralization of the Neyman-Pearson lemma exists for composite hypotheses, not even when required only asymptotically.

To counter such problems we can choose to evaluate testing power for a test sequence \((\phi_n)\) uniformly over \(\Theta_0\) and \(\Theta_1\), adding maximal type-I and -II error probabilities.

**Definition 2.25.** Given two disjoint model subsets \(\Theta_0, \Theta_1\), the **uniform testing power** \((\pi_n)\) for \(\Theta_0\) versus \(\Theta_1\) of a test sequence \((\phi_n)\) is given by,

\[
\pi_n = \sup_{\theta \in \Theta_0} P_{\theta,n}(1 - \phi_n), \quad \sup_{\eta \in \Theta_1} P_{\eta,n}(1 - \phi_n),
\]

Clearly there is a stronger, uniform version of consistency too, which incorporates a testing rate quite naturally.

**Definition 2.26.** Given a sequence \((a_n)\), \(a_n > 0, a_n \to 0\), we say that a test sequence for hypotheses \(\Theta_0, \Theta_1\) is **uniformly consistent at uniform testing rate** \(a_n\), if

\[
\sup_{\theta \in \Theta_0} P_{\theta,n}(1 - \phi_n) = o(a_n).
\]

A test sequence is simply **uniformly consistent** if it is uniformly consistent at some rate.

Again, depending on the subsets \(\Theta_0\) and \(\Theta_1\), there arises the question whether a uniformly consistent test sequence for the pair exists or not. The answer in the case of i.i.d. sampling, which is given in part II, characterizes those \(\Theta_0, \Theta_1\) that can be tested consistently, as precisely those subsets that can be **separated uniformly** (where the relevant model uniformity requires further discussion).

For fixed \(n \geq 1\), one wonders about the existence of a test of optimal uniform testing power, i.e. a test function \(\phi_n : \mathcal{X}_n \to [0, 1]\) such that,

\[
\sup_{\theta \in \Theta_0} P_{\theta,n}(1 - \phi_n) = \inf_{\psi} \left( \sup_{\theta \in \Theta_0} P_{\theta,n}(1 - \psi) + \sup_{\eta \in \Theta_1} P_{\eta,n}(\psi) \right),
\]

where the infimum runs over all measurable \(\psi : \mathcal{X}_n \to [0, 1]\). In that case, the test \((\phi_n)\) is said to be **minimax optimal** (and the test sequence is said to be minimax optimal if this holds for every \(n \geq 1\)). Such tests exist as a consequence of the so-called **minimax theorem** (see theorem 2.4 in the next section): under certain convexity, continuity and compactness conditions,

\[
\inf_{\psi} \sup_{\theta, \eta} P_{\theta,n}(1 - \psi) = \sup_{\theta, \eta} \inf_{\psi} (P_{\theta,n}(1 - \psi) + P_{\eta,n}(\psi)),
\]

and, furthermore, that there exists a test \((\phi_n)\) satisfying (2.34). The difference between left-hand and right-hand sides in (2.35) is that for the \((\theta, \eta) \in \Theta_0 \times \Theta_1\) in the supremum, the test \(\psi\) on the right-hand side can be chosen equal to the likelihood ratio test of example 2.15, with \(P_n = P_{\theta,n}\) and \(Q_n = P_{\eta,n}\). As a consequence, the uniform test \(\phi_n\) has the uniform testing power that is determined by (the supremum
over \((\theta, \eta)\) of the testing power for pointwise testing of example 2.15. Taking into account the convexity and compactness requirements of the minimax theorem (that we have neglected above), we arrive at the following useful proposition applicable under i.i.d. sampling.

**Lemma 2.5.** Let \(X^n = (X_1, \ldots, X_n)\) be an i.i.d. sample, \(X^n \sim P^n\) for some single-observation distribution \(P\) from a model \(\mathcal{P}\). Let \(\Theta_0, \Theta_1 \subset \mathcal{P}\) be model subsets with convex hulls \(K_0\) and \(K_1\) that are separated by non-zero Hellinger distance:

\[
H(K_0, K_1) = \inf_{P \in K_0} \inf_{Q \in K_1} H(P, Q) > 0.
\] (2.36)

Then, for every \(n \geq 1\), there exists a (minimax) test function \(\phi_n\) such that,

\[
\sup_{P \in \Theta_0} P^n(1 - \phi_n) + \sup_{Q \in \Theta_1} Q^n(1 - \phi_n) \leq e^{-nH(K_0, K_1)^2}.
\]

This proposition is proved in part II. Note that any Hellinger ball in \(\mathcal{P}\) is convex, so for hypotheses \(\Theta_0\) and \(\Theta_1\) that fit inside two disjoint Hellinger balls at non-zero Hellinger distance from one-another, a uniformly consistent test sequence of exponential uniform testing rate exists. This fact is used throughout statistics and in particular, applies in Bayesian limit theorems.

### 2.4.4 Posterior odds and Bayes factors

Bayesian hypothesis testing treats null and alternative hypotheses *symmetrically*. This poses an immediate conceptual difference with the most common frequentist methods of hypothesis testing, which are founded on the Neyman-Pearson perspective. It also leaves a lot of room for discussions involving both frequentist and Bayesian views to become confused (and heated) philosophical debates, in which neither side leaves room for the conceptual starting points of the other. Therefore any direct comparison between Bayesian and frequentist testing is difficult (see, however, [12]). In a frequentist analysis of Bayesian testing methods, true comparison is only possible with symmetric forms of frequentist testing.

In the Bayesian perspective, the subsets \(\Theta_0\) and \(\Theta_1\) of the parameter space have posterior and prior probabilities which are used directly to formulate the test: based on the proportions between those probabilities, we shall decide which hypothesis is the preferred one, based on the following definitions.

**Definition 2.27.** Let \((\Theta, \mathcal{G})\) a measurable space parameterizing a model \(\Theta \to \mathcal{P} : \theta \mapsto P_\theta\) for data \(Y \in \mathcal{Y}\), with prior \(\Pi : \mathcal{G} \to [0, 1]\). Let \(\{\Theta_0, \Theta_1\}\) be a measurable partition of \(\Theta\) such that \(\Pi(\Theta_0) > 0\) and \(\Pi(\Theta_1) > 0\). The *prior and posterior odds ratios* in favour of \(\Theta_0\) are defined by \(\Pi(\Theta_0) / \Pi(\Theta_1)\) and \(\Pi(\Theta_0 | Y) / \Pi(\Theta_1 | Y)\) respectively. The *Bayes factor* in favour of \(\Theta_0\) is defined to be
Bayesian basics

\[ B = \frac{\Pi(\Theta_0|Y) \Pi(\Theta_1)}{\Pi(\Theta_1|Y) \Pi(\Theta_0)}. \]

When doing Bayesian hypothesis testing, we have a choice of which ratio to use and that choice will correspond directly with a choice for subjectivist or objectivist philosophies. In the subjectivist’s view, the posterior odds ratio has a clear interpretation: if

\[ \frac{\Pi(\Theta_0|Y)}{\Pi(\Theta_1|Y)} > 1, \]

then the probability of \( \vartheta \in \Theta_0 \) is greater than the probability of \( \vartheta \in \Theta_1 \). Hence, if the posterior odds ratio exceeds one the subjectivist adopts \( H_0 \) rather than \( H_1 \); if, on the other hand, the posterior odds ratio lies below one, then the objectivist accepts \( H_1 \) and rejects \( H_0 \). The objectivist would object to this practice, saying that the relative prior weights of \( \Theta_0 \) and \( \Theta_1 \) can introduce a heavy bias in favour of one or the other in this approach (upon which the subjectivist would answer that that is exactly what he had in mind). The objectivist would prefer to use a criterion that is less dependent on the prior weights of \( \Theta_0 \) and \( \Theta_1 \). We look at a very simple example to illustrate the point.

**Example 2.16.** Let \( \Theta \) be a parameter space that consists of only two points, \( \theta_0 \) and \( \theta_1 \) and let \( \Theta_0 = \{ \theta_0 \}, \Theta_1 = \{ \theta_1 \} \), corresponding to simple null and alternative hypotheses \( H_0, H_1 \). Denote the prior by \( \Pi \) and assume that both \( \Pi(\{ \theta_0 \}) > 0 \) and \( \Pi(\{ \theta_1 \}) > 0 \). By Bayes’s rule, the posterior weights of \( \Theta_0 \) and \( \Theta_1 \) are

\[ \Pi(\vartheta \in \Theta_i|Y) = \frac{p_{\theta_i}(Y)\Pi(\Theta_i)}{p_{\theta_0}(Y)\Pi(\Theta_0) + p_{\theta_1}(Y)\Pi(\Theta_1)}, \]

for \( i = 0, 1 \). Therefore, the posterior odds ratio takes the form:

\[ \frac{\Pi(\vartheta \in \Theta_0|Y)}{\Pi(\vartheta \in \Theta_1|Y)} = \frac{p_{\theta_0}(Y)\Pi(\Theta_0)}{p_{\theta_1}(Y)\Pi(\Theta_1)}, \]

and the Bayes factor equals the likelihood ratio:

\[ B = \frac{p_{\theta_0}(Y)}{p_{\theta_1}(Y)}. \]

The objectivist prefers the Bayes factor to make a choice between two hypotheses: if \( B > 1 \) the objectivist adopts \( H_0 \) rather than \( H_1 \); if, on the other hand, \( B < 1 \), then the objectivist adopts \( H_1 \) rather than \( H_0 \). Note that the choice that results from this objective Bayesian testing procedure is identical to choice one makes based on the symmetric likelihood-ratio procedure of example 2.15.

We see that the Bayes factor does not depend on the prior weights of \( \Theta_0 \) and \( \Theta_1 \) but the posterior odds ratio does. Indeed, suppose we stack the prior odds heavily in favour of \( \Theta_0 \), by choosing \( \Pi(\Theta_0) = 1 - \varepsilon \) and \( \Pi(\Theta_1) = \varepsilon \) (for some small \( \varepsilon > 0 \)). Even if the likelihood ratio \( p_{\theta_0}(Y)/p_{\theta_1}(Y) \) is much smaller than one (but greater
than \( \epsilon / (1 - \epsilon) \), the subjectivist’s criterion favours \( H_0 \). In that case, the data clearly advocates hypothesis \( H_1 \) but the prior odds force adoption of \( H_0 \).

In example 2.16 the Bayes factor is independent of the choice of the prior. In general, the Bayes factor is not completely independent of the prior, but it does not depend on the relative prior weights of \( \Theta_0 \) and \( \Theta_1 \).

**Lemma 2.6.** Let \((\Theta, \mathcal{G})\) a measurable space parameterizing a model \( \Theta \to \mathcal{P} : \theta \mapsto P_\theta \) for data \( Y \in \mathcal{Y} \), with prior \( \Pi : \mathcal{G} \to [0, 1] \). Let \( \{ \Theta_0, \Theta_1 \} \) be a partition of \( \Theta \) such that \( \Pi(\Theta_0) > 0 \) and \( \Pi(\Theta_1) > 0 \). Then the Bayes factor \( B \) in favour of \( \Theta_0 \) does not depend on the prior odds ratio.

**Proof.** For any prior such that \( \Pi(\Theta_0) > 0 \) and \( \Pi(\Theta_1) > 0 \),

\[
\Pi(A) = \Pi(A|\Theta_0) \Pi(\Theta_0) + \Pi(A|\Theta_1) \Pi(\Theta_1),
\]

(2.37)

for all \( A \in \mathcal{G} \). In other words, \( \Pi \) is decomposed as a convex combination of two probability measures on \( \Theta_0 \) and \( \Theta_1 \) respectively. The Bayes factor is then rewritten (see (2.4)):

\[
B = \frac{\Pi(\Theta_0|Y) \Pi(\Theta_1)}{\Pi(\Theta_1|Y) \Pi(\Theta_0)} = \frac{\Pi(Y|\Theta_0)}{\Pi(Y|\Theta_1)},
\]

where, in a dominated model,

\[
\Pi(Y|\Theta_i) = \int_{\Theta_i} p_\theta(Y) d\Pi(\theta|\Theta_i),
\]

for \( i = 0, 1 \). In terms of the decomposition (2.37), \( B \) depends on \( \Pi(\cdot|\Theta_0) \) and \( \Pi(\cdot|\Theta_1) \), but not on \( \Pi(\Theta_0) \) and \( \Pi(\Theta_1) \).

So the difference between Bayes factors and posterior odds is exactly the bias introduced by non-zero prior odds; as such, it represents directly the difference between objectivist and subjectivist Bayesian philosophies.

**Example 2.17.** Consider a data vector \( Y = (X_1, \ldots, X_n) \) modelled as an i.i.d. sample from a uniform distribution \( U[\theta, \theta + 1] \), with \( \theta \in \Theta = [-1, 1] \). We formulate hypotheses,

\[
H_0 : \theta \geq 0, \quad H_1 : \theta < 0.
\]

and, to show how prior odds influence posterior odds but not Bayes factors, we use a prior with a Lebesgue density of the form,

\[
\pi(\theta) = \lambda 1\{\theta < 0\} + (1 - \lambda) 1\{\theta \geq 0\},
\]

for some \( 0 < \lambda < 1 \) (where it is noted that \( \lambda = 0 \) or \( \lambda = 1 \) would not be valid choices). Consequently, the prior odds in favour of \( \Theta_0 \) are \( 1 - 1/\lambda \). The likelihood is given by,

\[
L_n(\theta; X_1, \ldots, X_n) = \prod_{i=1}^{n} 1\{\theta \leq X_i \leq \theta + 1\},
\]
Bayesian basics

and the posterior density (relative to the Lebesgue measure on $\Theta = [-1, 1]$) is proportional to,

$$
\pi(\theta|X_1, \ldots, X_n) \propto \lambda 1\{\theta < 0\} 1\{\theta \leq X_{(1)}\} 1\{X_{(n)} \leq \theta + 1\} \\
+ (1 - \lambda) 1\{\theta \geq 0\} 1\{\theta \leq X_{(1)}\} 1\{X_{(n)} \leq \theta + 1\},
$$

where $X_{(1)}$ and $X_{(n)}$ denote first and last order statistics of the sample respectively. To calculate the posterior odds we do not need the normalization factor in the posterior and we see immediately that,

$$
\frac{\Pi(\theta \geq 0 | X_1, \ldots, X_n)}{\Pi(\theta < 0 | X_1, \ldots, X_n)} = \frac{1 - \lambda}{\lambda} \frac{\int_{0}^{1} 1\{X_{(n)} - 1 \leq \theta \leq X_{(1)}\} d\theta}{\int_{-1}^{0} 1\{X_{(n)} - 1 \leq \theta \leq X_{(1)}\} d\theta}
$$

Note the proportionality to the prior odds: the Bayes factor $B$ is equal to only the latter fraction in the expression on the right-hand side of the above display and is insensitive to the subjective choice for $\lambda$.

To conclude this section we make the following important remark.

**Remark 2.8.** The condition that both $\Theta_0$ and $\Theta_1$ receive prior mass strictly above zero is important since Bayes factors and odds ratios are based on conditioning of $\vartheta$. Bayesian hypothesis testing is sensible only if both $\Theta_0$ and $\Theta_1$ receive non-zero prior mass. This remark plays a role particularly when comparing a simple null hypothesis to an alternative, as illustrated in exercise 2.17.

## 2.5 Decision theory and classification

Many practical problems require that we make an observation and based on the outcome, make a decision of some kind. For instance when treating a patient, a doctor will observe a variety of diagnostic variables to come to a decision regarding the patient’s therapy. In financial markets, historical prices and volumes are analysed to decide how to (re-)position portfolios optimally. In a chemical plant, regulation of a chemical process amounts to a succession of decisions to control and optimize conditions, based on the measurement of thermodynamical quantities such as pressure, temperature and concentrations of chemicals involved. In this section, we look at problems of this nature, first from a frequentist perspective and then with the Bayesian approach.

Practical problems like those described above involve optimality criteria prescribed by the context of the problem (rather than just its statistical side, which concerns the observations only). For example, any statistical procedure meant to assist in medical diagnosis, should reflect that the misdiagnosis of a cancer patient has far more serious consequences than that of an influenza case. The methods of statistical inference that we have discussed thus far concentrate only on the stochastic description of the observations: the accuracy of an estimation procedure, coverage
probabilities for confidence intervals or the probability of Type-I and type-II errors in a testing procedure. By contrast, statistical decision theory formalizes optimality of decision-taking in terms of the (contextual) consequences of wrong decisions.

In statistical decision theory the nomenclature is slightly different from that introduced earlier. We consider a system that is in an unknown state \( \theta \in \Theta \), where \( \Theta \) is called the state space. The observation \( Y \) still takes its values in a measurable sample space \( (Y, \mathcal{B}) \) and is still considered stochastic. Its distribution \( P_\theta : \mathcal{B} \rightarrow [0,1] \) is a function of the state \( \theta \) of the system. The observation does not reveal the state of the system completely or with certainty. Based on the observation \( Y \), we take a decision \( a \in \mathcal{A} \) (or perform an action \( a \), as some prefer to say), where \( \mathcal{A} \) is the called the decision space. For each state \( \theta \) of the system there may be an optimal decision but since observation of \( Y \) does not give us the state \( \theta \) of the system with certainty, the decision is stochastic and may be suboptimal. The goal of statistical decision theory is to arrive at a rule that decides in the best possible way given only the data \( Y \).

If \( a \in \mathcal{A} \) is defined as a function of the state \( \theta \), the above does not add anything new to the approach we were already following: aside from the names, the concepts introduced here are those used in the usual problem of statistically estimating \( \theta(Y) \) based on data \( Y \sim P_\theta \). What sets decision theory apart is the formal introduction of the decision \( a \) and the associated notion of optimality, the loss-function.

**Definition 2.28.** Any lower-bounded function \( L : \Theta \times \mathcal{A} \rightarrow \mathbb{R} \) is a loss-function.

The loss-function has the following interpretation: if a decision \( a \) is taken while the state of the system is \( \theta \), then a loss \( L(\theta, a) \) is incurred. To illustrate, in systems where observation of the state is direct (i.e. \( Y = \theta \)) and non-stochastic, the optimal decision \( a(\theta) \) given the state \( \theta \) is any value of \( a \) that minimizes the loss \( L(\theta, a) \). The current problem is more difficult because the state \( \theta \) is unknown and can not be observed directly; all we have is the \( P_\theta \)-distributed observation \( Y \).

**Definition 2.29.** Let \( \mathcal{A} \) be a measurable space with \( \sigma \)-algebra \( \mathcal{H} \). A measurable \( \delta : \mathcal{H} \rightarrow \mathcal{A} \) is called a decision rule.

A decision-rule is a prescribed procedure to arrive at a decision \( \delta(y) \), for any possible realisation of the observation \( Y = y \). We denote the collection of all decision rules under consideration by \( \Delta \). Clearly our goal will be to find decision rules in \( \Delta \) that “minimize the loss” in an appropriate sense.

**Definition 2.30.** The risk-function \( R : \Theta \times \Delta \rightarrow \mathbb{R} \) is defined as the expected loss under \( Y \sim P_\theta \) when using \( \delta \),

\[
R(\theta, \delta) = \int L(\theta, \delta(y)) dP_\theta.
\]  

The above basic ingredients of decision-theoretic problems play a role in both the frequentist and Bayesian analysis. We consider the frequentist approach first and then look at decision theory from a Bayesian perspective.
2.5.1 Frequentist decision theory

Assuming the perspective of the frequentist, we suppose that $Y \sim P_{\theta_0}$ for some state $\theta_0 \in \Theta$ and would like to assess any decision rule $\delta$ according to the risk $P_{\theta_0}L(\theta_0, \delta)$ at $\theta_0$. But $\theta_0$ is unknown, so we are forced to consider all values of $\theta$ and look at the risk-function $\theta \mapsto R(\theta, \delta)$.

**Definition 2.31.** Let the state-space $\Theta$, states $P_{\theta}$, $(\theta \in \Theta)$, decision space $\mathcal{A}$ and loss $L$ be given. Choose $\delta_1, \delta_2 \in \Delta$. The decision rule $\delta_1$ is $R$-better than $\delta_2$, if

$$\forall \theta \in \Theta : \ R(\theta, \delta_1) < R(\theta, \delta_2). \tag{2.39}$$

A decision rule $\delta$ is admissible if there exists no $\delta' \in \Delta$ that is $R$-better than $\delta$ (and inadmissible if such a $\delta'$ does exist).

It is clear that the definition of $R$-better decision-rules is intended to order decision rules: if the risk-function associated with a decision-rule is relatively small, then that decision rule is preferable. Note, however, that the ordering we impose by definition 2.31 may be partial rather than complete: pairs $\delta_1, \delta_2$ of decision rules may exist such that neither $\delta_1$ nor $\delta_2$ is $R$-better than the other. This is due to the fact that $\delta_1$ may perform better (in the sense that $R(\theta, \delta_1) \leq R(\theta, \delta_2)$) for values of $\theta$ in some $\Theta_1 \subset \Theta$, while $\delta_2$ performs better in $\Theta_2 = \Theta \setminus \Theta_1$, resulting in a situation where (2.39) is true for neither. For that reason, it is important to find a way to compare risks (and thereby decision rules) in a $\theta$-independent way and thus arrive at a complete ordering among decision rules. This motivates the following definition.

**Definition 2.32.** (Minimax decision principle) Let the state-space $\Theta$, states $P_{\theta}$, $(\theta \in \Theta)$, decision space $\mathcal{A}$ and loss $L$ be given. The function

$$\Delta \rightarrow \mathbb{R} : \delta \mapsto \sup_{\theta \in \Theta} R(\theta, \delta)$$

is called the minimax risk. Let $\delta_1, \delta_2 \in \Delta$ be given. The decision rule $\delta_1$ is minimax-preferred to $\delta_2$, if

$$\sup_{\theta \in \Theta} R(\theta, \delta_1) < \sup_{\theta \in \Theta} R(\theta, \delta_2).$$

If $\delta^M \in \Delta$ minimizes $\delta \mapsto \sup_{\theta} R(\theta, \delta)$ then $\delta^M$ is called a minimax decision-rule.

One of the corner stones of decision theory is the so-called minimax theorem which guarantees the existence of minimax decision rules under very general conditions (see Sion (1958) [206]).

**Theorem 2.4.** Assume that $\Delta$ and $\Theta$ are convex sets. Furthermore, assume that the map $\Delta \rightarrow \mathbb{R} : \delta \mapsto R(\theta, \delta)$ is convex on $\Delta$ for every $\theta \in \Theta$ and that the map $\Theta \rightarrow \mathbb{R} : \theta \mapsto R(\theta, \delta)$ is concave on $\Theta$ for every $\delta$. Assume also that the topology on $\Delta$ is such that $\Delta$ is compact and $\delta \mapsto R(\theta, \delta)$ is continuous for all $\theta$. Then there exists a minimax decision rule $\delta^M$,

$$\sup_{\theta \in \Theta} R(\theta, \delta^M) = \inf_{\delta \in \Delta} \sup_{\theta \in \Theta} R(\theta, \delta) = \sup_{\theta \in \Theta} \inf_{\delta \in \Delta} R(\theta, \delta). \tag{2.40}$$
Proof. See Strasser (1985) [208], p. 239.

Since many loss-functions used in practice satisfy the convexity requirements, the minimax theorem has broad applicability in statistical decision theory and many other fields, particularly econometrics.

Note that the minimax theorem holds only for convex $\Delta$. In other words, if we want to guarantee the existence of an minimax-optimal decision rule, we are forced to consider convex combinations of decision rules. Unless $\mathcal{A}$ is a convex set, convex combinations of decision rules have no interpretation, so we adjust the definition of $\delta$ slightly.

**Definition 2.33.** Let $(\mathcal{A}, \mathcal{H})$ be a measurable space. A map $\delta$ that associates a random variable $\delta(y) \in \mathcal{A}$ with every possible realisation of the data $Y = y$, is called a randomised decision rule.

(A technical note: such maps have to be Markov kernels, that is, satisfy requirements 1 and 2 of definition B.16.) The decision procedure is adapted by randomisation: having seen $Y = y$ realised, we draw a random point in $\mathcal{A}$ from the distribution of $\delta(y)$. To complete the generalisation we revise the definition of the risk function to include an expectation over the distribution of $\delta(Y)$. The space $\Delta$ then concerns all maps that take $\mathcal{Y}$ into $\mathcal{M}_1(\mathcal{A})$, the space of all probability distributions on $(\mathcal{A}, \mathcal{H})$. Since $\Delta$ is convex, the minimax theorem asserts that there exists a randomised decision rule that optimises minimax risk. The fact that randomisation may lower risk is perhaps a bit counter-intuitive, but it is certainly in accordance with the fact that minimisation over a larger set produces a lower infimum.

The decision-theoretic approach can also be used to formulate estimation problems in a generalized way, if we choose the decision space $\mathcal{A}$ equal to the state-space $\Theta = \mathbb{R}$.

**Example 2.18. (Decision theoretic $L_2$-estimation)** Let $Y \sim N(\theta_0, 1)$ for some unknown $\theta_0 \in \Theta$. Choose $\mathcal{A} = \Theta$ and $L : \Theta \times \Theta \to \mathbb{R}$ equal to the quadratic difference, $L(\theta, a) = (\theta - a)^2$.

A choice referred to as an $L_2$-loss. Consider the decision-space

$$\Delta = \{ \delta_c : \mathcal{Y} \to \mathcal{A} : \delta_c(y) = cy, \ c \geq 0 \}.$$ 

Note that $\Delta$ plays the role of a family of estimators for $\theta_0$ here. The risk-function takes the form:
\[ R(\theta, \delta_c) = \int L(\theta, \delta_c(Y)) dP_\theta = \int (\theta - cy)^2 dN(\theta, 1)(y) \]
\[ = \int_{\mathbb{R}} (c(\theta - y) + (1-c)\theta)^2 dN(\theta, 1)(y) \]
\[ = \int_{\mathbb{R}} (c^2(y - \theta)^2 + 2c(1-c)\theta(y - \theta) + (1-c)^2\theta^2) dN(\theta, 1)(y) \]
\[ = c^2 + (1-c)^2\theta^2. \]

It follows that \( \delta_1 \) is \( R \)-better than all \( \delta_c \) for \( c > 1 \), so that for all \( c > 1 \), \( \delta_c \) is inadmissible. If we had restricted \( c \) to be greater than or equal to 1, \( \delta_c \) would have been admissible. However, since \( c \) may lie in \([0,1]\) as well, admissibility in the uniform sense of (2.39) does not apply to any \( \delta_c \). To see this, note that \( R(\theta, \delta_1) = 1 \) for all \( \theta \), whereas for \( c < 1 \) and some \( \theta > c/(1-c) \), \( R(0, \delta_c) < 1 < R(\theta, \delta_c) \). Therefore, there is no admissible decision rule in \( \Delta \).

The minimax criterion does give rise to a preference. However, in order to guarantee its existence, we need to bound (or rather, compactify) the parameter space: let \( M > 0 \) be given and assume that \( \Theta = [-M, M] \). The minimax risk for \( \delta_1 \) is given by
\[ \sup_{\theta \in \Theta} R(\theta, \delta_1) = c^2 + (1-c)^2M^2, \]
which is minimal iff \( c = M^2/(1+M^2) \), i.e. the (unique) minimax decision rule for this problem (or, since we are using decision theory to estimate a parameter in this case, the minimax estimator in \( \Delta \) with respect to \( L_2 \)-loss) is therefore,
\[ \delta^M(Y) = \frac{M^2}{1+M^2}Y. \]

Note that if we let \( M \to \infty \), this estimator for \( \theta \) converges to the MLE for said problem.

As demonstrated in the above example, uniform admissibility of a decision rule (c.f. (2.39)) is hard to achieve, but in many such cases a minimax decision rule does exist. One important remark concerning the use the minimax decision principle remains: considering (2.40), we see that the minimax principle chooses the decision rule that minimizes the maximum of the risk \( R(\cdot, \delta) \) over \( \Theta \). As such, the minimax criterion takes into account only the worst-case scenario and prefers decision rules that perform well under those conditions. In practical problems, that means that the minimax principle tends to take a rather pessimistic perspective on decision problems.
2.5 Decision theory and classification

2.5.2 Bayesian decision theory

Bayesian decision theory presents a more balanced perspective because instead of maximizing the risk function over $\Theta$, the Bayesian has the prior to integrate over $\Theta$. Optimization of the resulting integral takes into account more than just the worst case, so that the resulting decision rule is based on a less pessimistic perspective than the minimax decision rule.

**Definition 2.34.** Let the state-space $\Theta$, states $P_\theta, (\theta \in \Theta)$, decision space $A$ and loss $L$ be given. Additionally, assume that $(\Theta, \mathcal{G})$ and $(A, \mathcal{F})$ are measurable spaces, with prior $\Pi : \mathcal{G} \to \mathbb{R}$, and that $L$ and $\delta$ are measurable. The map

$$r(\Pi, \delta) = \int_\Theta R(\theta, \delta) d\Pi(\theta), \quad (2.41)$$

is called the *Bayesian risk function*. Let $\delta_1, \delta_2 \in \Delta$ be given. The decision rule $\delta_1$ is Bayes-preferred to $\delta_2$, if

$$r(\Pi, \delta_1) < r(\Pi, \delta_2).$$

If $\delta^\Pi \in \Delta$ minimizes $\delta \mapsto r(\Pi, \delta)$, i.e.

$$r(\Pi, \delta^\Pi) = \inf_{\delta \in \Delta} r(\Pi, \delta). \quad (2.42)$$

then $\delta^\Pi$ is called a *Bayes rule*. The quantity $r(\Pi, \delta^\Pi)$ is called the *Bayes risk*.

**Lemma 2.7.** Let $Y \in \mathcal{Y}$ denote data in a decision theoretic problem with state space $\Theta$, decision space $A$ and loss $L : \Theta \times A \to \mathbb{R}$. For any prior $\Pi$ and all measurable decision rules $\delta : \mathcal{Y} \to A$,

$$r(\Pi, \delta) \leq \sup_{\theta \in \Theta} R(\theta, \delta),$$

i.e. the Bayesian risk is always upper bounded by the minimax risk.

The proof of this lemma follows from the fact that the minimax risk is an upper bound for the integrand in the Bayesian risk function.

**Example 2.19.** (see example 2.18) Let $\Theta = \mathbb{R}$ and $Y \sim N(\theta_0, 1)$ for some unknown $\theta_0 \in \Theta$. Choose the loss-function $L : \Theta \times A \to \mathbb{R}$ and the decision space $\Delta$ as in example 2.18. We choose a prior $\Pi = N(0, \tau^2)$ (for some $\tau > 0$) on $\Theta$. Then the Bayesian risk function is given by:

$$r(\Pi, \delta_c) = \int_\Theta R(\theta, \delta_c) d\Pi(\theta) = \int_\mathbb{R} (c^2 + (1-c)^2 \theta^2) dN(0, \tau^2)(\theta)$$

$$= c^2 + (1-c)^2 \tau^2,$$

which is minimal iff $c = \tau^2/(1 + \tau^2)$. The (unique) Bayes rule for this problem and corresponding Bayes risk are therefore,
\[ \delta^\Pi(Y) = \frac{\tau^2}{1 + \tau^2} Y, \quad r(\Pi, \delta^\Pi) = \frac{\tau^2}{1 + \tau^2}. \]

In the Bayesian case, there is no need for a compact parameter space \( \Theta \), since we do not maximize but integrate over \( \Theta \).

In the above example, we could find the Bayes rule by straightforward optimization of the Bayesian risk function, because the class \( \Delta \) was rather restricted. If we extend the class \( \Delta \) to contain all non-randomized decision rules, the problem of finding the Bayes rule seems to be far more complicated at first glance. However, as we shall see in theorem 2.1, the following definition turns out to be the solution to this question.

**Definition 2.35.** (The conditional Bayes decision principle) Let the state-space \( \Theta \), states \( P_\theta, (\theta \in \Theta) \), decision space \( \mathcal{A} \) and loss \( L \) be given. In addition, assume that \((\Theta, \mathcal{G})\) and \((\mathcal{A}, \mathcal{H})\) are measurable spaces, with prior \( \Pi : \mathcal{G} \to \mathbb{R} \), and that \( L \) is measurable. We define the decision rule \( \delta^* : \mathcal{Y} \to \mathcal{A} \) to be such that for \( P^\Pi \)-almost-all \( y \in \mathcal{Y} \),

\[
\int_{\Theta} L(\theta, \delta^*(y)) \, dP^\Pi(\theta | Y = y) = \inf_{a \in \mathcal{A}} \int_{\Theta} L(\theta, a) \, dP^\Pi(\theta | Y = y). \tag{2.43}
\]

Point-wise for almost-all \( y \), the decision rule \( \delta^*(y) \) is assumed to minimize the posterior expected loss. This defines the decision rule \( \delta^* \) implicitly as a point-wise minimizer, which raises the usual questions concerning existence and uniqueness, of which little can be said in any generality. However, if existence (and measurability) of \( \delta^* \) is established, \( \delta^* \) is Bayes-risk optimal.

**Theorem 2.5.** Let the state-space \( \Theta \), states \( P_\theta, (\theta \in \Theta) \), decision space \( \mathcal{A} \) and loss \( L \) be given. Assume that \((\Theta, \mathcal{G})\) and \((\mathcal{A}, \mathcal{H})\) are measurable spaces, with prior \( \Pi : \mathcal{G} \to \mathbb{R} \), and that \( L \) is measurable. If the decision rule \( \delta^* : \mathcal{Y} \to \mathcal{A} \) is a well-defined, measurable map then \( \delta^* \) is a Bayes rule.

**Proof.** Denote the class of all decision rules for this problem again by \( \Delta \). According to theorem 2.1 (more particularly, exercise ??) for any decision rule \( \delta : \mathcal{Y} \to \mathcal{A} \),

\[
r(\Pi, \delta) = \int_{\Theta} R(\theta, \delta) \, dP^\Pi(\theta) = \int_{\Theta} \int_{\mathcal{Y}} L(\theta, \delta(y)) \, dP_\theta(y) \, dP^\Pi(\theta)
\]

\[
= \int_{\mathcal{Y}} \int_{\Theta} L(\theta, \delta(y)) \, dP^\Pi(\theta | Y = y) \, dP^\Pi(y).
\]

By assumption, the conditional Bayes decision rule \( \delta^* \) exists. Since \( \delta^* \) satisfies (2.43) point-wise for all \( y \in \mathcal{Y} \), we have

\[
\int_{\Theta} L(\theta, \delta^*(y)) \, dP^\Pi(\theta | Y = y) = \inf_{\delta \in \Delta} \int_{\Theta} L(\theta, \delta(y)) \, dP^\Pi(\theta | Y = y).
\]

Substituting this in (2.18), we obtain
2.5 Decision theory and classification

\[ r(\Pi, \delta^*) = \int_{\mathcal{Y}} \inf_{\delta \in \Delta} \int_{\Theta} L(\theta, \delta(y)) d\Pi(\theta | Y = y) dP^\Pi(y) \]
\[ \leq \inf_{\delta \in \Delta} \int_{\mathcal{Y}} \int_{\Theta} L(\theta, \delta(y)) d\Pi(\theta | Y = y) dP^\Pi(y) = \inf_{\delta \in \Delta} r(\Pi, \delta). \]

which proves that \( \delta^* \) is a Bayes rule.

To conclude, it is noted that randomization of the decision is not needed when optimizing with respect to the Bayes risk. The conditional Bayes decision rule is non-randomized and optimal.

2.5.3 Frequentist versus Bayesian classification

Many decision-theoretic questions take the form of a classification problem: under consideration is a population \( \Omega \) of objects that each belong to one of a finite number of classes \( \mathcal{A} = \{1, 2, \ldots, L\} \). The class \( K \) of the object is the unknown quantity of interest. Observing a vector \( Y \) of features of the object, the goal is to classify the object, i.e. estimate which class it belongs to. We formalize the problem in decision-theoretic terms: the population is a probability space \((\Omega, \mathcal{F}, P)\); both the feature vector and the class of the object are random variables, \( Y : \Omega \to \mathcal{Y} \) and \( K : \Omega \to \mathcal{A} \) respectively. The state-space in a classification problem equals the decision space \( \mathcal{A} \): the class can be viewed as a “state” in the sense that the distribution \( P(Y | K = k) \) of \( Y \) given the class \( K = k \) depends on \( k \). Based on the feature vector \( Y \), we decide to classify in class \( \delta(Y) \), i.e. the decision rule (or classifier, as it is usually referred to in the context of classification problems) maps features to classes by means of a map \( \delta : \mathcal{Y} \to \mathcal{A} \). A classifier \( \delta \) can be viewed equivalently as a finite partition of the feature-space \( \mathcal{Y} \): for every \( k \in \mathcal{A} \), we define

\[ \mathcal{Y}_k = \{ y \in \mathcal{Y} : \delta(y) = k \} \]

and note that if \( k \neq l \), then \( \mathcal{Y}_k \cap \mathcal{Y}_l = \emptyset \) and \( \mathcal{Y}_1 \cup \mathcal{Y}_2 \cup \ldots \cup \mathcal{Y}_L = \mathcal{Y} \). The partition of the feature space is such that if \( Y = y \in \mathcal{Y}_k \) for certain \( k \in \mathcal{A} \), then we classify the object in class \( k \).

Depending on the context of the classification problem, a loss-function \( L : \mathcal{A} \times \mathcal{A} \to \mathbb{R} \) is defined (see the examples in the introduction to this section, e.g. the example on medical diagnosis). Without context, the loss function in a classification problem can be chosen as follows

\[ L(k, l) = 1_{\{k \neq l\}}. \]

i.e. we incur a loss equal to one for each misclassification. Using the minimax decision principle, we look for a classifier \( \delta^M : \mathcal{Y} \to \mathcal{A} \) that minimizes:

\[ \delta \mapsto \sup_{k \in \mathcal{A}} \int_{\mathcal{Y}} L(k, \delta(y)) dP(y | K = k) = \sup_{k \in \mathcal{A}} P(\delta(Y) \neq k | K = k), \]
i.e. the minimax decision principle prescribes that we minimize the probability of misclassification uniformly over all classes.

In a Bayesian context, we need a prior on the state-space, which equals $\mathcal{A}$ in classification problems. Note that if known (or estimable), the marginal probability distribution for $K$ to be used as the prior for the state $k$, in accordance with definition 2.2. In practical problems, frequencies of occurrence for the classes $\{1, \ldots, L\}$ in $\Omega$ are often available or easily estimable; in the absence of information on the marginal distribution of $K$ equal prior weights can be assigned. Here, we assume that the probabilities $P(K = k)$ are known and use them to define the prior density with respect to the counting measure on the (finite) space $\mathcal{A}$:

$$\pi(k) = P(K = k).$$

The Bayes rule $\delta^* : \mathcal{Y} \rightarrow \mathcal{A}$ for this classification problem is defined as the minimizer of

$$\delta \mapsto \sum_{k \in \mathcal{A}} L(k, \delta(y)) d\Pi(k|Y = y) = \sum_{k=1}^{L} \Pi(\delta(y) \neq K \mid Y = y)$$

for every $y \in \mathcal{Y}$. According to theorem 2.5, the classifier $\delta^*$ minimizes the Bayes risk, which in this situation is given by:

$$r(\Pi, \delta) = \sum_{k \in \mathcal{A}} R(k, \delta) \pi(k) = \sum_{k \in \mathcal{A}} \int_{\mathcal{Y}} L(k, \delta(y)) dP(y|K = k) \pi(k)$$

$$= \sum_{k \in \mathcal{A}} P(k \neq \delta(Y) \mid K = k) P(K = k) = P(K \neq \delta(Y)).$$

Summarizing, the Bayes rule $\delta^*$ minimizes the overall probability of misclassification, i.e. without referring to the class of the object. (Compare this with the minimax classifier.) Readers interested in the statistics of classification and its applications are encouraged to read B. Ripley’s “Pattern recognition and neural networks” (1996) [195].

To close the chapter, the following remark is in order: when we started our comparison of frequentist and Bayesian methods, we highlighted the conflict in philosophy. However, now that we have seen some of the differences in more detail by considering estimation, confidence sets, testing and decision theory in both schools, we can be far more specific. Statistical problems can be solved in both schools; whether one chooses for a Bayesian or frequentist solution is usually not determined by deeply felt belief in either philosophy but by much more practical considerations. Perhaps the classification example of this subsection illustrates this point most clearly: if one is concerned about correct classification for objects in the most difficult class, one should opt for the minimax decision rule. If, on the other hand, one wants to minimize the overall misclassification probability (disregarding misclassification per class), one should choose to adopt the conditional Bayes decision rule. In other words, depending on the risk to be minimized (minimax risk and Bayes risk are different!) one arrives at different classifiers.
An argument in favour of Bayesian methods is the control that the statistician gains over a statistical procedure by choice of a well-chosen bias, expressed through a suitable prior (the subjective bias intended in example 1.8). To give another example of this nature, Bayesian methods are popular in forensic statistics because the freedom to choose a prior leaves room to incorporate background information and common-sense. From a more technical point of view, bias may be required for regularization purposes (like a penalty in frequentist terms, see remark 2.6). Prior bias may even be guided in a data-dependent way, e.g. when we employ empirical Bayesian methods (see section 3.4), to optimize the procedure.

Another reason to use one or the other may be computational advantages or useful theoretical results that exist for one school but have no analog in the other. Philosophical preference should not play a role in the choice for a statistical procedure, practicality should (and usually does).

2.6 Exercises

2.1. CALIBRATION
A physicist prepares for measurement of a physical quantity $Z$ in his laboratory. To that end, he installs a measurement apparatus that will give him an outcome of the form $Y = Z + e$ where $e$ is a measurement error due to the inaccuracy of the apparatus, assumed to be stochastically independent of $Z$. Note that if the expectation of $e$ equals zero, long-run sample averages converge to the expectation of $Z$; if $Pe \neq 0$, on the other hand, averaging does not cancel out the resulting bias. The manufacturer of the apparatus says that $e$ is normally distributed with known variance $\sigma^2 > 0$. The mean $\theta$ of this normal distribution depends on the way the apparatus is installed and thus requires calibration. The following questions pertain to the calibration procedure.

The physicist decides to conduct the following steps to calibrate his measurement. First, he makes certain that the apparatus receives no input signal, $Z = 0$. Then he repeats measurement of $Y$, generating an i.i.d. sample of size $n$, which amounts to an i.i.d. sample from the distribution of $e$ used to estimate the unknown mean $\theta$. The physicist expects that $Ee$ lies close to zero.

a. Explain why, from a subjectivist point of view, the choice $\theta \sim N(0, \tau^2)$ forms a suitable prior in this situation. Explain the role of the parameter $\tau^2 > 0$.

b. With the choice of prior as in part a., calculate the posterior density for $\theta$.

c. Interpret the influence of $\tau^2$ on the posterior, taking into account your answer under part a. (Hint: take limits $\tau^2 \downarrow 0$ and $\tau^2 \uparrow \infty$ in the expression you have found under b.)

d. What is the influence of the sample size $n$? Show that the particular choice of the constant $\tau^2$ becomes irrelevant in the large-sample limit $n \to \infty$.

2.2. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the uniform distribution $U[0, \theta]$, with unknown parameter $\theta \in \Theta = (1, \infty)$. As a prior for $\theta$, choose the Pareto distribu-
tion with exponent $\alpha > 0$. Calculate the posterior density for $\theta$ with respect to the Lebesgue measure on $[0, \infty)$.

2.3. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Poisson distribution $\text{Poisson}(\lambda)$, with unknown parameter $\lambda > 0$. As a prior for $\lambda$, let $\lambda \sim \Gamma(2, 1)$. Calculate the posterior density for $\lambda$ with respect to the Lebesgue measure on $[0, \infty)$.

2.4. Let $X_1, \ldots, X_n$ be an i.i.d. sample from a binomial distribution $\text{Bin}(k, \theta)$, with known $k \geq 1$ and unknown $\theta \in \Theta = [0, 1]$. As a prior for $\theta$, use a beta distribution, $\theta \sim \beta(2, 2)$. Calculate the posterior density for $\theta$ with respect to the Lebesgue measure on $[0, 1]$.

2.5. Let $X_1, \ldots, X_n$ be an i.i.d. sample from a normal distribution $\mathcal{N}(0, \sigma^2)$, with unknown $\sigma^2 > 0$. We define the prior for the variance $\sigma^2$ implicitly, by stating that the inverse $1/\sigma^2$ is distributed according to a $\Gamma(\alpha, \beta)$ distribution. Calculate the posterior density for $\sigma^2$ with respect to the Lebesgue measure on $[0, \infty)$.

2.6. Let $(\mathcal{P}, \mathcal{F}, \Pi)$ be a model with prior for i.i.d. $X_1, \ldots, X_n$ taking values in a sample space $(\mathcal{X}, \mathcal{B})$. Suppose that the model is dominated by a $\sigma$-finite measure $\mu$ on $(\mathcal{X}, \mathcal{B})$ and that the prior is dominated by a $\sigma$-finite measure $\nu$ on $(\mathcal{P}, \mathcal{G})$. Show that if $\mu'$ is another $\sigma$-finite measure on $(\mathcal{X}, \mathcal{B})$, such that $\mathcal{P} \ll \mu' \ll \mu$, and $\nu'$ is another $\sigma$-finite measure on $(\mathcal{P}, \mathcal{G})$, such that $\Pi \ll \nu' \ll \nu$, then the MAP estimator $\hat{\theta}_2$ does not change with $\mu'$ (compare with exercise 1.1), but $\hat{\theta}_2$ does change with $\nu'$.

2.7. In the model of exercise 2.2, calculate the maximum-likelihood estimator, the posterior mean and the maximum-a-posteriori estimator.

2.8. In the model of exercise 2.3, calculate the maximum-likelihood estimator, the posterior mean and the maximum-a-posteriori estimator.

2.9. In the model of exercise 2.4, calculate the maximum-likelihood estimator, the posterior mean and the maximum-a-posteriori estimator.

2.10. In the model of exercise 2.5, calculate the maximum-likelihood estimator, the posterior mean and the maximum-a-posteriori estimator.

2.11. Consider the following questions in the context of exercise 2.3, after exercise 2.8.

a. Let $n \to \infty$ both in the MLE and MAP estimator and conclude that the difference vanishes in the limit, $P_\lambda$-almost-surely.

b. Following remark 2.6, explain the difference between ML and MAP estimators exclusively in terms of the prior.

c. Consider and discuss the choice of prior $\lambda \sim \Gamma(2, 1)$ twice, once in a qualitative, subjectivist Bayesian fashion, and once following the frequentist interpretation of the log-prior-density.
2.12. Let \( Y \sim P_0 \) denote the data and \( \mathcal{P} \) a model with metric \( d \). Suppose that \( \mathcal{P} \) is endowed with a prior defined on the Borel \( \sigma \)-algebra induced by the metric topology. Assume that \( P_0 \ll P_\Pi \) and that \( \mathcal{P} \) is compact. The following questions pertain to the small-ball estimators defined in definition 2.11 and remark 2.5. We assume that the posterior distribution is such that for all \( \varepsilon > 0 \) and all \( P \in \mathcal{P} \), the (topological) boundary of the ball \( B_{d}(P, \varepsilon) \) receives mass equal to zero: \( \Pi(\partial B_{d}(P, \varepsilon)|Y) = 0 \), \( P_0 \)-almost surely.

a. Show that, for any \( p \in (1/2, 1) \) and large enough \( \varepsilon > 0 \), the small-ball estimator \( \hat{P} \) of exists, \( P_0 \)-almost-surely.

b. Show that for any two measurable model subsets \( A, B \subset \mathcal{P} \),

\[
\left| \Pi(A|Y) - \Pi(B|Y) \right| \leq \Pi(A \cup B|Y) - \Pi(A \cap B|Y),
\]

\( P_0 \)-almost-surely.

c. Show that for every \( \varepsilon > 0 \), the map \( P \mapsto \Pi(B_{d}(P, \varepsilon)|Y) \) is continuous, \( P_0 \)-almost-surely.

d. Show that for every \( \varepsilon > 0 \), the small-ball estimator of definition 2.11 exists.

e. Let some \( p \in (1/2, 1) \) be given. Suppose that \( \varepsilon > 0 \) denotes some radius for which there exists a ball \( B_{d}(P, \varepsilon) \subset \mathcal{P} \) of posterior probability greater than or equal to \( p \). Show that, if both \( \hat{P}_1 \) and \( \hat{P}_2 \) are centre points of such balls, then \( d(\hat{P}_1, \hat{P}_2) < 2\varepsilon \), \( P_0 \)-almost-surely.

2.13. Let \( Y_n = (X_1, \ldots, X_n) \) be an i.i.d. sample from the normal distribution \( N(\mu, \sigma^2) \) for certain \( \mu \in \mathbb{R}, \sigma^2 > 0 \). Show that the sample average is distributed according to the normal distribution,

\[
P_{n}X \sim N(\mu, \sigma^2_n),
\]

with variance \( \sigma^2_n = \sigma^2/n \).

2.14. Let \( Y \) be normally distributed with known variance \( \sigma^2 > 0 \) and unknown location \( \theta \). As a prior for \( \theta \), choose \( \Pi = N(0, \tau^2) \). Let \( \alpha \in (0, 1) \) be given. Using the posterior density with respect to the Lebesgue measure, express the level-\( \alpha \) HPD-credible set in terms of \( Y, \sigma^2, \tau^2 \) and quantiles of the standard normal distribution. Consider the limit \( \tau^2 \to \infty \) and compare with level-\( \alpha \) confidence intervals centred on the ML estimate for \( \theta \).

2.15. Let \( Y \sim \text{Bin}(n; p) \) for known \( n \geq 1 \) and unknown \( p \in (0, 1) \). As a prior for \( p \), choose \( \Pi = \text{Beta}(\frac{1}{2}, \frac{1}{2}) \). Calculate the posterior distribution for the parameter \( p \). Using the Lebesgue measure on \( (0, 1) \) to define the posterior density, give a level-\( \alpha \) (non-HPD!) credible interval for \( p \) in terms of \( Y, n \) and the quantiles of beta-distributions.

2.16. Let \( \Theta \) be a subset of \( \mathbb{R} \) and let \( \Theta \to \mathcal{P} : \theta \mapsto P_{\theta} \) describe an identifiable parametrization of the model \( \mathcal{P} \) for an i.i.d. sample \( X_1, \ldots, X_n \), and assume that there exists a \( \theta_0 \in \Theta \) such that \( P_{\theta_0} \) is the marginal distribution for each of the \( X_i \). Let \( \theta \) and \( \theta' \) with \( \theta' > \theta \) from \( \Theta \) be given and consider the hypotheses.
Given a significance level $\alpha \in (0,1)$, write down the Neyman-Pearson test for $H_0$ versus $H_1$ (see lemma 2.4), in each of the following cases,

a. for all $\theta \in [0,1], P_\theta = \text{Bernoulli}(\theta)$;

b. for all $\theta \in (0,\infty), P_\theta = \text{Poisson}(\theta)$;

c. for all $\theta \in [0,1], P_\theta = \text{Bin}(\theta,k)$ for some known integer $k \geq 1$.

2.17. Consider a dominated model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ for data $Y$, where $\Theta \subset \mathbb{R}$ is an interval. For certain $\theta_0 \in \Theta$, consider the simple null-hypothesis and alternative:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0.$$ 

Show that if the prior $\Pi$ is absolutely continuous with respect to the Lebesgue measure on $\Theta$, then the posterior odds ratio in favour of the hypothesis $H_0$ equals zero. [Remark: conclude that calculation of posterior odds ratios makes sense only if both hypotheses receive non-zero prior mass. Otherwise, the statistical question we ask is rendered invalid e x a n t e by our beliefs concerning $\theta$, as expressed through the choice of the prior. (See example 2.3.)]

2.18. Let $X_1, \ldots, X_n$ be an i.i.d. sample from a binomial distribution $\text{Bin}(\theta,k)$, for some known integer $k \geq 1$ and an unknown parameter $\theta \in \Theta = [0,1]$. Let the prior $\Pi$ for $\theta$ be a Beta distribution $\text{B}(\alpha,\beta)$, with certain parameters $\alpha, \beta > 1$.

a. Calculate the posterior distribution for $\theta$.

b. Write down the equations that determine the two end-points of the HPD credible interval (based on the density of the posterior relative to Lebesgue measure on $\Theta$), for given credible level $\delta \in (0,1)$.

Consider the hypotheses,

$$H_0 : \theta \leq \frac{1}{2}, \quad H_1 : \theta > \frac{1}{2}.$$ 

c. Give the prior odds, posterior odds and Bayes factor for the hypotheses $H_0$ and $H_1$.

2.19. PRISONER’S DILEMMA

Two men have been arrested on the suspicion of burglary and are held in separate cells awaiting interrogation. The prisoners have been told that burglary carries a maximum sentence of $x$ years. However, if they confess, their prison terms are reduced to $y$ years (where $0 < y < x$). If one of them confesses and the other does not, the first receives a sentence of $y$ years while the other is sentenced to $x$ years.

Guilty of the crime he is accused of, our prisoner contemplates whether to confess to receive a lower sentence, or to deny involvement in the hope of escaping justice altogether. He cannot confess without implicating the other prisoner. If he keeps his mouth shut and so does his partner in crime, they will both walk away free. If he keeps his mouth shut but his partner talks, he gets the maximum sentence. If he talks, he will always receive a sentence of $y$ years and the other prisoner
receives \( y \) or \( x \) years depending on whether he confessed or not himself. To talk or not to talk, that is the question.

There is no data in this problem, so we set \( \theta \) equal to 1 or 0, depending on whether the other prisoner talks or not. Our prisoner can decide to talk (\( t = 1 \)) or not (\( t = 0 \)). The loss function \( L(\theta, t) \) equals the prison term for our prisoner. In the absence of data, risk and loss are equal.

a. Calculate the minimax risk for both \( t = 0 \) and \( t = 1 \). Argue that the minimax-optimal decision for our prisoner is to confess.

As argued in section 2.5, the minimax decision can be overly pessimistic. In the above, it assumes that the other prisoner will talk and chooses \( t \) accordingly.

The Bayesian perspective balances matters depending on the chance that the other prisoner will confess when interrogated. This chance finds its way into the formalism as a prior for the trustworthiness of the other prisoner. Let \( p \in [0, 1] \) be the probability that the other prisoner confesses, i.e. \( \Pi(\theta = 1) = p \) and \( \Pi(\theta = 0) = 1 - p \).

b. Calculate the Bayes risks for \( t = 0 \) and \( t = 1 \) in terms of \( x \), \( y \) and \( p \). Argue that the Bayes decision rule for our prisoner is as follows: if \( y / x > p \) then our prisoner does not confess, if \( y / x < p \), the prisoner confesses. If \( y / x = p \), the Bayes decision criterion does not have a preference.

So, depending on the degree to which our prisoner trusts his associate and the ratio of prison terms, the Bayesian draws his conclusion. The latter is certainly more sophisticated and perhaps more realistic, but it requires that our prisoner quantifies his trust in his partner in the form of a prior Bernoulli(\( p \)) distribution.
Bayesian procedures have been the object of much criticism, often focusing on the choice of the prior as an undesirable source of ambiguity. The answer of the subjectivist that the prior represents the “belief” of the statistician or “expert knowledge” pertaining to the measurement elevates this ambiguity to a matter of principle, thus setting the stage for a heated debate between “pure” Bayesians and “pure” frequentists concerning the philosophical merits of either school within statistics. As said, the issue is complicated further by the fact that the Bayesian procedure does not refer to the “true” distribution $P_0$ for the observation (see section 2.1), providing another point of fundamental philosophical disagreement for the fanatically pure to lock horns over. Leaving the philosophical argumentation to others, we shall try to discuss the choice of a prior at a more conventional, practical level.

In this chapter, we look at the choice of the prior from various points of view: in section 3.1, we consider the priors that emphasize the subjectivist’s prior “belief”. In section 3.2 we construct priors with the express purpose not to emphasize any part of the model, as advocated by objectivist Bayesians. Hierarchical prior construction and Bayesian modelling are the subject of section 3.3, and methods that choose priors by frequentist means (commonly known as empirical Bayes methods) forms the subject of section 3.4. Because it is mathematically desirable and computationally advantageous to have closed-form expressions for posterior distributions, so-called conjugacy of families of distributions over the model is considered in section 3.5. Special attention goes to the Dirichlet distributions of section 3.6 because they describe a conjugate family of probability distributions on spaces of probability measures (rather than their parametrizing spaces). As will become clear in the course of the chapter, the choice of a “good” prior is highly dependent on the model under consideration, as well as on the purpose of the analysis.

All of the material presented in the first sections of this chapter applies only in parametric models. To find a suitable prior for a non-parametric model can be surprisingly difficult. The concept of a measure on an infinite-dimensional space has technical subtleties that do not play a role in parametric models (e.g. the Radon property and the lack of default dominating measures like Lebesgue measure or counting measure in finite-dimensional or discrete models). This makes the defi-
nition of priors relatively hard and can lead to unexpected behaviour of posterior distributions, as we shall see in part II. One construction stands out as completely natural, however, and it is built with priors on finite-dimensional spaces: in subsection ?? we define a conjugate family of priors/posteriors on (fully) non-parametric models in terms of a directed set of finite partitions of the sample space, with associated finite-dimensional Dirichlet distributions, the so-called Dirichlet process prior.

3.1 Subjective priors

As was explained in chapters 1 and 2, all statistical procedures require the statistician to make certain choices, e.g. for model and method of inference. The subjectivist chooses the model as a collection of stochastic explanations of the data that he finds “reasonable”, based on criteria no different from those frequentists and objectivist Bayesians would use.

3.1.1 Motivation for the subjectivist approach

Bayesians then proceed to choose a prior, in such a manner that the support of this prior is not essentially smaller than the model itself. But even when the support of the prior is fixed, there is a large collection of possible priors left to be considered, each leading to a different posterior distribution. The objectivist Bayesian will choose from those possibilities a prior that is “homogeneous” (in a suitable sense), in the hope of achieving unbiased inference. The subjectivist, however, chooses his prior such as to emphasize parts of the model that he believes in stronger than others, thereby introducing a bias in his inferential procedure explicitly. Such a prior is called a subjective prior, or informative prior. The reason for this approach is best explained by examples like 1.8, which demonstrate that intuitive statistical reasoning is not free of bias either.

Subjectivity finds its mathematical expression when high prior “belief” is translated into “relatively large” amounts of assigned prior mass to certain regions of the model. However, there is no clear rule directing the exact fashion in which prior mass is to be distributed. From a mathematical perspective this is a rather serious shortcoming, because it leaves us without a precise definition of the subjectivist approach. Often the subjectivist will have a reasonably precise idea about his “beliefs” at the roughest level (e.g. concerning partitions of the model into a few subsets) but none at more detailed levels. When the parameter space $\Theta$ is unbounded this lack of detail becomes acute, given that the tail of the prior is hard to fix by subjective reasoning, yet highly influential for the inferential conclusions based on it. In practice, a subjectivist will often choose his prior without mathematical precision. He considers the problem, interprets the parameters in his model and chooses a prior
3.1 Subjective priors

to reflect all the (background) information at his disposition, ultimately filling in remaining details in an ad-hoc manner. It is worthwhile to mention that studies have been conducted focused on the ability of people to make a realistic guess at a probability distribution: they have shown that without specific training or practice, people tend to be overconfident in their assessment, assigning too much mass to possibilities they deem most likely and too little to others [3]. This suggests that people tend to formulate their “beliefs” on a deterministic basis and deviate from that point of view only slightly (or, too little) when asked to give a realistic probabilistic perspective. (For more concerning the intricacies of choosing subjective prior distributions, see Berger (1985) [19].)

Remark 3.1. For this reason it is imperative that a subjectivist prior choice is fully described alongside inferential conclusions based upon it. Reporting on methods is important in any statistical setting, but if chosen methods lead to express bias, explanation is even more important. Indeed, not only the prior but also the reasoning leading to its choice should be reported, because in a subjectivist setting, the motivation for the choice of a certain prior (and not any other) is part of the analysis rather than an external consideration.

3.1.2 Methods for the construction of subjective priors

If the model \( \Theta \) is one-dimensional and the parameter \( \theta \) has a clear interpretation, it is often not exceedingly difficult to find a reasonable prior \( \Pi \) expressing the subjectivist’s “belief” concerning the value of \( \theta \).

Example 3.1. If one measures the speed of light in vacuo \( c \) (a physical constant, approximately equal to 299792458 m/s), the experiment will be subject to random perturbations outside the control of the experimenter. For example, imperfection of the vacuum in the experimental equipment, small errors in timing devices, electronic noise and countless other factors may influence the resulting measured speed \( Y \). We model the perturbations collectively as a normally distributed error \( e \sim N(0, \sigma^2) \) where \( \sigma \) is known as a characteristic of the experimental setup. The measured speed is modelled as \( Y = c + e \), i.e. the model \( \mathcal{P} = \{N(c, \sigma^2) : c > 0\} \) is used to infer on \( c \). Based on experiments in the past (most famous is the Michelson-Morley experiment (1887)), the experimenter knows that \( c \) has a value close to \( 3 \cdot 10^8 \) m/s, so he chooses his prior to reflect this: a normal distribution located at 30000000 m/s with a standard deviation of (say) 1000000 m/s will do. The latter choice is arbitrary, just like the choice for a normal distribution over other possible choices.

The situation changes when the parameter has a higher dimension, \( \Theta \subset \mathbb{R}^d \): first of all, interpretability of each of the \( d \) components of \( \theta = (\theta_1, \theta_2, \ldots, \theta_d) \) can be far from straightforward, so that concepts like prior “belief” or “expert knowledge” become inadequate guidelines for the choice of a prior. Additionally, the choice for a prior in higher-dimensional models also involves choices concerning the dependence structure between parameters.
Remark 3.2. Often, subjectivist inference employs exceedingly simple, parametric models for the sake of interpretability of the parameter (and to be able to choose a prior accordingly). Most frequentists would object to such choices for their obvious lack of realism, since they view the data as being generated by a “true, underlying distribution”, usually assumed to be an element of the model. By contrast, the subjectivist does not have the ambition to be strictly realistic and calls for interpretability instead: to the subjectivist, inference is a personal rather than a universal matter. As such, the preference for simple parametric models is a matter of subjective interpretation rather than an assumption concerning reality or realistic distributions for the data.

When confronted with the question which subjective prior to use on a higher-dimensional model, it is often of help to define the prior in several steps based on a choice for the dependence structure between various components of the parameter. Suppose that the subjectivist can formulate a reasonable distribution for the first component $\theta_1$, provided he can think about the other components $\theta_2, \ldots, \theta_d$ as being fixed at any value: this prescribes the conditional prior distribution $\Pi_{\theta_1 | \theta_2, \ldots, \theta_d}$ of $\theta_1$ given the other components. Next suppose that a reasonable subjective prior for the second component may be found, conditional on $\theta_3, \ldots, \theta_d$. This amounts to specification of the conditional distribution $\Pi_{\theta_2 | \theta_3, \ldots, \theta_d}$. If we continue like this, eventually defining the marginal prior $\Pi_{\theta_d}$ for the last component $\theta_d$, we have found a dependent prior for the full parameter $\theta$, because for all $A_1, \ldots, A_d \in \mathcal{B}$,

$$
\Pi(\theta_1 \in A_1, \ldots, \theta_d \in A_d) = 
\Pi(\theta_1 \in A_1 | \theta_2 \in A_2, \ldots, \theta_d \in A_d) \times \ldots \times \Pi(\theta_{d-1} \in A_{d-1} | \theta_d \in A_d) \Pi(\theta_d \in A_d).
$$

Because prior beliefs may be more easily expressed when imagining a situation where other parameters have fixed values, one eventually succeeds in defining the prior for the high-dimensional model.

Example 3.2. Suppose that we measure pairs $(W_i, L_i)$, $1 \leq i \leq n$, where $W_i$ is the weight and $L_i$ is the length of the $i$-th draw from an i.i.d. sample of monkeys from a certain population. As our model for $(W, L)$, we choose model distributions that are products of Gamma distributions with shape parameter $k = 3$: $(W, L) \sim \Gamma(w/3, 3) \times \Gamma(l/3, 3)$, so that the model distribution $(W, L) | (w, l)$ has expectation $(w, l)$. To choose a prior for the parameter $(w, l)$, we note that a relatively tall monkey will also be relatively heavy. We express this through the specification of the prior: we assume that weight $w$ and length $l$ approximately follow the relationship $w = Kl^\alpha$, for some known $K > 0$ and a known exponent $\alpha$. In that case, given the length $l$, we specify (for some choice $\lambda > 0$),

$$
w | l \sim \Gamma(\lambda^{-1} Kl^\alpha, \lambda),
$$

so that the conditional prior expectation for $w | l$ is $K l^\alpha$. For $l$, we choose a marginal prior that depends on the same constant $\lambda$,

$$
l \sim \Gamma(\lambda^{-1} l, \lambda),
$$
3.1 Subjective priors

where \( \ell \) is an approximate mean length for a monkey from the population. The prior variances of \( l \) and \( w/l \) are equal to \( \ell^2/\lambda \) and \( K^2\ell^2\alpha/\lambda \) respectively, i.e. \( \lambda \) has the interpretation of being inversely proportional to uncertainty expressed in the prior (higher values of \( \lambda \) bias the prior (and hence also the posterior) more to the prior expectation \((\ell, \ell')\)).

The construction indicated here is reminiscent of that of a so-called hyperprior, which is discussed in section 3.5. The difference is, that components of \( \theta \) all occur in the definition of model distributions \( \mathcal{P}_\theta \), whereas hyperparameters do not. Note that it is important to choose a parametrization of the model in which the independence between \( \theta_i \) and \((\theta_1, \ldots, \theta_i-1)\), \textit{given} \((\theta_{i+1}, \ldots, \theta_d)\), is plausible for all \( i \geq 1 \).

In certain situations, the subjectivist has more factual information at his disposal when defining the prior for his analysis. In particular, if a probability distribution on the model reflecting the subjectivist’s “beliefs” can be found by other statistical means, it can be used as a prior. Suppose the statistician is planning to measure a quantity \( Y \) and infer on a model \( \mathcal{P} \); suppose also that this experiment repeats or extends an earlier analysis. From the earlier analysis, the statistician may have obtained a posterior distribution on \( \mathcal{P} \). For the new experiment, this posterior may serve as a prior.

Example 3.3. Let \( \Theta \rightarrow \mathcal{P} : \theta \mapsto \mathcal{P}_\theta \) be a parametrized model for \( i.i.d. X_1, X_2, \ldots, X_n \) with prior measure \( \Pi_1 : \mathcal{G} \rightarrow [0, 1] \). Let the model be dominated (see definition 1.3), so that the posterior \( \Pi_1(\cdot | X_1, \ldots, X_n) \) satisfies (2.14). Suppose that this experiment has been conducted, with the sample realised as \((X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)\). Next, consider a new, independent experiment in which a quantity \( X_{n+1} \) is measured (with the same model). As a prior \( \Pi_2 \) for the new experiment, we use the (realised) posterior of the earlier experiment, i.e. for all \( G \in \mathcal{G} \),

\[
\Pi_2(G) = \Pi_1(G | X_1 = x_1, \ldots, X_n = x_n).
\]

The posterior for the second experiment then satisfies:

\[
d\Pi_2(\theta | X_{n+1}) = \frac{p_\theta(X_{n+1}) d\Pi_1(\theta | X_1 = x_1, \ldots, X_n = x_n)}{\int_\Theta p_\theta(X_{n+1}) d\Pi_1(\theta | X_1 = x_1, \ldots, X_n = x_n)}
= \frac{p_\theta(X_{n+1}) \prod_{i=1}^n p_\theta(x_i) d\Pi_1(\theta)}{\int_\Theta p_\theta(X_{n+1}) \prod_{i=1}^n p_\theta(x_i) d\Pi_1(\theta)} \quad \text{(3.1)}
\]

The latter form is comparable to the posterior that would have been obtained if we had conducted a single experiment with an \( i.i.d. \) sample \( X_1, X_2, \ldots, X_{n+1} \) of size \( n+1 \) and prior \( \Pi_1 \). In that case, the posterior would have been of the form:
\[ \Pi(\cdot | X_1, \ldots, X_{n+1}) = \frac{\prod_{i=1}^{n+1} p_\theta(X_i) d\Pi_1(\theta)}{\int_\Theta \prod_{j=1}^{n+1} p_\theta(X_j) d\Pi_1(\theta)}, \quad \text{(3.2)} \]

i.e. the only difference is the fact that the posterior \( \Pi_1(\cdot | X_1 = x_1, \ldots, X_n = x_n) \) is realised. As such, we may interpret independent consecutive experiments as a single, interrupted experiment and the posterior \( \Pi_1(\cdot | X_1, \ldots, X_n) \) can be viewed as an intermediate result.

Clearly, there are other ways to obtain a distribution on the model that can be used as an informative prior. One example is the distribution that is obtained when a previously obtained frequentist estimator \( \hat{\theta} \) for \( \theta \) is subject to a procedure called the bootstrap \[83\]. Although the bootstrap gives rise to a distribution that is interpreted (in the frequentist sense) as the distribution of the estimator \( \hat{\theta} \) rather than \( \theta \) itself, a subjectivist may reason that the estimator provides him with the “expert knowledge” on \( \theta \) that he needs to define a prior on \( \Theta \).

### 3.2 Non-informative priors

Objectivist Bayesians agree with frequentists that the “beliefs” of the statistician analyzing a given measurement should play a minimal role in the methodology. Obviously, the model choice already introduces a bias, but rather than embrace this necessity and expand upon it like subjectivists, they seek to keep the remainder of the procedure unbiased. In particular, they aim to use priors that do not introduce additional information (in the form of prior “belief”) in the procedure. Subjectivists introduce their “belief” by concentrating prior mass in certain regions of the model; correspondingly, objectivists prefer priors that are “homogeneous” in an appropriate sense.

#### 3.2.1 Uniform priors

At first glance, one may be inclined to argue that a prior is objective (or non-informative) if it is uniform over the parameter space: if we are inferring on parameter \( \theta \in [0, 1] \) and we do not want to favour any part of the model over any other, we would choose the Lebesgue measure on \([0, 1]\) for a prior. Attempts to minimize the amount of subjectivity introduced by the prior therefore focus on uniformity (argumentation that departs from the Shannon entropy in discrete probability spaces reaches the same conclusion (see, for example, Ghosh and Ramamoorthi (2003) \[103\], p. 47)). The original references on Bayesian methods (e.g. Bayes (1763) \[13\], Laplace (1774) \[151\]) use uniform priors as well. But there are several
3.2 Non-informative priors

problems with this approach: first of all, one must wonder how to extend such rea-
soning when \( \theta \in \mathbb{R} \) (or any other unbounded subset of \( \mathbb{R} \)). In that case, Lebesgue
measure is infinite and cannot be normalized to a probability measure. Any attempt
to extend \( \Pi \) to such unbounded models as a probability measure would eventually
lead to inhomogeneity, \textit{i.e.} go at the expense of the unbiasedness of the procedure.

The compromise some objectivists are willing to make, is to relinquish the inter-
pretation that subjectivists give to the prior: they do not express any prior "degree
of belief" in \( A \in \mathcal{G} \) through the subjectivist statement that the (prior) probability
of finding \( \vartheta \in A \) equals \( \Pi(A) \). Although they maintain the Bayesian interpre-
tation of the posterior, they view the prior as a mathematical definition rather than a
philosophical concept. Then, the following definition can be made without further
reservations.

**Definition 3.1.** Given a model \((\Theta, \mathcal{G})\), a prior measure \( \Pi : \mathcal{G} \to [0, \infty] \) such that
\( \Pi(\Theta) = \infty \) is called an \textit{improper} prior.

Note that any the normalization factor for a prior cancels in the expression for the
posterior, \textit{c.f.} (2.4) or (2.6): any finite multiple of a (finite) prior is equivalent to
the original prior as far as the posterior is concerned. However, this argument does
not extend to the improper case: integrability problems or other infinities may ruin
the procedure, even to the point where the posterior measure becomes infinite or
ill-defined. So not just the philosophical foundation of the Bayesian approach is
lost, mathematical integrity of the procedure can no longer be guaranteed either.
When confronted with an improper prior, the entire procedure must be checked for
potential problems. In particular, one must verify that the posterior is a well-defined
probability measure.

**Remark 3.3.** Throughout these notes, whenever we refer to a prior measure, it is
 implied that this measure is a probability measure.

But even if one is willing to accept that objectivity of the prior requires that we
restrict attention to models on which “uniform” probability measures exist (\textit{e.g.} with
\( \Theta \) a bounded subset of \( \mathbb{R}^d \)), a more fundamental problem exists: the very notion
of uniformity is dependent on the parametrization of the model! To see this we
look at a model that can be parametrized in two ways and we consider the way
in which uniformity as seen in one parametrization manifests itself in the other
parametrization. Suppose that we have a \( d \)-dimensional parametric model \( \mathcal{P} \) with
two different parametrizations, on \( \Theta_1 \subset \mathbb{R}^d \) and \( \Theta_2 \subset \mathbb{R}^d \) respectively,
\[
\phi_1 : \Theta_1 \to \mathcal{P}, \quad \phi_2 : \Theta_2 \to \mathcal{P}
\] (3.3)
both of which are bijective. Assume that \( \mathcal{P} \) is a measurable space with \( \sigma \)-algebra \( \mathcal{G} \).
Require that \( \phi_1 \) and \( \phi_2 \) are Borel-to-\( \mathcal{G} \) measurable and assume that their inverses \( \phi_1^{-1} \)
and \( \phi_2^{-1} \) are measurable as well. Assuming that \( \Theta_1 \) is bounded, we consider the uni-
form prior \( \Pi_1 \) on \( \Theta_1 \), \textit{e.g.} the normalized Lebesgue measure \( \Pi_1(A) = \mu(\Theta_1)^{-1} \mu(A) \),
for all \( A \in \mathcal{B}_1 \). This induces a prior \( \Pi'_1 \) on \( \mathcal{P} \): for all \( B \in \mathcal{G} \),
\[
\Pi'_1(B) = (\Pi_1 \circ \phi_1^{-1})(B).
\] (3.4)
In turn, this induces a prior $\Pi''_1$ on $\Theta_2$: for all $C \in \mathcal{B}_2$,

$$\Pi''_1(C) = (\Pi'_1 \circ (\phi_2^{-1})^{-1})(C) = (\Pi'_1 \circ \phi_2)(C) = (\Pi_1 \circ (\phi_1^{-1} \circ \phi_2))(C).$$

Even though $\Pi_1$ is uniform, generically $\Pi''_1$ is not, because, effectively, we are mapping (a subset of) $\mathbb{R}^d$ to $\mathbb{R}^d$ by $\phi_2^{-1} \circ \phi_1 : \Theta_1 \rightarrow \Theta_2$. (Differentiable counterparts to such measurable re-coordinatizations are used extensively in differential geometry, where a smooth manifold is parametrized in various ways by sets of maps called charts.)

**Example 3.4.** Consider the model of all normal distributions centred on the origin with unknown variance between 0 and 1. We may parametrize this model in many different ways, but we consider only the following two:

$$\phi_1 : (0,1) \rightarrow \mathcal{P} : \tau \mapsto N(0, \tau), \quad \phi_2 : (0,1) \rightarrow \mathcal{P} : \sigma \mapsto N(0, \sigma^2).$$

(3.5)

Although used more commonly than $\phi_1$, parametrization $\phi_2$ is not special in any sense: both parametrizations describe exactly the same model. Now, suppose that we choose to endow the first parametrization with a uniform prior $\Pi_1$, equal to the Lebesgue measure $\mu$ on $(0,1)$. By (3.4), this induces a prior on $\mathcal{P}$. Let us now see what this prior looks like if we consider $\mathcal{P}$ parametrized by $\sigma$: for any constant $C \in (0,1)$ the point $N(0,C)$ in $\mathcal{P}$ is the image of $\tau = C$ and $\sigma = \sqrt{C}$, so the relation between $\tau$ and corresponding $\sigma$ is given by

$$\tau(\sigma) = (\phi_2^{-1} \circ \phi_1)(\sigma) = \sigma^2.$$

Since $\Pi_1$ equals the Lebesgue measure, we find that the density of $\Pi''_1$ with respect to the Lebesgue measure equals:

$$\pi''_1(\sigma) d\sigma = \pi_1(\tau(\sigma)) \left| \frac{d\tau}{d\sigma} \right| (\sigma) d\sigma = 2\sigma d\sigma.$$

This density is non-constant and we see that $\Pi''_1$ is non-uniform. In a subjectivist sense, the prior $\Pi''_1$ places higher prior “belief” on values of $\sigma$ close to 1 than on values close to 0.

From the above argument and example 3.4, we see that uniformity of the prior is entirely dependent on the parametrization: what we call “uniform” in one parametrization, may be highly non-uniform in another. Consequently, what is deemed “objective” in one parametrization may turn out to be highly subjective in another.

What matters, is the model $\mathcal{P}$ itself and not its parametrization in terms of any specific parameter. The parametrization is a mere choice made by the statistician analysing the problem. Any statistical concept that depends on the parametrization is flawed from the outset: through $\mathcal{P}$ and only through $\mathcal{P}$ do the parameters $\sigma$ and $\tau$ have any bearing on (the law of) the observation in example 3.4. If we could define what is meant by uniformity on the model $\mathcal{P}$ itself, instead of on its parametrizing spaces, one would obtain a viable way to formalize objectivity. But spaces of probability measures do not have an intrinsic notion of uniformity (like translation-
invariance of Lebesgue measure on $\mathbb{R}^d$, or more generally, left-invariance of the Haar measure on locally compact topological groups).

### 3.2.2 Jeffreys prior and reference priors

Once it is clear that uniformity on any parametrizing space does not have intrinsic meaning in the model $\mathcal{P}$, the very definition of objectivity in terms of uniformity of the prior is void. A subjectivist can use any parametrization to formulate his prejudice but an objectivist has to define his notion of “objectivity” regardless of the parametrization used. Therefore, the emphasis is shifted: instead of looking for uniform priors, we look for priors that are well-defined on $\mathcal{P}$ and declare them objective. For differentiable parametric models, a construction from Riemannian geometry can be used to define a parameterisation-independent prior (see Jeffreys (1946), (1961) [127, 128]) if we interpret the Fisher information as a Riemannian metric on the model (as first proposed by Rao (1945) [191] and extended by Efron (1975) [82]; for an overview, see Amari (1990) [4]) and use the square-root of its determinant as a density with respect to the Lebesgue measure.

**Definition 3.2.** Let $\Theta \subset \mathbb{R}^d$ be open and let $\mathcal{P}$ be a dominated model with identifiable, differentiable parametrization $\Theta \rightarrow \mathcal{P}$. Assume that for every $\theta \in \Theta$, the score-function $\dot{\ell}_\theta$ is square-integrable with respect to $P_\theta$. The Jeffreys prior $\Pi$ has the square root of the determinant of the Fisher information $I_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T$ as its density with respect to the Lebesgue measure on $\Theta$:

$$d\Pi(\theta) = \sqrt{\det(I_\theta)} d\theta. \quad (3.6)$$

The expression for Jeffreys prior has the appearance of being highly dependent on the parametrization of $\mathcal{P}$ in terms of $\theta \in \Theta$. However, the form (3.6) of this prior is the **same in any** parametrization (a property referred to sometimes as *co-variance* with respect to diffeomorphisms). In other words, no matter which parametrization we use to calculate $\Pi$ in (c.f. (3.6)), the induced measure $\Pi'$ on $\mathcal{P}$ is always the same one. As such, Jeffreys prior is a measure defined on $\mathcal{P}$ rather than a parametrization-dependent measure.

**Example 3.5.** We calculate the density of Jeffreys prior in the normal model of example 3.4. The score-function with respect to the parameter $\sigma$ in parametrization $\phi_2$ of $\mathcal{P}$ is given by:

$$\dot{\ell}_\sigma(X) = \frac{1}{\sigma} \left( \frac{X^2}{\sigma^2} - 1 \right).$$

The Fisher information (which is a $1 \times 1$-matrix in this case), is then given by:

$$I_\sigma = P_\sigma \dot{\ell}_\sigma \dot{\ell}_\sigma = \frac{1}{\sigma^2} P_\sigma \left( \frac{X^2}{\sigma^2} - 1 \right)^2 = \frac{2}{\sigma^2}$$

Therefore, the density for Jeffreys prior $\Pi$ takes the form
for all $\sigma \in \Theta_2 = (0, 1)$. A similar calculation using the parametrization $\phi_1$ shows that, in terms of the parameter $\tau$, Jeffreys prior takes the form:

$$d\Pi(\tau) = \frac{1}{\sqrt{2\tau}} d\tau,$$

for all $\tau \in \Theta_1 = (0, 1)$. That both densities give rise to the same measure on $\mathcal{P}$ is the assertion of the following lemma.

**Lemma 3.1. (Parameterization-independence of Jeffreys prior)**

Consider the parametrizations $\phi_1$ and $\phi_2$ of (3.3) and assume that they satisfy the conditions of definition 3.2. In addition, we require that the map $\phi_1^{-1} \circ \phi_2 : \Theta_2 \to \Theta_1$ is differentiable. Then the densities (3.6), calculated in coordinates $\phi_1$ and $\phi_2$ induce the same measure on $\mathcal{P}$, Jeffreys prior.

**Proof.** Since the Fisher information can be written as:

$$I_{\theta_1} = P_{\theta_1}(\ell_{\theta_1} \ell_{\theta_1}^T),$$

and the score $\ell_{\theta_1}(X)$ is defined as the gradient of $\theta_1 \mapsto \log p_{\theta_1}(X)$ with respect to $\theta_1$, the change of parametrization $\theta_1(\theta_2) = (\phi_1^{-1} \circ \phi_2)(\theta_2)$ induces a transformation of the form,

$$I_{\theta_2} = S_{1,2}(\theta_2) I_{\theta_1(\theta_2)} S_{1,2}(\theta_2)^T,$$

on the Fisher information matrix, where $S_{1,2}(\theta_2)$ is the total derivative of $\theta_2 \mapsto \theta_1(\theta_2)$ in the point $\theta_2$ of the model. Therefore,

$$\sqrt{\det I_{\theta_2}} d\theta_2 = \sqrt{\det(S_{1,2}(\theta_2) I_{\theta_1(\theta_2)} S_{1,2}(\theta_2)^T)} d\theta_2$$

$$= \sqrt{\det(S_{1,2}(\theta_2))^2} \sqrt{\det(I_{\theta_1(\theta_2)})} d\theta_2$$

$$= \sqrt{\det(I_{\theta_1(\theta_2)})} \left| \det(S_{1,2}(\theta_2)) \right| d\theta_2 = \sqrt{\det(I_{\theta_1})} d\theta_1$$

i.e. the form of the density of the Jeffreys prior is such that reparametrization leads exactly to the Jacobian for the transformation of $d\theta_2$ to $d\theta_1$.

Ultimately, the above construction derives from the fact that the Fisher information $I_0$ (or, in fact, any Hessian of a twice-differentiable convex function) can be viewed as a Riemann metric on the “smooth manifold” $\mathcal{P}$. The definition of a measure with Lebesgue density (3.6) is then a standard construction of a measure on the manifold in differential geometry.

**Example 3.6.** To continue with the normal model of examples 3.4 and 3.5, we note that $\sigma(\tau) = \sqrt{\tau}$, so that $d\sigma/d\tau(\tau) = 1/(2\sqrt{\tau})$. As a result,
\[ \sqrt{\det I_{\theta_2}} d\theta_2 = \frac{\sqrt{\sigma}}{\sigma} d\sigma = \frac{\sqrt{\sigma}}{\sigma(\tau)} \left| \frac{d\sigma}{d\tau} \right| (\tau) d\tau = \frac{1}{\sqrt{2\tau}} d\tau = \sqrt{\det(I_{\theta_1})} d\theta_1, \]

which verifies the assertion of lemma 3.1 explicitly.

Other constructions and criteria for the construction of non-informative priors exist: very popular is the use of so-called *reference priors*, as introduced in Lindley (1956) [168] and rediscovered in Bernardo (1979) [24] (see also Berger and Bernardo (1992) [20]). By defining principle, a reference prior is required to maximize the Kullback-Leibler divergence between prior and posterior. To motivate this condition, we have to look at information theory, from which Kullback-Leibler divergence emerges as a (popular but by no means unique) way to quantify the notion of the “amount of information” contained in a probability distribution. Sometimes called the *Shannon entropy*, the Kullback-Leibler divergence of the counting measure with respect to a distribution \( P \) in discrete probability spaces,

\[
S(P) = \sum_{\omega \in \Omega} p(\omega) \log(p(\omega)),
\]

can be presented as such convincingly (see Bolzmann (1895, 1898) [40], Shannon (1948) [203]). For lack of a default dominating measure, the argument does not extend formally to continuous probability spaces but is generalized nevertheless.

**Definition 3.3.** A *reference prior* \( \Pi \) on a dominated, parametrized model \( \Theta \to \mathcal{P} : \theta \mapsto P_\theta \) for an observation \( Y \) is a maximizer of the so-called *Lindley entropy*,

\[
S_L = \int \int \log\left( \frac{\pi(\theta | Y = y)}{\pi(\theta)} \right) dP(\theta | Y = y) d\Pi(\theta),
\]

which measures the prior-predictive expectation of the Kullback-Leibler divergence of the prior with respect to the posterior.

Note that Bayes’s Rule (2.4) (see also exercise ??) allows us to rewrite the Lindley entropy in the form,

\[
S_L = \int \int \log\left( \frac{\pi(\theta | Y = y)}{\pi(\theta)} \right) dP_{\theta}(y) d\Pi(\theta),
\]

Usually, the derivation of a reference prior [24] is performed in the large-sample limit where the posterior for a sufficiently smooth model becomes asymptotically normal, in accordance with the Bernstein-von Mises theorem of chapter 4. For certain models, Jeffreys prior emerges as a reference prior.

For an overview of various objective methods of constructing priors, the reader is referred to Kass and Wasserman (1995) [132]. When using non-informative priors, however, the following general warning should be heeded.

**Remark 3.4.** In many models, non-informative priors, including Jeffries prior and reference priors, are improper.
3.3 Hierarchical priors

Consider again the problem of estimating the mean of a single, normally distributed observation $Y$ with known variance. The model consists of all normal distributions $P_0 = N(\theta, \sigma^2)$, where $\theta \in \mathbb{R}$ is unknown and $\sigma^2 > 0$ is known. Imposing a normal prior on the parameter $\theta$, $\Pi = N(0, \tau^2)$, for some choice of $\tau^2 > 0$, we calculate the posterior distribution,

$$
\Pi(\theta \in A | Y) = N\left( \frac{\tau^2 Y}{\sigma^2 + \tau^2}, \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2} \right) (A),
$$

for every $A \in \mathcal{B}$. The posterior mean, a point-estimator for $\theta$, is then given by,

$$
\hat{\theta}(Y) = \frac{\tau^2 Y}{\sigma^2 + \tau^2}.
$$

As long as expert knowledge is available in some sufficiently strong form, the subjectivist choice for a certain value of $\tau^2$ is well-motivated. However in situations where no prior belief or information on the parameter $\theta$ is available, or if the parameter itself does not have a clear interpretation, the subjectivist has no answer. Yet a choice for $\tau^2$ is required! One may express ignorance about $\tau^2$ either by choosing a prior on objectivist grounds, or by considering more and more homogeneous (but still normal) priors by means of the limit $\tau \to \infty$.

**Remark 3.5.** From a statistical perspective there exists a better answer to the question regarding $\tau^2$: since $\tau$ is not known, estimate its value from the data.

In this section and the next, we consider this answer from the Bayesian and from the frequentist’s angle respectively, giving rise to procedures known as hierarchical Bayesian modelling and empirical Bayesian estimation.

3.3.1 Hyperparameters and hyperpriors

First we turn to the Bayesian answer to remark 3.8: the Bayesian views a parameter to be estimated as just another random variable in the probability model. In case we want to estimate the parameter for a family of priors, then that parameter is to be included in the probability space from the start. Going back to the example with which we started this section, this means that we still use normal distributions $P_0 = N(\theta, \sigma^2)$ to model the uncertainty in the data $Y$, supply $\theta \in \mathbb{R}$ with a prior $\Pi_1 = N(0, \tau^2)$ and then proceed to choose a another prior $\Pi_2$ for $\tau^2 \in (0, \infty)$:

$$
Y|\theta, \tau^2 = Y|\theta \sim P_0 = N(\theta, \sigma^2), \quad \theta|\tau^2 \sim \Pi_1 = N(0, \tau^2), \quad \tau^2 \sim \Pi_2,
$$

Note that the parameter $\tau^2$ has no direct bearing on the model distributions: conditionally on $\theta$, $Y|\theta$ is independent of $\tau^2$. In a sense, the hierarchical Bayesian
approach to prior choice combines subjective and objective philosophies: whereas
the subjectivist will make a definite, informed choice for $\tau^2$ and the objectivist will
keep himself as uncommitted as possible by striving for uniformity, the choice for
a hierarchical prior expresses uncertainty about the value of $\tau^2$ in the form of a
probability distribution $\Pi_2$. As such, the hierarchical Bayesian approach allows for
intermediate prior choices: if $\Pi_2$ is chosen highly concentrated around one point in
the model, resembling a degenerate measure, the procedure will be close to subjec-
tive; if $\Pi_2$ is spread widely and is far from degenerate, the procedure will be less
biased and closer to objective. Additionally, the flexibility gained through introduc-
tion of $\Pi_2$ offers a much wider freedom of modelling. In particular, we may add
several levels of modelled parameter uncertainty to build up a hierarchy of priors
for parameters of priors. Such structures are used to express detailed subjectivist
beliefs, much in the way graphical models are used to build intricate dependency
structures for observed data. The origins of the hierarchical approach go back, at
least, to Lindley and Smith (1972) [169].

**Definition 3.4.** Let the data $Y$ be random in $(\mathcal{Y}, \mathcal{B})$. A hierarchical Bayesian model
for $Y$ consists of a collection of probability measures $\mathcal{P} = \{P_0 : \theta \in \Theta_0\}$, with
$(\Theta_0, \mathcal{G}_0)$ measurable and endowed with a prior $\Pi : \mathcal{G}_0 \to [0, 1]$ built up in the fol-
lowing way: for some $k \geq 1$, we introduce measurable spaces $(\Theta_i, \mathcal{G}_i)$,
$i = 1, 2, \ldots, k$ and conditional priors
$G_i \times \Theta_{i+1} \to [0, 1] : (G, \theta_{i+1}) \mapsto \Pi_i(G|\theta_{i+1})$,
for $i = 1, \ldots, k-1$ and a marginal $\Pi_k : \Theta_k \to [0, 1]$ on $\Theta_k$. The prior for the original
parameter $\theta$ is then defined by,
$$
\Pi(\theta \in G) = \int_{\Theta_1 \times \ldots \times \Theta_k} \Pi_0(\theta \in G|\theta_1) d\Pi(\theta_1|\theta_2) \ldots d\Pi(\theta_{k-1}|\theta_k) d\Pi_k(\theta_k),
$$
for all $G \in \mathcal{G}_0$. The parameters $\theta_1, \ldots, \theta_k$ and the priors $\Pi_1, \ldots, \Pi_k$ are called hyper-
parameters and their hyperpriors.

Definition 3.4 is very close to the general Bayesian model that incorporates all pa-
rameters $(\theta, \theta_1, \ldots, \theta_k)$ as modelling parameters. What distinguishes hierarchical
modelling from the general situation is the dependence structure imposed on the
parameters. The parameter $\theta$ is distinct from the hyperparameters by the fact that
conditional on $\theta$, the data $Y$ is independent of all hyperparameters $\theta_1, \ldots, \theta_k$. This
distinction is repeated at higher levels in the hierarchy, i.e. levels are separate from
one another through the conditional independence of $\theta_i|\theta_{i+1}$ from $\theta_{i+2}, \ldots, \theta_k$.

**Remark 3.6.** The hierarchy indicated in definition 3.4 inherently loses interpretabil-
ity as we ascend in level. One may be able to give a viable interpretation to the
parameter $\theta$ and to the hyperparameter $\theta_1$, but higher-level parameters $\theta_2, \theta_3, \ldots$
become harder and harder to understand heuristically. Since the interpretation of the
hierarchy requires a subjective motivation of the hyperpriors, interpretability of each
level is imperative, or left as a non-informative choice. In practice, Bayesian hierar-
3 Choice of the prior

Hierarchical models are rarely more than two levels deep ($k = 2$) and the last hyperprior $\Pi_k$ is often chosen by objective criteria.

### 3.3.2 Hierarchical prior construction in an example

This subsection considers a more-or-less practical example of a hierarchical prior in some detail.

**Example 3.7.** We observe the number $Y$ of surviving offspring from a bird’s litter and aim to estimate the number of eggs the bird has laid: the bird lays $N \geq 0$ eggs, distributed according to a Poisson distribution with parameter $\lambda > 0$. For the particular species of bird in question, the Poisson rate $\lambda$ is not known exactly: the uncertainty in $\lambda$ can be modelled in many ways; here we choose to model it by a Gamma-distribution $\Gamma(\alpha, \beta)$, where $\alpha$ and $\beta$ are chosen to reflect our imprecise knowledge of $\lambda$ as well as possible. Each of the eggs then comes out, producing a viable chick with known probability $p \in [0, 1]$, independently. Hence, the total number $Y$ of surviving chicks from the litter is distributed according to a binomial distribution, conditional on $N$,

$$Y|N \sim \text{Bin}(N, p), \quad N|\lambda \sim \text{Poisson}(\lambda), \quad \lambda \sim \Gamma(\alpha, \beta).$$

The posterior distribution is obtained as follows: conditional on $N = n$, the probability of finding $Y = k$ is binomial,

$$P(Y = k|N = n) = \binom{n}{k} p^k (1 - p)^{n-k},$$

so Bayes’s rule tells us that the posterior is given by:

$$P(N = n|Y = k) = \frac{P(N = n)}{P(Y = k)} \binom{n}{k} p^k (1 - p)^{n-k}.$$

Since $\sum_{n \geq 0} P(N = n|Y = k) = 1$ for every $k$, the marginal $P(Y = k)$ (that is, the denominator or normalization factor for the posterior given $Y = k$) can be read off once we have the expression for the numerator. We therefore concentrate on the marginal for $N = n, (n \geq 0)$:

$$P(N = n) = \int_{\mathbb{R}} P(N = n|\lambda) p_{\alpha, \beta}(\lambda) d\lambda = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty \frac{e^{-\lambda} \lambda^n}{n!} \lambda^{\alpha-1} e^{-\lambda/\beta} d\lambda.$$

The integral is solved using the normalization constant of the $\Gamma(\alpha + n, \beta/\beta + 1)$-distribution:

$$\int_0^\infty e^{-\lambda} \frac{\beta^{\beta+1}}{\Gamma(\beta)} \lambda^{\alpha+n-1} d\lambda = \Gamma(\alpha + n) \left(\frac{\beta}{\beta + 1}\right)^{\alpha+n}.$$

Substituting and using the identity $\Gamma(\alpha + 1) = \alpha \Gamma(\alpha)$, we find:
3.4 Empirical prior choice

\[ P(N = n) = \frac{\Gamma(\alpha + n)}{\Gamma(\alpha)} \frac{1}{n!} \frac{1}{\beta} \left( \frac{\beta}{\beta + 1} \right)^{\alpha + n} \]

\[ = \frac{1}{n!} \left( \frac{\beta}{\beta + 1} \right)^n \frac{1}{(\beta + 1)^\alpha} \prod_{l=1}^{n} (\alpha + l - 1) \]  

Although not in keeping with the subjective argumentation of the introduction to this example, for simplicity we consider \( \alpha = \beta = 1 \) and find that in that case, 

\[ P(N = n) = (1/2)^{n+1}. \]

The posterior for \( N = n \) given \( Y = k \) then takes the form:

\[ P(N = n|Y = k) = \frac{1}{2^n} \binom{n}{k} p^k (1-p)^{n-k} / \sum_{m=0}^{\infty} \frac{1}{2^m} \binom{m}{k} p^k (1-p)^{m-k}. \]

The eventual form of the posterior illustrates that the hierarchy contributes only to the construction of the prior: in case we choose \( \alpha = \beta = 1 \), the posterior we find from the hierarchical approach does not differ from the posterior that we would have found if we had started from the model that incorporates a geometric prior for \( N \),

\[ Y|N \sim \text{Bin}(N, p), \quad N \sim \text{Geo}(1/2). \]

Indeed, even if we leave \( \alpha \) and \( \beta \) free, the marginal distribution for \( N \) we found in (3.10) is none other than the prior (3.9) for this problem.

The hierarchical approach to prior construction allows for greater freedom and a more solid foundation to motivate the choice for certain prior over other possibilities. This point is all the more significant in light of remark 3.1: the motivation of a subjectivist choice for the prior is part of the statistical analysis rather than an external aspect of the procedure. Hierarchical Bayesian modelling helps to refine and justify motivations for subjectivist priors.

3.4 Empirical prior choice

More unexpected is the frequentist perspective on remark 3.8, which goes by the general name empirical Bayes: point-estimate \( \tau^2 \) first based on available data and then perform the Bayesian analysis with the estimate as a "plug-in" for the unknown \( \tau^2 \). Critical notes can be placed with the philosophical foundations for this practice, since it appears to combine the methods of two contradictory schools of statistics. Be that as it may, the method is used routinely based on its practicality: eventually, the justification comes from the subjectivist who does not reject frequentist methods to obtain expert knowledge on his parameters, as required in his own paradigm.

Any common-sense statistician will tell you that it is crucial for the statistical analysis that one first obtains a certain feeling for the statistical problem by inspection of the data, before making decisions on how to analyse it. This alternative form
of “expert knowledge” is at odds with another common-sense practical rule: good statistical practice dictates that one may not use the data to decide which statistical method to use for the analysis of the same data. The rationale behind this dictum is the potential for the introduction of bias in the analysis. If we take both points seriously, the choice for a prior (read, the point-estimate for $\tau^2$) should be data-dependent, but not depend on the same data $Y$ that is to be used later to derive the posterior distribution for $\theta|Y$.

Ideally, one splits the sample in two independent parts, making any data-driven choice for a prior based on one sub-sample and performing the analysis proper with the other. Independence between the sub-samples guarantees the absence of bias. If the data-set consists of an i.i.d. sample $X_1, \ldots, X_n$, for example, one is presented with the question which fraction of the data is to be used for estimation of hyperparameters like $\tau^2$ and how much data should be preserved for the calculation of the posterior. Cross-validation, in which splitting of the sample is done many times over, appears to provide a valid method to make such choices.

However this warning is ignored almost customarily: it is common practice to calculate a posterior $\Pi(\cdot|Y, \tau)$ for fixed values of a hyperparameter $\tau^2$, find out that it is biased or otherwise flawed and subsequently use “plug-in” estimates $\hat{\tau}(Y)$ based on the same data to correct. The quantity $\Pi(\cdot|Y, \hat{\tau}(Y))$ is not a posterior, however: although expression (2.6) for the posterior suggests immediate generalization to,

$$\Pi(\theta \in B | Y, \hat{\tau}(Y)) = \left[ \int_B p_\theta(Y) d\Pi_{\tau|Y}(\theta) \right] / \left[ \int_{\Theta} p_\theta(Y) d\Pi_{\tau|Y}(\theta) \right], \quad (3.11)$$

the definition of the posterior (2.4) shows that there is no room for $Y$-dependence in the prior choice. The random measure (3.11) may still be a probability measure on the model, but it is not the conditional distribution of $\theta|Y$ in any way, and cannot be viewed as an implementation of Bayes’s rule any more. If the sample is split in two independent sub-samples, this objection is lifted, so sample-splitting saves the notion of a posterior with data-driven prior choice, and, as such, empirical Bayes methods.

In line with the concluding remarks of section 3.1, a natural split of the sample occurs in the following situation: when we analyse data pertaining to an individual from a larger population and it is reasonable to assume that the prior can be inferred from the population, then one may estimate parameters like $\tau^2$ above from population data and use the estimates in the prior for the individual.

A more sophisticated form of applying empirical Bayesian methods is the use of hyperparameters estimated by maximum-likelihood, applied to the prior predictive distribution. Recall that the prior expectation of the distribution of the data in the subjectivist Bayesian formulation (see definition 2.4) predicts a distribution for the marginal distribution of the data. This prediction may be reversed to decide which value for the hyperparameter leads to the best explanation of the observed data, where our notion of “best” is based on the likelihood principle.

Denote the data by $Y$ and assume that it takes its values in a measurable space $(\mathcal{Y}, \mathcal{B})$. Denote the model by $\mathcal{P} = \{P_\theta : \theta \in \Theta_0\}$. Consider a family of priors
parametrized by a hyperparameter $\eta \in H$, $\{\Pi_\eta : \eta \in H\}$. For every $\eta$, the prior predictive distribution $P_\eta$ is given by:

$$P_\eta(A) = \int_\Theta P_{\theta}(A) d\Pi_\eta(\theta),$$

for all $A \in \mathcal{B}$, i.e. we obtain a new model for the observation $Y$, given by $\mathcal{P}' = \{P_\eta : \eta \in H\}$, contained in the closed convex hull of the original model (see proposition 2.3). Note that this new model is parametrized by the hyperparameter; hence if we close our eyes to the rest of the problem and we follow the maximum-likelihood procedure for estimation of $\eta$ in this new model, we find the value of the hyperparameter that best explains the observation $Y$. Assuming that the model $\mathcal{P}'$ is dominated, with densities $\{p_\eta : \eta \in H\}$, the maximum-likelihood estimate is found as the point $\hat{\eta}(Y) \in H$ such that

$$p_\eta(Y) = \sup_{\eta \in H} p_\eta(Y).$$

under the assumptions of existence and uniqueness, by the usual methods for maximum-likelihood estimation.

**Definition 3.5.** The estimator $\hat{\eta}(Y)$ is called the **ML-II estimator**, provided it exists and is unique.

**Remark 3.7.** There is one caveat that applies to the ML-II approach: in case the data $Y$ consists of an i.i.d.-distributed sample, the prior predictive distribution describes the sample as exchangeable, but not i.i.d.! Hence, comparison of prior predictive distributions with the data suffer from the objection raised in remark 2.1. The frequentist who assumes that the true, underlying distribution $P_{n0}$ of the sample is i.i.d., has to keep in mind that the ML-II model is misspecified.

### 3.4.1 Model selection with empirical methods

A situation where empirical Bayes methods are often used, is in model selection: suppose that there are several models $\mathcal{P}_1, \mathcal{P}_2, \ldots$ with priors $\Pi_1, \Pi_2, \ldots$, each of which may serve as a reasonable explanation of the data, depending on an unknown parameter $K \in \{1, 2, \ldots\}$. The choice to use model-prior pair $(\mathcal{P}_K, \Pi_K)$ in the determination of the posterior is made after estimation of $K$. Where the Bayesian chooses a hyperprior for the hyperparameter $K$, frequentist ways to estimate $K$ leads to empirical Bayes methods.

**Example 3.8.** Consider the situation where we are provided with a specimen from a population that is divided into an unknown number of classes. Assume that all we know about the classes is that they occur with equal probabilities in the population. The particular class of our specimen remains unobserved. We perform a real-valued measurement $Y$ on the specimen, which is normally distributed conditionally on its
class, with variance 1 and unknown mean $\mu_k \in \mathbb{R}$ that depends on the class $k \geq 1$. Given that the number of classes $K$ equals $k$, $Y$ is distributed according to a discrete mixture of normal distributions,

$$Y|K, \mu \sim \mathcal{P}_K; \mu_1, \ldots, \mu_K = \frac{1}{K} \sum_{i=1}^{K} N(\mu_i, 1),$$

where $\mu = (\mu_1, \ldots, \mu_k) \in \mathbb{R}^k$ are unknown and satisfy $\mu_1 < \ldots < \mu_k$ (to keep the parameterizations in terms of $\mu$ identifiable). For every $k \geq 1$, we have a model of the form,

$$\mathcal{P}_k = \{P_k; \mu_1, \ldots, \mu_k : (\mu_1, \ldots, \mu_k) \in \mathbb{R}^k, \mu_1 < \ldots < \mu_k \}$$

Each of these models can be endowed with a prior $\Pi_k$ on $\mathbb{R}^k$, for example, by drawing an i.i.d. sample $\mu' = (\mu'_1, \ldots, \mu'_k)$ from a standard normal distribution:

$$\mu' \sim \Pi_k = N(0, I_k),$$

and then order the result, $\mu = (\mu'_1, \ldots, \mu'_k)$.

At this point, a Bayesian would choose a hyperprior $\Pi_2$ for the discrete hyperparameter $K \geq 1$ and proceed to calculate the posterior using all models $\mathcal{P}_k$, weighed by the prior masses $\Pi_2(K = k)$ for all $k \geq 1$. Alternatively, the Bayesian can use posterior odds to make a decision as to which value of $K$ to use, reducing the analysis to a selected, or estimated value for $K$. The frequentist estimates $K$ to select one of the models $\mathcal{P}_K$. Inspection of the data may reveal which number of classes is most appropriate, if one observes clearly separated peaks in the observations. Otherwise, many frequentist procedures (called clustering methods) exist to estimate $K$, for instance from a larger population of specimens. After we have an estimate $\hat{K}$ for $K$ that is independent of the data proper, we are in a position to calculate the posterior for $\mu$ based on $(\mathcal{P}_{\hat{K}}, \Pi_{\hat{K}})$.

There are two remarks to be made with regard to the estimation of $K$ from a larger population of specimens: first of all, maximization of the likelihood with an unbounded number of classes picks a number of classes equal to (or in the order of) the sample-size, simply because assigning each data-point its own class leads to the largest likelihood function. A similar phenomenon arises in regression, where it is called over-fitting: if we allow regression polynomials of arbitrary degree, the MLE will fit the data perfectly by choosing a polynomial of degree equal to the sample-size. The fit is perfect, residuals are zero and any associated measure for quality (like the $R^2$) will reflect this. But we are no longer doing statistics, because we are not distinguishing signal from noise (in fact, we have included all noise in the fit). Like for the degree of the regression polynomial, one would like to have a sensible way to regularize the estimate for the number of clusters $K$, and then estimate $\mu_1, \ldots, \mu_k$.

In such questions of model selection, penalized likelihood criteria are employed which favour smaller choices for $K$ over larger ones. Note that it is not clear, neither intuitively nor mathematically, how the penalty should depend on $K$ (and things like the sample-size $n$), nor which proportionality between penalty and likelihood
is appropriate. A well-known standard choice comes in the form of the so-called Akaike information criterion (AIC) for model selection [202]: it argues for maximization of the (\(k\)-dependent) likelihood minus twice the dimension of the \(k\)'th parameter space (here \(2k\)), motivated from information theory and large sample sizes. The Bayesian faces the same problem when he chooses a prior for \(K\): if he assigns too much prior weight to the higher-dimensional models, his estimators (or, equivalently, the bulk of the resulting posterior’s mass) will get the chance to “run off” to infinity with growing samplesize, indicating inconsistency from over-fitting. The so-called Bayesian information criterion (BIC) [202] weighs the AIC penalty by the logarithm of the sample size, motivated by the Bernstein-von Mises limit of chapter 4, maximizing likelihood minus \(2k \log(n)\). Indeed, the correspondence between the frequentist’s necessity for a penalty in maximum-likelihood methods on the one hand, and the Bayesian’s need for a prior expressing sufficient bias for the lower-dimensional model choices on the other, is explained in remark 2.6.

It is difficult to indicate which regularization method is preferred, as long as the argument is to be made for each sample-size \(n \geq 1\) separately. Matters organise themselves in the large-sample limit, where one would like to select the model consistently: if we observe larger and larger i.i.d. samples \(X^n = (X_1, X_2, \ldots, X_n)\), with each \(X_i\) distributed like \(Y\) above marginally, for some unobserved value \(K = k\), we would like to have a model selection method that selects the correct number of clusters \(k\) with probability growing to one as \(n \to \infty\).

Example 3.9. In part II, we shall see that in the model of example 3.8, consistent selection of \(K\) is possible, if we restrict the model to consist of an upper-bounded number \(f\) clusters and the locations \(\mu_i\) all lie in a fixed, compact subset of \(\mathbb{R}\), at some fixed minimal distance from one-another. We can summarize these requirements in terms of a single integer \(M \geq 1\) such that, \(1 \leq K \leq M\) and,

\[ \mathcal{P}_l = \left\{ p_{l_1, \ldots, l_M} : (\mu_1, \ldots, \mu_l) \in [-M, M]^l, \mu_1 < \ldots < \mu_l, \mu_{l+1} - \mu_l > 1/M \right\}. \]

Then any convex combination of priors \(\Pi_l\) that are of full support on their respective submodels \(\mathcal{P}_l\),

\[ \Pi = \sum_{l=1}^{M} \pi_l \Pi_l, \]

for \(0 < \pi_1, \ldots, \pi_M < 1\) such that \(\sum_{l=1}^{M} \pi_l = 1\), will lead to a sequence of posteriors on \(\mathcal{P} = \bigcup_{l=1}^{M} \mathcal{P}_l\) that concentrate all mass in the correct component \(\mathcal{P}_k\) with probability growing to one as \(n \to \infty\); consequently, posterior odds can be used to model select consistently. The restriction that all classes are represented in equal numbers in the population is not necessary (although consistent selection with posterior odds requires a minimal value \(1/M\) for each of the fractions). And the question also arises, what if we use upper bounds \(M_n\) that grow larger with growing sample-size \(n\)? How fast can \(M_n\) go to infinity, while still achieving a consistent posterior?
3.4.2 Bias and the James-Stein phenomenon

What is clear in the clustering and regression examples, is that model selection can also be viewed as correction of a bias inherent to our estimation method, a bias towards models with a high number of clusters or high order of a regression polynomial. Such views are particularly fruitful in the Bayesian case, because often, Bayesian point estimators that are expectations with respect to the posterior like the posterior predictive distribution \( P_{\Pi|Y} \) (but also the posterior mean of a parameter) can be decomposed into an unbiased, consistent estimate \( \hat{P}_n(Y) \) and a bias ascribed to the prior, like the prior predictive distribution \( \Pi \) (but also the prior mean of a parameter),

\[
P_{\Pi|Y} = (1 - \lambda_n) \hat{P}_n(Y) + \lambda_n \Pi.
\]

Refer to decomposition 6.13 for an example in the context of the Dirichlet process prior and posterior. If \( \lambda_n \to 0 \) the posterior predictive distribution follows the unbiased, frequentist estimate asymptotically (and will be consistent if \( \hat{P}_n \) is). There are also cases where \( \lambda_n \) does not go to zero and bias persists in the limit, leading to inconsistency of the Bayesian estimator.

With empirical Bayes methods to estimate which value of \( \alpha \) is most appropriate, however, the inherent prior bias in (3.12) may be repairable: if we use the data to de-bias the prior predictive distribution \( \Pi \) itself, such problems can be mitigated or eliminated altogether.

Example 3.10. (Univariate normal mean)

Consider the simpler case of \( X_1, X_2, \ldots \) that are i.i.d. \( N(\theta, \sigma^2) \)-distributed (with known \( \sigma^2 > 0 \)) and a normal, non-central prior for the parameter \( \theta \in \mathbb{R} \), that is, \( \Pi = N(\alpha, \tau^2) \) for some \( \alpha \in \mathbb{R} \) and \( \tau^2 > 0 \). The normal family is conjugate for this model, so the posterior distribution is a normal distribution and it is easily seen that the posterior mean is,

\[
\theta_n(X_1, \ldots, X_n) = \left( \frac{\sigma^2 \tau^2}{n \tau^2 + \sigma^2} \right) \left( \frac{\alpha \sigma^2 + \sum_{i=1}^n X_i}{\sigma^2} \right) = \left( 1 + \frac{\sigma^2}{n \tau^2} \right)^{-1} \bar{X}_n + \left( 1 + \frac{n \tau^2}{\sigma^2} \right)^{-1} \alpha.
\]

Note that the sample average \( \bar{X}_n \) is an unbiased, consistent estimator for the location \( \theta \), while the prior expectation \( \alpha \) introduces a bias. As \( n \to \infty \), the difference between the posterior mean and the sample average goes to zero and the bias introduced by the prior disappears.

If we use empirical Bayes here, and we estimate \( \alpha \) from an independent i.i.d. sample \( X'_1, \ldots, X'_n \), with the sample average,

\[
\hat{\alpha}(X'_1, \ldots, X'_n) = \bar{X}'_n,
\]

then both contributions in (3.13) are unbiased. So, although the non-empirical estimate in (3.13) and its empirical-Bayes version are asymptotically equivalent, the
empirical Bayes has the attractive property that it de-biases the posterior mean *at finite values of* $n$.

In case that the point estimator $\hat{p}_n$ has a bias, empirical Bayes methods can be used to correct. That idea is applied somewhat unexpectedly in the following example.

**Example 3.11. (Multivariate normal mean)**

Suppose that $d \geq 3$ and we consider a data vector $Y = (Y_1, \ldots, Y_d)$ with components $Y_i$ that are modelled as independent, and distributed according to a multivariate normal distribution with a covariance matrix that is a known multiple ($\sigma^2 > 0$) of the identity,

$$Y_i \mid \theta \sim N(\theta_i, \sigma^2),$$

for each $1 \leq i \leq d$. A moment’s thought shows that the ML estimator for $\theta$ based only on $Y$, is given by $\hat{\theta}_{ML}(Y) = Y$. A prior for the parameter $\theta \in \mathbb{R}^d$ is chosen as follows: we view the components $(\theta_1, \ldots, \theta_d)$ as an i.i.d. sample from the one-dimensional normal distribution $N(\mu, \tau^2)$ with hyperparameters $\mu \in \mathbb{R}$ and $\tau^2 > 0$.

The prior predictive distribution for the data vector $Y$, given $\mu, \tau^2$, has Lebesgue density,

$$p_{\mu, \tau^2}^\Pi(y_1, \ldots, y_d) = \prod_{j=1}^d \frac{1}{\sqrt{2\pi(\sigma^2 + \tau^2)^{1/2}}} \exp\left(-\frac{1}{2} \frac{(y_j - \mu)^2}{\sigma^2 + \tau^2}\right).$$

The ML-II method prescribes that we maximize the prior predictive likelihood $p_{\mu, \tau^2}^\Pi(Y_1', \ldots, Y_d')$ based on an independent copy $Y'$ of $Y$ (in principle, but with $Y' = Y$ commonly). Taking derivatives of $\log p_{\mu, \tau^2}^\Pi$ with respect to $\mu$ and $\tau^2$, we find ML-II estimates for the hyperparameters,

$$\bar{y}' = \frac{1}{d} \sum_{j=1}^d y_j', \quad \bar{\tau}^2 = (s_d^2 - \sigma^2)_+ = \max\{0, s_d^2 - \sigma^2\},$$

with $s_d^2 = d^{-1} \sum (Y_j' - \bar{y}')^2$. Essentially, the resulting empirical prior imposes, for each component $\theta_j$, $1 \leq j \leq d$, a bias towards the average value $\bar{y}'$ of the observed components $Y_j'$ through $Y_d'$. Then, there are two distinct cases: when $s_d^2 \geq \sigma^2$, differences between (observed) components are relatively large, and the prior is normal with a variance that adds with $\sigma^2$ to the observed $s_d^2$; when $s_d^2 < \sigma^2$, differences between (observed) components $Y_1', \ldots, Y_d'$ are relatively small, indicating that their average $\bar{y}'$ may be of use when estimating the means $\theta_j$, $1 \leq j \leq d$, for the individual components. In the first case, the bias formulated by the prior expectation $\bar{y}'$ is mitigated by a prior variance that leaves room for doubt; in the second case, the prior is concentrated all the way on $\delta_{\bar{y}'=\bar{y}}$ (as in example 2.3). This is reflected in the posterior: when $s_d^2 \geq \sigma^2$, the empirical Bayes posterior for the $j$-th component of $\theta$ is (see (3.7)),

$$\Pi(\theta_j \in A \mid Y, \mu = \bar{y}', \tau^2 = \bar{\tau}^2) = N_d(\hat{\lambda} \bar{y}' + (1 - \hat{\lambda})Y_j, (1 - \hat{\lambda})\sigma^2)(A),$$
where $\lambda = d s^2 / \sigma^2$; when $s^2 < \sigma^2$, all posterior mass is *shrunk* into one point. In both cases, we can write the empirical Bayes posterior expectation as,

$$\hat{\theta}_{EB}(Y; Y') = (1 - \hat{\kappa}) Y' + \hat{\kappa} Y_j,$$

(3.14)

where $\hat{\kappa} = (1 - d \sigma^2 / s^2)_+$. Perhaps somewhat surprisingly, the empirical Bayes estimator $\hat{\theta}_{EB}$ outperforms the ML estimate $\hat{\theta}_{ML}$ and other unbiased estimators for the problem (with respect to mean-squared error). In fact, a slightly different estimator that shrinks the unbiased estimate was written down without reference to any Bayesian methods in Stein’s 1956 work [207] and then shown to be $R$-better than the usual estimates in James and Stein (1961) [126]: for $d \geq 3$, the James-Stein estimator is the shrinkage estimator,

$$\hat{\theta}_{JS}(Y) = \left(1 - \frac{(d-2)\sigma^2}{s^2}ight)(Y - \bar{Y}) + \bar{Y}.$$  

(3.15)

It came as quite a shock that for all $\theta \in \mathbb{R}^d$,

$$P_\theta(\hat{\theta}_{JS} - \theta)^2 \leq P_\theta(\hat{\theta}_{ML} - \theta)^2,$$

with strict inequality for most values of $\theta$ (for proof, see corollary 4.7.2 in Lehmann and Casella (1998) [153]): the James-Stein estimator generally has mean-squared error smaller than the mean-squared error of the ML estimator (and, according to the Lehmann-Scheffe theorem, also of all other unbiased estimators). For $d \geq 3$, the ML estimator for the location of a multivariate normal random variable is *inadmissible* in the sense of definition 2.31 and example 2.18. It was recognized in Efron and Morris (1973) [81] that point-estimators that result from empirical Bayes posteriors introduce the type of bias that the James-Stein estimator has and, correspondingly, outperform unbiased estimators in the given example. Indeed, if we use an unbiased estimator for $\sigma^2/(\sigma^2 + \tau^2)$ rather than the ML-II estimate (see problem 4.7.1 in [153]), the resulting empirical Bayes estimator is the James-Stein estimator. Because the ML estimate is optimal in the large class of regular estimators (see section 4.1), it is called *efficient*. Correspondingly, the James-Stein and empirical Bayes estimators are called *superefficient*.

To provide some counterweight to that remarkable conclusion, let us consider some of the drawbacks of shrinkage estimation: first of all, the improvement occurs only if we assess performance using the mean squared error. For example, it can be shown that the James-Stein estimator estimates individual components of $\theta$ with larger errors than the non-shrunk estimate [153]. Similarly, the James-Stein estimator does not outperform if we use loss-functions other than the squared-error loss. Secondly, the squared-error *Bayes* risk function (combine definition 2.34 and example 2.18) for shrinkage estimators is generally higher than that of their non-shrunk versions (for a discussion in the context of the multivariate mean problem discussed above, see example 4.7.3 in [153]). This may be explained by the fact that risk functions of estimators with fixed points of shrinkage tend to display wild fluctuations around the point of shrinkage (although this phenomenon is well-understood.
only if the model is one-dimensional \([155]\)). Finally, posterior variance tends to be shrunk too (see (3.7)), which leads to credible sets (and confidence sets based on the James-Stein estimator) that are too small. This can be understood from the fact that the empirical Bayes posterior does not account for the inaccuracies in the estimation of \(\tau^2\), it only quantifies the uncertainty in the subsequent estimation of \(\theta\), thus underestimating the overall error. Notwithstanding their practical usefulness \([81]\], perhaps this is the most serious short-coming of shrinkage estimators: improved estimation accuracy comes at the cost of impaired uncertainty quantification and other forms of inference (like testing with shrunk statistics).

\section*{3.5 Conjugate families}

In this section, we consider a type of prior choice that is motivated primarily by mathematical convenience, rather than philosophy or statistical utility. Recall that if we model the data with normal distributions of known variance but unknown location \(\theta\) and we supply \(\theta\) with a normal prior, then the posterior for \(\theta\) is again a normal distribution. Since the calculation of the posterior is tractable, any choice for the parameters of the normal prior can immediately be updated to values for location and variance of the normal posterior upon observation of \(Y = y\). Not only does this signify ease of manipulation in calculations with the posterior, it also reduces the computational burden dramatically since numerical integration or simulation from the posterior is no longer necessary.

\subsection*{3.5.1 Basic definitions and some examples}

The subject of this section revolves around the following definition.

**Definition 3.6.** Let \((\mathcal{P}, \mathcal{A})\) be a measurable model for an observation \(Y \in \mathcal{Y}\). Let \(M\) denote a collection of probability distributions on \((\mathcal{P}, \mathcal{A})\). The set \(M\) is called a **conjugate family** for the model \(\mathcal{P}\), if the posterior based on a prior from \(M\) again lies in \(M\):

\[
\Pi \in M \implies \Pi(\cdot | Y = y) \in M,
\]

for all almost all \(y \in \mathcal{Y}\).

(Like before, the phrase “almost all” in the above definition refers to the prior predictive distribution for Bayesians, and to the true \(P_0\) for frequentists.) Such structure was first proposed by Raiffa and Schlaifer (1961) \([189]\). Their method for the prior choice is usually classified as objectivist because it does not rely on subjectivist notions and is motivated without reference to outside factors.

**Remark 3.8.** Often in the literature, a prior is referred to as a “**conjugate prior**” if the posterior is of the same form. This is somewhat misleading, since it is the family
Choice of the prior

Example 3.12. Consider an experiment in which we observe $n$ independent Bernoulli trials and consider the total number of successes, $Y \sim \text{Bin}(n, p)$ with unknown parameter $p \in [0, 1]$, 

$$P_p(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$  

For the parameter $p$ we choose a prior $p \sim \text{Beta}(\alpha, \beta)$ from the Beta-family, for some $\alpha, \beta > 0$,

$$d\Pi(p) = B(\alpha, \beta) p^{\alpha-1} (1-p)^{\beta-1} dp,$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{(\Gamma(\alpha) \Gamma(\beta))}$ normalizes $\Pi$. Then the posterior density with respect to the Lebesgue measure on $[0, 1]$ is proportional to:

$$d\Pi(p | Y) \propto p^Y (1-p)^{n-Y} p^{\alpha-1} (1-p)^{\beta-1} dp = p^{\alpha+Y-1} (1-p)^{\beta+n-Y-1} dp,$$

We conclude that the posterior again lies in the Beta-family, with parameters equal to a data-amended version of those of the prior, as follows:

$$\Pi(\cdot | Y) = \text{Beta}(\alpha + Y, \beta + n - Y).$$

So the family of Beta-distributions is a conjugate family for the binomial model. Depending on the available amount of prior information on $\theta$, the prior’s parameters may be chosen on subjective grounds. However, in the absence thereof, the parameters $\alpha, \beta$ suffer from the same ambiguity that plagues the parameter $\tau^2$ featuring in the example with which we opened this section.

Example 3.12 indicates a strategy to find conjugate families for a given parametrized, dominated model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Customarily, we view densities $y \mapsto p_\theta(y)$ as functions of the outcome $Y = y$ but they are functions of the parameter $\theta$ as well and their dependence $\theta \mapsto p_\theta(y)$ determines which prior densities $\theta \mapsto \pi(\theta)$ preserve their functional form when multiplied by the likelihood $p_\theta(Y)$ to yield the posterior density.

3.5.2 Exponential families

Although we shall encounter an example of a conjugate family for a non-parametric model in section 3.6, conjugate families are, by and large, part of parametric statistics. Many models are so-called exponential families, for which conjugate families of priors can be found readily.

Definition 3.7. A Lebesgue-dominated collection of probability measures $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ (with densities $p_\theta$) is called a $k$-parameter exponential family, if there exists a $k \geq 1$ such that for all $\theta \in \Theta$, 

$$P_\theta(Y) = \int p_\theta(y) d\Pi(y)$$

where $\Pi$ is a $k$-parameter exponential family.
3.5 Conjugate families

\[ p_\theta(x) = \exp \left( \sum_{i=1}^{k} \eta_i(\theta) T_i(x) - B(\theta) \right) h(x), \quad (3.17) \]

where \( h \) and \( T_i, \ i = 1, \ldots, k \), are statistics and \( B, \eta_i, \ i = 1, \ldots, k \) are real-valued functions on \( \Theta \).

Any exponential family can be parametrized such that the exponent in (3.17) is linear in the parameter: by the mapping \( \Theta \rightarrow H: \eta_i = \eta_i(\theta) \) (a bijection if the original parametrization is identifiable), taking \( \Theta \) into \( H = \eta(\Theta) \) and \( B \) into \( A(\eta) = B(\theta(\eta)) \), any exponential family can be rewritten in its so-called canonical form.

**Definition 3.8.** An exponential family \( \mathcal{P} = \{ P_\eta : \eta \in H \} \), \( H \subset \mathbb{R}^k \) is said to be in its canonical representation, if

\[ p_\eta(x) = \exp \left( \sum_{i=1}^{k} \eta_i T_i(x) - A(\eta) \right) h(x). \quad (3.18) \]

In addition, \( \mathcal{P} \) is said to be of full rank if the interior of \( H \subset \mathbb{R}^k \) is non-void, i.e. \( \hat{H} \neq \emptyset \).

Although they are parametric models, exponential families are versatile modelling tools and have properties that are mathematically tractable; many common models, like the Bernoulli-, normal-, binomial-, Gamma-, Poisson-models, etcetera, can be rewritten in the form (3.17). To give an example of a type of parameter that cannot be accommodated in an exponential family, consider models in which the support of model distributions is parameter-dependent, like the family of all uniform distributions on \( \mathbb{R} \), or the parameter that describes the domain offset in the Pareto-model. Their statistical practicality stems primarily from the fact that for an exponential family of full rank, the statistics \( T_i, \ i = 1, \ldots, k \) are sufficient and complete, enabling the use of the Lehmann-Scheffé theorem for minimal-variance unbiased estimation (see, for instance, Lehmann and Casella (1998) [153]). Their versatility can be understood in many ways, e.g. by the Pitman-Koopman-Darmois theorem (see, Jeffreys (1961) [128]), which says that a family of distributions whose support does not depend on the parameter, is exponential, if and only if in the models describing its i.i.d. samples, there exist sufficient statistics whose dimension remains bounded asymptotically (i.e. as we let the sample size diverge to infinity).

**Example 3.13.** The model of all normal distributions \( \mathcal{P} = \{ N(\mu, \sigma^2) : \theta \in \mathbb{R}, \sigma^2 > 0 \} \) on \( \mathbb{R} \) forms an exponential family. To see this, write \( \theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty) \) and rewrite the usual parametrization in the form (3.17), as follows,

\[ p_{\mu,\sigma^2}(x) = (2\pi)^{-1/2} \sigma^{-1} \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right), \]

\[ = \exp \left( -\frac{1}{2\theta_2} x^2 + \frac{\theta_1}{\theta_2} x - \frac{\theta_1}{2\theta_2} - \frac{1}{2} \log \theta_2 - \frac{1}{2} \log (2\pi) \right), \]

and, comparing with (3.17), we read off,
\[ \eta_1(\theta) = \frac{\theta_1}{\theta_2}, \quad \eta_2(\theta) = -\frac{1}{2\theta_2}, \quad B(\theta) = \frac{\theta_1^2}{2\theta_2} + \frac{1}{2} \log \theta_2, \]
\[ T_1(x) = x, \quad T_2(x) = x^2, \quad h(x) = (2\pi)^{-1/2}. \]

The map \( \eta : \Theta \to H : \theta \mapsto (\eta_1, \eta_2)(\theta) \) takes the original parameter into the canonical parameter \( \eta \in H = \mathbb{R} \times (-\infty, 0) \). Note that the inverse of \( \eta \) takes the form,
\[ (\theta_1, \theta_2)(\eta) = \left( -\frac{\eta_1}{2\eta_2}, -\frac{1}{2}\eta_2 \right), \]
from which we deduce that,
\[ A(\eta) = B(\theta(\eta)) = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2} \log \left( -\frac{1}{2\eta_2} \right), \]
for the new normalization. Expressed in these new parameters \( \eta \), the density takes the form (3.18). Note that \( H = \mathbb{R} \times (-\infty, 0) \) has non-empty interior, so the normal model is an exponential family of full rank. In case we had started with the model \( \mathcal{P} = \{ N(\theta, \theta) : \theta > 0 \} \), for example, the analysis would have been largely analogous; however, the latter \( \mathcal{P} \) is not of full rank.

Presently our interest lies with the following theorem which says that if a model \( \mathcal{P} \) constitutes an exponential family, there exists a conjugate family of priors for \( \mathcal{P} \).

**Theorem 3.1.** Let \( \mathcal{P} \) be a model that can be written as an exponential family, c.f. definition 3.7. Then there exists a parametrization of \( \mathcal{P} \) of the form (3.18) and the family of distributions \( \Pi_{\mu, \lambda} \), defined by Lebesgue probability densities
\[ \pi_{\mu, \lambda}(\eta) = K(\mu, \lambda) \exp \left( \sum_{i=1}^{k} \eta_i \mu_i - \lambda A(\eta) \right), \quad \text{(3.19)} \]
(\( \mu \in \mathbb{R}^k \) and \( \lambda \in \mathbb{R} \) are such that \( 0 < K(\mu, \lambda) < \infty \)), is a conjugate family for \( \mathcal{P} \).

**Proof.** It follows from the argument preceding definition 3.8 that \( \mathcal{P} \) can be parametrized as in (3.18). Choosing a prior on \( H \) of the form (3.19), we find that the posterior again takes the form (3.19),
\[ \pi(\eta|X) \propto \exp \left( \sum_{i=1}^{k} \eta_i (\mu_i + T_i(X)) - (\lambda + 1) A(\eta) \right) \]
(the factor \( h(X) \) arises both in numerator and denominator of (2.6) and is \( \eta \)-independent, so that it cancels). The data-amended versions of the parameters \( \mu \) and \( \lambda \) that emerge from the posterior are therefore given by:
\[ (\mu + T(X), \lambda + 1), \]
and we conclude that the distributions \( \Pi_{\mu, \lambda} \) form a conjugate family for \( \mathcal{P} \).
From a frequentist perspective, it is worth noting the import of the factorization theorem, which says that the parameter-dependent factor in the likelihood is a function of the data only through the sufficient statistic. Since the posterior is a function of the likelihood, in which data-dependent factors that do not depend on the parameter can be cancelled between numerator and denominator, the posterior is a function of the data $Y$ only through the sufficient statistic $T(Y)$. Specifying to the case of exponential families, we note the following conclusion regarding $L_2$-optimality.

**Proposition 3.1.** If $\mathcal{P}$ is an exponential family of full rank, $T$ is not just sufficient but also complete for $\mathcal{P}$, and any point-estimator $\hat{\theta}$ based on the posterior (e.g. the posterior mean, see definition 2.7) that is unbiased and quadratically integrable (for all $\theta \in \Theta$), is optimal in the sense that, for all $\theta \in \Theta$,

$$P_\theta(\hat{\theta} - \theta)^2 \leq P_\theta(\hat{\eta} - \theta)^2,$$

for any other unbiased, quadratically integrable estimator $\hat{\eta}$ for $\theta$.

We omit the proof because it is a direct consequence of the famous Lehmann-Scheffé theorem (see, e.g., Lehmann and Casella (1998) [153]). The usefulness of proposition 3.1 is limited, because like the maximum-likelihood estimator, point-estimators based on the posterior tend to be biased, as we have seen in subsection 3.4.2.

### 3.6 Dirichlet priors

The construction of priors on non-parametric models is far from trivial. Broadly, there are two mathematical reasons for this: whereas the usual norm topology on $\mathbb{R}^k$ is unique [172], infinite-dimensional vector spaces support many different norm topologies and various other topologies besides. Similarly, whereas on $\mathbb{R}^k$ the (unique shift-invariant) Lebesgue measure provides a solid foundation for the definition of models in terms of densities, no such default uniform dominating measure exists in infinite-dimensional spaces.

Nevertheless many examples of probability measures on infinite-dimensional spaces exist. Some of the constructions and properties of the measures they result in, are discussed in detail in Ghosh and Ramamoorthi (2003) [103]. In this section, we look at a class of priors first proposed by Ferguson (1973) [89], which have become known as Dirichlet process priors.

#### 3.6.1 Dirichlet distributions

Let $\mathcal{X} = \{1, 2, \ldots, k\}$ (with its powerset $2^\mathcal{X}$ as a $\sigma$-algebra) and consider the collection $M(\mathcal{X})$ of all probability measures on $\mathcal{X}$. Every $P \in M(\mathcal{X})$ has a density $p : \mathcal{X} \rightarrow [0, 1]$ (with respect to the counting measure on $\mathcal{X}$) and we denote
Choice of the prior

\[ p_i = p(i) = P\{\{i\}\}, \text{ so that for every } A \in 2^\mathcal{X}, \]
\[ P(A) = \sum_{i \in A} p_i. \]

Therefore, the space \( M(\mathcal{X}) \) can be parametrized as follows,

\[ M(\mathcal{X}) = \{ P : 2^\mathcal{X} \to [0,1] : \sum_{i=1}^{k} p_i = 1, p_i \geq 0, (1 \leq i \leq k) \}, \]

and is in bijective correspondence with the simplex in \( \mathbb{R}^k \). For reasons to be discussed shortly, we consider the following family of distributions on \( M(\mathcal{X}) \).

**Definition 3.9. (Finite-dimensional Dirichlet distribution)**

Let \( \alpha = (\alpha_1, \ldots, \alpha_k) \) with \( \alpha_i > 0 \) for all \( 1 \leq i \leq k \). A stochastic vector \( p = (p_1, \ldots, p_k) \) is said to have Dirichlet distribution \( D_\alpha \) with parameter \( \alpha \), if the density \( \pi \) for \( p \) satisfies:

\[ \pi(p) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} p_{\alpha_1-1} p_{\alpha_2-1} \cdots p_{\alpha_k-1-1} \left(1 - \sum_{i=1}^{k-1} p_i\right)^{\alpha_k-1} \]

If \( \alpha_i = 0 \) for some \( i, 1 \leq i \leq k \), then we set \( D_\alpha(p_i = 0) = 1 \) marginally and we treat the remaining components of \( p \) as \((k-1)\)-dimensional.

**Example 3.14.** Consider the case where \( k = 2 \) (so that \( p_2 = 1 - p_1 \)): in that case, the density of the Dirichlet distribution takes the form:

\[ \pi(p_1, p_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) \Gamma(\alpha_2)} p_1^{\alpha_1-1} (1 - p_1)^{\alpha_2-1}, \]

i.e. \( p_1 \) has a Beta distribution \( B(\alpha_1, \alpha_2) \).

We also note the following two well-known facts on the Dirichlet distribution (proofs can be found in [103]).

**Lemma 3.2. (Gamma-representation of \( D_\alpha \))**

If \( Z_1, \ldots, Z_k \) are independent and each marginally distributed according to a Gamma-distribution with parameter \( \alpha_i \), i.e.

\[ Z_i \sim \Gamma(\alpha_i), \]

for all \( 1 \leq i \leq k \), then the normalized vector

\[ \left( \frac{Z_1}{S}, \ldots, \frac{Z_k}{S} \right) \sim D_\alpha, \]

(3.20)

with \( S = \sum_{i=1}^{k} Z_i \).
Lemma 3.2 shows that we may think of a $D_\alpha$-distributed vector as being composed of $k$ independent, $\Gamma$-distributed components, normalized to form a probability distribution through division by $S$ in (3.20). This division should be viewed as an $L_1$-projection from the positive cone in $\mathbb{R}^k$ onto the $k-1$-dimensional simplex. The following property can also be viewed as a statement on the effect of a projection on a distribution, this time from the simplex in $\mathbb{R}^k$ to lower-dimensional simplices. It is this property (related to a property called infinite divisibility of Gamma distributions) that motivates the choice for the Dirichlet distribution made by definition 3.9.

Lemma 3.3. Let $\mathcal{X}$ be a finite point-set. If the density $p : \mathcal{X} \to [0,1]$ of a distribution $P$ is itself distributed according to a Dirichlet distribution with parameter $\alpha$, $p \sim D_\alpha$, then for any partition $\{A_1, \ldots, A_m\}$ of $\mathcal{X}$, the vector of probabilities $(P(A_1), P(A_2), \ldots, P(A_m))$ has a Dirichlet distribution again, 

$$(P(A_1), P(A_2), \ldots, P(A_m)) \sim D_{\alpha'},$$

where the parameter $\alpha'$ is given by:

$$\left(\alpha'_1, \ldots, \alpha'_m\right) = \left(\sum_{l \in A_1} \alpha_l, \ldots, \sum_{l \in A_m} \alpha_l\right). \tag{3.21}$$

The identification (3.21) in lemma 3.3 suggests that we adopt a slightly different perspective on the definition of the Dirichlet distribution: we view $\alpha$ as a finite measure on $\mathcal{X}$, so that $P \sim D_\alpha$, if and only if, for every partition $(A_1, \ldots, A_m)$,

$$(P(A_1), \ldots, P(A_m)) \sim D_{(\alpha(A_1), \ldots, \alpha(A_m))}. \tag{3.22}$$

Property (3.22) serves as the point of departure of the generalization to the non-parametric model, because it does not depend on the finite nature of $\mathcal{X}$.

Definition 3.10. Let $\mathcal{X}$ be a finite point-set; denote the collection of all probability measures on $\mathcal{X}$ by $M(\mathcal{X})$. The Dirichlet family $\mathcal{D}(\mathcal{X})$ is defined to be the collection of all Dirichlet distributions on $M(\mathcal{X})$, i.e. $\mathcal{D}(\mathcal{X})$ consists of all $D_\alpha$ with $\alpha$ a finite measure on $\mathcal{X}$.

The proof of lemma 6.1 makes use of the following direct consequence.

Lemma 3.4. Let $\alpha$ be a finite measure on a finite point-set $\mathcal{X}$ and let $B \subset \mathcal{X}$ be given. Then,

(i) If $\alpha(B) = 0$, then $P(B) = 0$, $D_\alpha - a.s.$

(ii) If $\alpha(B) > 0$, then $P(B) > 0$, $D_\alpha - a.s.$

(iii) The $D_\alpha$-expectation of $P$ is,

$$\int P(B) \, dD_\alpha(P) = \frac{\alpha(B)}{\alpha(\mathcal{X})}.$$
Proof. Consider the partition \((B_1, B_2)\) of \(\mathcal{X}\), where \(B_1 = B, B_2 = \mathcal{X} \setminus B\). According to (3.22),
\[
(P(B_1), P(B_2)) \sim D(\alpha(B), \alpha(\mathcal{X}) - \alpha(B)),
\]
so that \(P(B) \sim B(\alpha(B), \alpha(\mathcal{X}) - \alpha(B))\). Stated properties then follow from the properties of the Beta-distribution.

The following property of the Dirichlet distribution describes two independent Dirichlet-distributed quantities in convex combination, which form a new Dirichlet-distributed quantity if mixed by means of an (independent) Beta-distributed parameter.

**Lemma 3.5.** Let \(\mathcal{X}\) be a finite point-set and let \(\alpha_1, \alpha_2\) be two measures on \((\mathcal{X}, 2^\mathcal{X})\). Let \((P_1, P_2)\) be independent and marginally distributed as
\[
P_1 \sim D_{\alpha_1}, \quad P_2 \sim D_{\alpha_2}.
\]
Furthermore, let \(\lambda\) be independent of \(P_1, P_2\) and marginally distributed according to \(\lambda \sim B(\alpha_1(\mathcal{X}), \alpha_2(\mathcal{X}))\). Then the convex combination \(\lambda P_1 + (1 - \lambda) P_2\) again has a Dirichlet distribution with base measure \(\alpha_1 + \alpha_2\):
\[
\lambda P_1 + (1 - \lambda) P_2 \sim D_{\alpha_1 + \alpha_2}.
\]

Many other properties of the Dirichlet distribution could be considered here, most notably the so-called tail-free property and neurality to the right (see [103]). We do not provide details because both are rather technical and we do not use them in following chapters.

A most important property of the family of Dirichlet distributions is its conjugacy for the full non-parametric model.

**Theorem 3.2.** Let \(\mathcal{X}\) be a finite point-set; let \(X_1, \ldots, X_n\) denote an i.i.d. sample of observations taking values in \(\mathcal{X}\). The Dirichlet family \(D(\mathcal{X})\) is a conjugate family: if the prior equals \(D_{\alpha}\), the posterior is a Dirichlet distribution \(D_{\alpha'}\) with,
\[
\alpha' = \alpha + \sum_{i=1}^{n} \delta_{X_i}, \quad (3.23)
\]
as a base measure.

Proof. Since \(\mathcal{X}\) is finite (denote the order of \(\mathcal{X}\) by \(k\)), \(M(\mathcal{X})\) is dominated (by the counting measure). If we denote the density \(p_l = P(\{l\})\) for every \(l \in \mathcal{X}\), the posterior can be written as in (2.14) and the likelihood takes the form:
\[
P \mapsto \prod_{i=1}^{n} p_{X_i} = \prod_{l=1}^{k} p_l^{n_l},
\]
where \(n_l\) denotes the number of \(X_i\) equal to \(l\), for all \(1 \leq l \leq k\). Multiplying by the prior density for \(\Pi = D_{\alpha}\), we find that the posterior density is proportional to,
\[\pi(p_1, \ldots, p_k|X_1, \ldots, X_n) \propto \pi(p_1, \ldots, p_k) \prod_{i=1}^{n} p_{X_i} \]
\[\propto \prod_{l=1}^{k} p_{l}^{n_l} \prod_{l=1}^{k} p_{l}^{\alpha_l-1} \left(1 - \sum_{i=1}^{k-1} p_i\right)^{\alpha_k-1} = \prod_{l=1}^{k} p_{l}^{\alpha_l+n_l-1} \left(1 - \sum_{i=1}^{k-1} p_i\right)^{\alpha_k+n_k-1},\]

which is again a Dirichlet density (but with changed base measure). Since the posterior is a probability distribution, we know that the normalization factor follows suit. Note that we may view \(n_l\) as the density of the measure,

\[n_l = \sum_{i=1}^{n} 1\{X_i = l\} = \sum_{i=1}^{n} \delta_{X_i}(l),\]

for every \(1 \leq l \leq k\). So the posterior is the Dirichlet distribution \(D_{\alpha'}\), with base measure (3.23).

### 3.7 Exercises

#### 3.1. A PROPER JEFFREYS PRIOR
Let \(X\) be a random variable, distributed \(Bin(n; p)\) for known \(n\) and unknown \(p \in (0, 1)\). Calculate Jeffreys prior for this model, identify a standard family of probability distributions that this prior would belong to, if it were normalized as a probability distribution.

#### 3.2. JEFFREYS AND UNIFORM PRIORS
Let \(\mathcal{P}\) be a model parametrized according to some mapping \(\Theta \rightarrow \mathcal{P} : \theta \mapsto P_\theta\). Assuming differentiability of this map, Jeffreys prior \(\Pi\) takes the form (3.6). In other parametrizations, the form of this expression remains the same, but the actual dependence on the parameter changes. This makes it possible that there exists another parametrization of \(\mathcal{P}\) such that Jeffreys prior is equal to the uniform prior. We shall explore this possibility below.

For each of the following models in their 'standard' parametrizations \(\theta \mapsto P_\theta\), find a parameter \(\eta \in H, \eta = \eta(\theta)\), such that the Fisher information \(I_\eta\), expressed in terms of \(\eta\), is constant.

a. Find \(\eta\) for the model of all Poisson distributions.

b. In the cases \(\alpha = 1, 2, 3\), find \(\eta\) for the model \(\mathcal{P}\) consisting of all \(\Gamma(\alpha, \theta)\)-distributions, with \(\theta \in (0, \infty)\).

c. Find \(\eta\) for the model \(\mathcal{P}\) of all \(Bin(n; \theta)\) distributions, where \(n\) is known and \(\theta \in (0, 1)\).

Note that if the Fisher information \(I_\eta\) is constant, Jeffries prior is uniform. Therefore, if a parametrization like above exists and \(H\) is unbounded, Jeffries prior is improper (in all parametrizations).

#### 3.3. OPTIMALITY OF UNBIASED BAYESIAN POINT ESTIMATORS
Let \(\mathcal{P}\) be a dominated, parametric model, parametrized identifiably by \(\Theta \rightarrow \mathcal{P}:\)
3 Choice of the prior

Let \( \theta \mapsto P_\theta \), for some \( \Theta \subset \mathbb{R}^k \). Assume that \( (X_1, \ldots, X_n) \in \mathcal{X}^n \) form an i.i.d. sample from a distribution \( P_0 = P_{\theta_0} \in \mathcal{P} \), for some \( \theta_0 \in \Theta \). Let \( \Pi \) be a prior on \( \Theta \) and denote the posterior by \( \Pi(\cdot | X_1, \ldots, X_n) \). Assume that \( T : \mathcal{X}^n \to \mathbb{R}^m \) is a sufficient statistic for the model \( \mathcal{P} \).

a. Use the factorization theorem to show that the posterior depends on the data only through the sufficient statistic \( T(X_1, \ldots, X_n) \).

b. Let \( \hat{\theta}_n : \mathcal{X}^n \to \Theta \) denote a point-estimator derived from the posterior. Use a. above to argue that there exists a function \( \hat{\theta}_n : \mathbb{R}^m \to \Theta \), such that

\[
\hat{\theta}_n(X_1, \ldots, X_n) = \hat{\theta}_n(T(X_1, \ldots, X_n)).
\]

Bayesian point-estimators share this property with other point-estimators that are derived from the likelihood function, like the maximum-likelihood estimator and penalized versions thereof. Next, assume that \( T \) is complete, that \( P_0^0(\hat{\theta}_n)^2 < \infty \) and that \( \hat{\theta}_n \) is unbiased, i.e. \( P_0^0 \hat{\theta}_n = \theta_0 \).

c. Apply the Lehmann-Scheffé theorem to prove that, for any other unbiased estimator \( \tilde{\theta}_n : \mathcal{X}^n \to \Theta \),

\[
P_0^0(\hat{\theta}_n - \theta_0)^2 \leq P_0^0(\tilde{\theta}_n - \theta_0)^2.
\]

The message of this exercise is, that Bayesian point-estimators that happen to be unbiased and quadratically integrable, are automatically \( L_2 \)-optimal in the class of all unbiased estimators for \( \theta \). They share this remarkable property with maximum-likelihood estimators.

3.4. Conjugate Model-Prior Pairs

In this exercise, conjugate model-prior pairs \((\mathcal{P}, \Pi)\) are provided. In each case, we denote the parameter we wish to estimate by \( \theta \) and assume that other parameters have known values. Let \( X \) denote a single observation.

In each case, derive the posterior distribution to prove conjugacy and identify the \( X \)-dependent transformation of parameters that takes prior into posterior.

a. \( X|\theta \sim N(\theta, \sigma^2) \) and \( \theta \sim N(\mu, \tau^2) \).

b. \( X|\theta \sim \text{Poisson}(\theta) \) and \( \theta \sim \Gamma(\alpha, \beta) \).

c. \( X|\theta \sim \Gamma(\rho, \theta) \) and \( \theta \sim \Gamma(\alpha, \beta) \).

d. \( X|\theta \sim \text{Bin}(n; \theta) \) and \( \theta \sim \beta(\alpha, \beta) \).

e. \( X|\theta \sim N(\mu, \theta^{-1}) \) and \( \theta \sim \Gamma(\alpha, \beta) \).

f. \( X|\theta_1, \ldots, \theta_k \sim M(n; \theta_1, \ldots, \theta_k) \) and \( \theta \sim D_\alpha \), where \( M \) denotes the multinomial distribution for \( n \) observations drawn from \( k \) classes with probabilities \( \theta_1, \ldots, \theta_k \) and \( D_\alpha \) is a Dirichlet distribution on the simplex in \( \mathbb{R}^k \) (see definition 3.9).

3.5. In this exercise, we generalize the setup of example 3.7 to multinomial rather than binomial context. Let \( k \geq 1 \) be known. Consider an observed random variable \( Y \) and an unobserved \( N = 1, 2, \ldots, \) such that, conditionally on \( N \), \( Y \) is distributed multinomially over \( k \) classes, while \( N \) has a Poisson distribution with hyperparameter \( \lambda > 0 \).
3.7 Exercises

\[ Y | N \sim M_k(N; p_1, p_2, \ldots, p_k), \quad N \sim \text{Poisson}(\lambda). \]

Determine the prior predictive distribution of \( Y \), as a function of the hyperparameter \( \lambda \).

3.6. Let \( X_1, \ldots, X_n \) form an i.i.d. sample from a Poisson distribution \( \text{Poisson}(\theta) \) with unknown \( \theta > 0 \). As a family of possible priors for the Bayesian analysis of this data, consider exponential distributions \( \theta \sim \Pi_\lambda = \text{Exp}(\lambda) \), where \( \lambda > 0 \) is a hyperparameter.

a. Calculate the prior predictive distribution for \( X \).
b. Give the ML-II estimate \( \hat{\lambda} \) for \( \lambda \).
c. With the estimated hyperparameter, give the posterior distribution \( \theta | X_1, \ldots, X_n \).
d. Calculate the posterior mean. Compare its data-dependence to that of the posterior mean we would have obtained if we had not made an empirical choice for the hyperparameter, but a fixed choice.

3.7. Let \( X_1, \ldots, X_n \) form an i.i.d. sample from a binomial distribution \( \text{Bin}(N; p) \), for known \( N \) and unknown \( p \in [0, 1] \). For the parameter \( p \) we take a prior \( p \sim \beta(\alpha, \beta) \) with hyperparameters \( \alpha, \beta > 0 \).

a. Show that the family of \( \beta \)-distributions is conjugate for binomial data.
b. Using (standard expressions for) the expectation and variance of \( \beta \)-distributions, give the posterior mean and variance in terms of the original \( \alpha \) and \( \beta \) chosen for the prior and the data.
c. Calculate the prior predictive distribution and discuss the steps one would perform in the ML-II procedure to estimate \( p \).
Chapter 4
The Bernstein-von Mises theorem

Given a model $\mathcal{P}$ and an infinite i.i.d. sample $X_1, X_2, \ldots$ drawn from $P_0 \in \mathcal{P}$, a point-estimation procedure gives rise to estimates $\hat{P}_n(X^n) \in \mathcal{P}$ calculated using only the first $n$ observations $X^n = (X_1, \ldots, X_n)$, for every $n \geq 1$. The resulting sequence of estimates $(\hat{P}_n)$ can be studied for stochastic convergence in the limit of growing sample size. More generally, any statistical procedure can be indexed by the size $n$ of the data sample used to calculate it, leading to sequences of (parameter) estimates $\hat{\theta}_n(X^n)$, tests $\phi_n(X^n)$, confidence regions $C_n(X^n)$, etcetera. Properties of such sequences reflect the behaviour of the procedure with growing sample-size, and an intuitively reasonable requirement of any estimation procedure is asymptotic consistency: the sequence $\hat{P}_n(X^n)$ approaches the true distribution $P_0$ to within arbitrary precision with high probability, if the sample size $n$ is high enough. Similarly, samples of arbitrarily large size should enable one to test with power arbitrarily close to a step-function and define arbitrarily small confidence regions. Further analysis of a consistent sequence $(\hat{P}_n)$ (or $\hat{\theta}_n$) concerns the (suitably rescaled) distribution of the estimator-sequence around its point of convergence. In the still parametric models that form the focus of this chapter, one studies the weak limit of differences $\sqrt{n}(\hat{\theta}_n - \theta_0)$ to refine the analysis to a notion of optimality (called efficiency) among so-called regular estimators, as borne out by Hajek’s 1970 convolution theorem.

The study of the asymptotic regime of an estimation procedure is interesting because asymptotic results provide approximations to exact values: finite-sample calculations are often intractable even in the simplest models, but the analysis of the large-sample limit often remains possible. The answer obtained in the large-sample limit may then be used as an approximation to the finite-sample answer. Asymptotic confidence sets obtained as credible sets with the Bernstein-von Mises theorem are an example of such an asymptotic approximation.

In this chapter, we consider the asymptotic behaviour of the Bayesian procedure in parametric models with likelihoods that depend smoothly on the parameter. The analysis is refined to the level of uncertainty quantification, showing that sequences of credible sets are approximations of of efficient confidence sets asymptotically. Recall from chapter 2 that at the conceptual level, the Bayesian posterior plays a role analogous to that of a frequentist sampling distribution: both are distributions on
the model or on its parameter space, supposedly informative at the inferential level. From that perspective it is natural to wonder whether credible sets and confidence sets have anything to do with each other. Since they are conceptually so close, could it be that they are close also mathematically, at least, when explained as frequentist devices?

The so-called Bernstein-von Mises theorem asserts that the sequence of posteriors on a smooth parametric model coincides more and more with a sharpening sequence of normal distributions centred on efficient point-estimators. The theorem demonstrates that the equivalence of credible sets and confidence sets holds in the asymptotic limit and shows that the relevant sets are efficient confidence intervals. Although the name of the theorem refers to the historical calculations of Bernstein (1917) [15] and von Mises (1931) [174], it is Le Cam (1953) [155] that truly deserves the credit. Certainly the most useful reference for this subject is Le Cam and Yang (1990) [166]. A version of the Bernstein-von Mises theorem based on Le Cam’s inequality is found in Le Cam (1986) [162].

4.1 Efficiency in smooth parametric models

First we consider frequentist estimation in smooth, parametric models and state Hájek’s convolution theorem, which characterizes efficiency. This paves the way for the Bernstein-von Mises theorem of the next section, which asserts that posterior distributions in smooth parametric models concentrate in an asymptotically normal way around efficient point-estimators. Essential to the development of efficient estimation are two concepts: smoothness of the model and regularity of the estimator. When properly defined and then combined, smoothness and regularity describe a notion of statistical optimality comparable (and related) to estimators that achieve minimal mean-squared error within the family of unbiased estimators in the Cramér-Rao sense.

4.1.1 Optimality in smooth, parametric estimation problems

The concept of efficiency has its origin in Fisher’s 1920’s claim of asymptotic optimality of the maximum-likelihood estimator in differentiable parametric models. Here, optimality of the ML estimate means that they are consistent, achieve optimal $n^{-1/2}$-rate of convergence and possessed a asymptotic sampling distribution of minimal variance. In 1930’s and –40’s, Fisher’s ideas on optimality in differentiable models were sharpened and elaborated upon. To illustrate, consider the following classical result from $M$-estimation (which can be found as theorem 5.23 in [219]).

**Theorem 4.1.** Let $Θ$ be open in $\mathbb{R}^k$ and assume that $\mathcal{P} = \{P_θ : θ ∈ Θ\}$ is a Lebesgue-dominated model for i.i.d. data $X_1, X_2, \ldots$, with densities $p_θ : \mathcal{X} \to \mathbb{R}$ such that $θ \mapsto \log p_θ(x)$ is differentiable at $θ_0$ for all $x ∈ \mathcal{X}$, with derivative $\ell_θ(x)$. 


Assume that there exists a function $\ell : \mathcal{X} \to \mathbb{R}$ such that $P_0 \ell^2 < \infty$ and,

$$\left| \log p_{\theta_1}(x) - \log p_{\theta_2}(x) \right| \leq \ell(x) \| \theta_1 - \theta_2 \|,$$

for all $\theta_1, \theta_2$ in an open neighbourhood of $\theta_0$. Furthermore, assume that $\theta \mapsto P_0 \log p_{\theta}$ has a second-order Taylor expansion around $\theta_0$ of the form,

$$P_{\theta_0} \log p_{\theta} = P_{\theta_0} \log p_{\theta_0} + \frac{1}{2} (\theta - \theta_0)^T I_{\theta_0}(\theta - \theta_0) + o(\| \theta - \theta_0 \|^2),$$

with non-singular $I_{\theta_0}$. If $(\hat{\theta}_n)$ is an estimator sequence satisfying,

$$\mathbb{P}_n \log p_{\hat{\theta}_n} \geq \sup_{\theta \in \Theta} \mathbb{P}_n \log p_{\theta} - o_p(n^{-1}),$$

such that $\hat{\theta}_n \xrightarrow{d} \theta_0$, then $(\hat{\theta}_n)$ is asymptotically linear,

$$n^{1/2}(\hat{\theta}_n - \theta_0) = n^{-1/2} \sum_{i=1}^n I_{\theta_0}^{-1} \ell_{\theta_0}(X_i) + o_p(n^{-1}).$$

In particular, $n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_{\theta_0}^{-1})$.

The last assertion of theorem 4.1 says that the (near-)maximum-likelihood estimators $(\hat{\theta}_n)$ are asymptotically consistent, converge at rate $n^{-1/2}$ and have the inverse Fisher information $I_{\theta_0}^{-1}$ as the covariance matrix for their (normal) limit distribution. At this stage of the discussion, we do not have an argument to show that this asymptotic behaviour is in any sense optimal. Nevertheless, let us take the opportunity to illustrate briefly how asymptotic behaviour translates into inference on $\theta$ by considering associated asymptotic confidence sets.

Recall definition 2.15 and example 2.11: an asymptotic confidence set is an approximate confidence set that is derived not from an exact sampling distribution, but from approximations implied by limit distributions, e.g. from $n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_{\theta_0}^{-1})$ in the above theorem. To demonstrate, first suppose that the model is one-dimensional and satisfies the conditions of theorem 4.1. Denoting quantiles of the standard normal distribution by $\xi_\alpha$, we see from the last assertion of the theorem that:

$$P_{\theta_0} \left( -\xi_\alpha I_{\theta_0}^{1/2} < n^{1/2}(\hat{\theta}_n - \theta_0) \leq \xi_\alpha I_{\theta_0}^{1/2} \right) \to 1 - 2\alpha.$$

If the Fisher information were known, this would give rise immediately to a confidence interval: the above display implies that,

$$\left[ \hat{\theta}_n - n^{-1/2} \xi_\alpha I_{\theta_0}^{1/2}, \hat{\theta}_n + n^{-1/2} \xi_\alpha I_{\theta_0}^{1/2} \right]$$

has asymptotic coverage probability $1 - 2\alpha$. Since the Fisher information is not known exactly, we substitute an estimator for it, for example the sample variance $S_n^2$, to arrive at a studentized version of the above, which has the same asymptotic cov-
The Bernstein-von Mises theorem erage and can therefore be used as an asymptotic confidence interval. But we could also have chosen to “plug in” the estimator \( \hat{\theta}_n \) for \( \theta_0 \) in the expression for the Fisher information to arrive at an estimate \( I_{\hat{\theta}_n} \). To generalize to higher-dimensional \( \Theta \subset \mathbb{R}^k \), recall that if \( Z \) has a \( k \)-dimensional multivariate normal distribution \( \mathcal{N}_k(0, \Sigma) \), then \( Z^T \Sigma^{-1} Z \) possess a \( \chi^2 \)-distribution with \( k \) degrees of freedom. Denoting quantiles of the \( \chi^2 \)-distribution with \( k \) degrees of freedom by \( \chi^2_{k, \alpha} \), we find that so-called Wald-type confidence sets, ellipsoids of the form,

\[
C_{\alpha}(X_1, \ldots, X_n) = \{ \theta \in \Theta : n(\theta - \hat{\theta}_n)^T I_{\hat{\theta}_n}(\theta - \hat{\theta}_n) \leq \chi^2_{k, \alpha} \},
\]

have coverage probabilities converging to \( 1 - \alpha \) and are therefore asymptotic confidence sets.

### 4.1.2 Regularity and efficiency

Theorem 4.1 requires a rather large number of smoothness properties of the model: log-densities are required to be differentiable and Lipschitz and the Kullback-Leibler divergence must display a second-order expansion with non-singular second derivative matrix. These sufficient conditions are there to guarantee that the ML estimator displays a property known as *regularity* and the conditions listed are usually referred to as “regularity conditions”. The prominence of regularity in the context of optimality questions was not fully appreciated until in 1951, J. Hodges discovered an estimator that displayed a property now known as *superefficiency*.

**Example 4.1.** Suppose that we estimate a parameter \( \theta \in \Theta = \mathbb{R} \) with an estimator sequence \( (\hat{\theta}_n) \), satisfying limiting behaviour described by,

\[
n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{P_{\theta-w.}} L_\theta,
\]

for some law \( L_\theta \), for all \( \theta \in \Theta \). In addition, we define a so-called *shrinkage estimator*,

\[
S_n(X^n) = \begin{cases} 
\hat{\theta}_n(X^n), & \text{if } |\hat{\theta}_n(X^n)| \geq n^{-1/4} \\
0, & \text{if } |\hat{\theta}_n(X^n)| < n^{-1/4}.
\end{cases}
\]

The estimator \( S_n \) has a *bias* towards 0: any realization of \( \hat{\theta}_n \) that is close enough to 0 is “shrunk” to 0 fully. One shows quite easily that \( S_n \) has the same asymptotic behaviour as \( \hat{\theta}_n \) as long as \( \theta \neq 0 \), i.e. \( n^{1/2}(S_n - \theta) \xrightarrow{P_{\theta-w.}} L_\theta \) if \( \theta \neq 0 \). But if \( \theta = 0 \), \( \varepsilon_n(S_n - 0) \xrightarrow{P_{\theta=0-w.}} 0 \) for any rate sequence \( \varepsilon_n \). In other words, the asymptotic quality of \( S_n \) is as good as that of \( \hat{\theta}_n \), or *strictly better* if \( \theta = 0 \). In a next step we could improve on \( S_n \), by constructing a version of \( S_n \) that displays shrinkage in another point. Generalisation of this construction to other estimators and other models essentially says that *any* estimator sequence can be improved upon in a strict sense, at least in one point, through some form of shrinkage. (*NB: In one-dimensional models [155],*
superefficiency comes at a price, paid in terms of the behaviour of risk functions in neighbourhoods of the point of shrinkage. Furthermore, in one-dimensional setting, superefficiency can be achieved on a subset of Lebesgue measure zero only. In models of dimension 3 or higher, this restriction does not apply, as demonstrated by the strict improvement on risk that the James-Stein shrinkage estimator represents.)

So at certain points in the parameter space, Hodges’s shrinkage estimators and other superefficient estimators outperform the MLE and other estimators like it asymptotically, while doing equally well for all other points. In 1951, Hodges’s superefficiency indicated that Fisher’s 1920’s claim was false without further refinement and that a comprehensive understanding of optimality in differentiable estimation problems remained elusive.

To resolve the issue and arrive at a sound theory of asymptotic optimality in the estimation of smooth parameters, we have to introduce two concepts. The first is a concise notion of smoothness that describes local behaviour of likelihood products directly in terms of score functions. The “local” aspect of the definition stems from the $n^{-1/2} (\theta - \theta_0)$.

**Definition 4.1.** (Local asymptotic normality (LAN), [7])

Let $\Theta \subset \mathbb{R}^k$ be open, parametrizing a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ for i.i.d. data $X_1, X_2, \ldots$ that is dominated by a $\sigma$-finite measure with densities $p_\theta$. The model is said to be locally asymptotically normal (LAN) at $\theta_0$ if, for any converging sequence $h_n \to h$ in $\mathbb{R}^k$:

$$
\log \prod_{i=1}^n \frac{p_{\theta_0 + n^{-1/2} h_n}(X_i)}{p_{\theta_0}} = h^T \Gamma_{n, \theta_0} - \frac{1}{2} h^T I_{\theta_0} h + o_p(1), \quad (4.2)
$$

for random vectors $\Gamma_{n, \theta_0}$ such that $\Gamma_{n, \theta_0} \xrightarrow{p_{\theta_0}} N_k(0, I_{\theta_0})$.

Typical parameters for which the LAN-expansion (7.14) holds are the parameters $\theta$ (or $\eta(\theta)$) in exponential families, and typical examples of parameters that are not LAN are domain boundaries, like the location parameter in exponential or Pareto models. The LAN property formulates a notion of smoothness in parameter dependence and it is useful to formulate sufficient conditions based on differentiability of the density $\theta \mapsto p_\theta(x)$ at $\theta_0$ for every $x$.

**Proposition 4.1.** Let $\Theta \subset \mathbb{R}^k$ be open, parametrizing a dominated model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ for i.i.d. data $X_1, X_2, \ldots \in \mathcal{X}$ with densities $p_\theta : \mathcal{X} \to [0, \infty)$. Assume that the map $\theta \mapsto \sqrt{p_\theta(x)}$ is continuously differentiable for every $x$. If elements of the matrix $I_\theta = P_\theta \ell_\theta \ell_\theta^T$ are finite and depend on $\theta$ continuously, then the model is LAN with respect to $\theta$, with

$$
\Gamma_{n, \theta_0} = n^{-1/2} \sum_{i=1}^n \ell_{\theta_0}(X_i).
$$
Proof. This proposition is a combination of lemma 7.6 and theorem 7.2 in [219].

But local asymptotic normality can be achieved under weaker conditions; well known is the following property, best characterized as Hadamard differentiability of square-roots of model densities relative to the $L_2(P_0)$ norm.

**Definition 4.2.** (Differentiability in quadratic mean (DQM))

Let $\Theta \subset \mathbb{R}^k$ be open. A dominated model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ for i.i.d. data $X_1, X_2, \ldots$ with densities $p_\theta$ is said to be differentiable in quadratic mean at $\theta_0 \in \Theta$, if there exists a score function $\ell_{\theta_0} \in L_2(P_{\theta_0})$ such that:

$$\int \left( p_{\theta_0 + h}^{1/2} - p_{\theta_0}^{1/2} - \frac{1}{2} h^T \ell_{\theta_0} p_{\theta_0}^{1/2} \right)^2 d\mu = o(\|h\|^2),$$

as $h \to 0$.

Theorem 7.2 in [219] shows that a model that is DQM at $\theta_0$, is LAN at $\theta_0$. However, in many situations, it is quite straightforward to demonstrate the LAN property directly, in i.i.d. context usually through application of the central limit for $\Gamma_n$, $\theta_0$ and the law of large numbers for the term that is second order in $h$.

The second essential concept is a property that characterizes the class of estimators over which optimality is achieved, in particular excluding Hodges’s shrinkage estimators (and all other examples of superefficiency, as becomes clear below). To prepare the definition heuristically, note that, given Hodges’s counterexample, it is not enough to have estimators with pointwise convergence to limit laws; we must restrict the behaviour of estimators over ($n^{-1/2}$-)neighbourhoods rather than allow the type of wild variations that make superefficiency possible.

**Definition 4.3.** Let $\Theta \subset \mathbb{R}^k$ be open. An estimator sequence $(T_n)$ for the parameter $\theta$ is said to be regular at $\theta$ if, for all $h \in \mathbb{R}^k$,

$$n^{1/2} \left( T_n - (\theta + n^{-1/2} h) \right) \overset{P}{\longrightarrow} L_\theta,$$

where $P_n = P_{\theta + n^{-1/2} h}$.

The point of definition 4.3 is the requirement that the limit law $L_\theta$ depends on $\theta$ but is independent of $h$, indicating that the limit law is insensitive to perturbation of the parameter of size $n^{-1/2} h$. Typical examples of regular estimators are sample-means that estimate expectations (provided a second moment exists), while typical non-regular estimators are shrinkage estimators (like those of example 4.1 and definition (3.15)) and estimators like $\hat{\theta}_n = \max \{ X_i : 1 \leq i \leq n \}$ for the parameter $\theta$ that represents the upper bound of the support for the distribution of a bounded, real-valued random variable $X$.

The two properties covered, local asymptotic normality and regularity, come together in the following theorem (see theorems 7.10, 8.3 and 8.4 in [219], or corollary 7.4.23 in [214]), which describes the foundation for the convolution theorem.
that follows: the models \( \mathcal{P}_n = \{ P_\theta^n : \theta \in \Theta \} \) for the \textit{i.i.d.} samples \( X^n \) have a “limiting model” (for a fully developed theory of this type of limits of experiments, see Le Cam (1964) [158]; see also [162, 166] and [214]) that describes a single observation for a normal distribution with unknown location, and regular sequences of estimators \((T_n)\) are matched with a random variable \( T \) in the limiting model, in an asymptotically unbiased way.

**Theorem 4.2.** Let \( \Theta \subset \mathbb{R}^k \) be open; let \( \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \) be LAN at \( \theta_0 \) with non-singular Fisher information \( I_{\theta_0} \). Let \((T_n)\) be regular estimators in the “localized models” \( \{ P_{\theta_0 + n^{-1/2} h} : h \in \mathbb{R}^k \} \). Then there exists a (randomized) statistic \( T \) in the normal location model \( \{ N_k(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k \} \) such that \( T \sim h \sim L_{\theta_0}, \) for all \( h \in \mathbb{R}^k \).

Theorem 4.2 provides every regular estimator sequence with a limit in the form of a statistic in a very simple statistical experiment involving only a single \( N_k(h, I_{\theta_0}^{-1}) \)-distributed observation \( X \) with unknown location \( h \): the (weak) limit distribution that describes the local asymptotics of the sequence \((T_n)\) under \( P_{\theta_0+n^{-1/2}h} \) equals the distribution of \( T \) under \( h \), for all \( h \in \mathbb{R}^k \). Moreover, regularity of the sequence \((T_n)\) implies that under \( N_k(h, I_{\theta_0}^{-1}) \), the distribution of \( T \) relative to \( h \) is independent of \( h \), an invariance usually known as equivariance-in-law. The class of equivariant-in-law estimators for location in the model \( \{ N_k(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k \} \) is fully known: for any equivariant-in-law estimator \( T \) for \( h \), there exists a probability distribution \( M \) such that \( T \sim N_k(h, I_{\theta_0}^{-1} \ast M) \). The most straightforward example is \( T = X \), for which \( M = \delta_0 \). This argument gives rise to the following central result in the theory of efficiency.

**Theorem 4.3.** (Convolution theorem (Hajék, 1970) [115])

Let \( \Theta \subset \mathbb{R}^k \) be open and let \( \{ P_\theta : \theta \in \Theta \} \) be LAN at \( \theta_0 \) with non-singular Fisher information \( I_{\theta_0} \). Let \((T_n)\) be a regular estimator sequence with limit distribution \( L_{\theta_0} \). Then there exists a probability distribution \( M_{\theta_0} \) such that,

\[
L_{\theta_0} = N_k(0,I_{\theta_0}^{-1} \ast M_{\theta_0}),
\]

In particular, if \( L_{\theta_0} \) has a covariance matrix \( \Sigma_{\theta_0} \), then \( \Sigma_{\theta_0} \succeq I_{\theta_0}^{-1} \).

The occurrence of the inverse Fisher information as an optimal lower-bound in asymptotic context is finally explained here: the estimator \( T \) is unbiased so it satisfies the Cramér-Rao lower bound for asymptotic variance in the limiting model \( \{ N_k(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k \} \). Convolution of \( N_k(0,I_{\theta_0}^{-1}) \) with any distribution \( M \) raises its variance unless \( M \) is degenerate: the last assertion of the convolution theorem says that, within the class of regular estimates, asymptotic variance is lower-bounded by the inverse Fisher information. A regular estimator that is optimal in this sense, is called \textit{best-regular}; an example is the ML estimator of theorem 4.1. Anderson’s lemma below broadens the notion of optimality, in the sense that best-regular estimators outperform other regular estimators with respect to many loss functions.

**Definition 4.4.** A sub-convex loss-function is a map \( \ell : \mathbb{R}^k \to [0, \infty) \) such that the level sets \( \{ x \in \mathbb{R}^k : \ell(x) \leq c \} \) are closed, convex and symmetric around the origin.
Examples of subconvex loss-functions are many and include, for example, the common choices $\ell(x) = \|x\|^p$, $p \geq 1$.

**Lemma 4.1.** (Anderson’s lemma)

For any $k \geq 1$, any sub-convex loss function $\ell$, any probability distribution $M$ on $\mathbb{R}^k$ and any $k$-dimensional covariance matrix $\Sigma$,

$$\int \ell dN_k(0, \Sigma) \leq \int \ell d(N_k(0, \Sigma) * M).$$

**Proof.** A proof of Anderson’s lemma can be found, for instance, in [124].

Based on Anderson’s lemma, we see that the extent of the convolution theorem is greater than mere optimality with respect to some specific loss function; Hajék’s conclusion establishes efficiency with respect to the entire class of all sub-convex loss functions. To conclude we mention the following equivalence, which characterizes efficiency concisely in terms of a weakly converging sequence.

**Lemma 4.2.** In a LAN model, estimators $(T_n)$ for $\theta$ are best-regular if and only if the $(T_n)$ are asymptotically linear, which means that for all $\theta$ in the model,

$$n^{1/2}(T_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I^{-1}_\theta \ell_\theta(X_i) + o_P(1). \quad (4.3)$$

The random sequence of differences on the $r.h.s.$ of (4.3) is denoted by $\Delta_n(\theta)$ in theorem 4.4. Coming back to theorem 4.1, we see that under stated conditions, a consistent MLE $(\hat{\theta}_n)$ is best-regular, finally giving substance to Fisher’s claim. Referring to the discussion on confidence sets with which we opened this section, we now know that in a LAN model confidence sets of the form (4.1), based on best-regular estimators $(\hat{\theta}_n)$, enjoy a similar form of optimality: according to the convolution theorem, the asymptotic sampling distributions of best-regular estimator sequences are all the same and sharpest among asymptotic sampling distributions for regular estimators. The question remains if we can somehow identify confidence sets and credible intervals; in the next section, that identification is made asymptotically, and in part II, we take this point a great deal further, showing that credible sets derived from a posterior can be always be enlarged slightly to obtain asymptotic confidence sets, if one is using the right prior.

### 4.2 Le Cam’s Bernstein-von Mises theorem

To address the question of efficiency in smooth parametric models from a Bayesian perspective, we turn to the Bernstein-von Mises theorem. The first results concerning limiting normality of a posterior distribution date back to Laplace (1820) [152]. Later, Bernstein (1917) [15] and Von Mises (1931) [174] proved results to a similar extent. Le Cam used the term ‘Bernstein-Von Mises theorem’ in 1953 [155] and
proved its assertion in greater generality. Walker (1969) [220] and Dawid (1970) [59] gave extensions to these results and Bickel and Yahav (1969) [27] proved a limit theorem for posterior means. Below we follow Le Cam and Yang (1990) [2]. It worth emphasizing that in this section we shall not be too strict in Bayesian, subjectivist orthodoxy and interpret the posterior as a frequentist device, in a role very close conceptually to that of a sampling distribution.

The (proof of the) Bernstein-Von-Mises theorem depends crucially on local asymptotic normality of the model at \( \theta_0 \). A quick sketch of the proof can be given as follows. Suppose that the prior has a Lebesgue density that is continuous and strictly positive at \( \theta_0 \). Also assume that the posterior concentrates in neighbourhoods of \( \theta_0 \) of sizes decreasing as \( n^{-1/2} \). Then it makes sense to consider the posterior for the local parameter \( h = \sqrt{n}(\theta - \theta_0) \in H \) (for some bounded \( H \)), with Lebesgue-density:

\[
\pi_n(h | X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} p_{\theta_0 + h/\sqrt{n}}(X_i) \pi(\theta_0 + h/\sqrt{n}) / \int \prod_{i=1}^{n} p_{\theta_0 + h'/\sqrt{n}}(X_i) \pi(\theta_0 + h'/\sqrt{n}) dh',
\]

almost-surely. Continuity of the Lebesgue density \( \pi \) of the prior at \( \theta_0 \) implies that \( \pi(\theta_0 + h/\sqrt{n}) \) converges to the constant \( \pi(\theta_0) \), which is strictly positive by assumption. This makes it plausible that upon substitution of the likelihood expansion (4.2), the posterior density converges to:

\[
\prod_{i=1}^{n} p_{\theta_0 + h/\sqrt{n}}(X_i) dh \approx \int \prod_{i=1}^{n} p_{\theta_0 + h'/\sqrt{n}}(X_i) dh' \rightarrow \int dN(h, I^{-1}_{\theta_0})(X) dh \
\]

(4.4)

(in a suitable sense with respect to \( P_0 \)). Here \( X \) is an observation in the normal limit model \( \{N(\theta, I^{-1}_{\theta_0}) : \theta \in \mathbb{R}^k\} \). The r.h.s. of the last display equals \( dN(X, I^{-1}_{\theta_0})(h) \) and is the posterior based on a sample consisting only of \( X \) and the Lebesgue prior on \( H \) for the limit model.

**Theorem 4.4. (Bernstein-Von Mises [155, 166])**

Assume that \( \Theta \subset \mathbb{R}^k \) is open and that the model \( \mathcal{M} = \{P_\theta : \theta \in \Theta\} \) is identifiable and dominated. Suppose \( X_1, X_2, \ldots \) forms an i.i.d. sample from \( P_{\theta_0} \) for some \( \theta_0 \in \Theta \). Assume that the model is locally asymptotically normal at \( \theta_0 \) with non-singular Fisher information \( I_{\theta_0} \). Furthermore suppose that the prior \( \Pi_{\theta_0} \) has a Lebesgue density that is continuous and strictly positive at \( \theta_0 \) and that for every \( \varepsilon > 0 \), there exists a test sequence \( (\phi_n) \) such that,

\[
P_{\theta_0}^\phi \phi_n \rightarrow 0, \quad \sup_{\|\theta - \theta_0\| > \varepsilon} P_{\theta_0}(1 - \phi_n) \rightarrow 0.
\]

Then the posterior distributions converge in total variation,

\[
\| \Pi(h \in \cdot | X_1, \ldots, X_n) - N(\Delta_n, \theta_0, I^{-1}_{\theta_0}) \|_{P_\theta} \rightarrow 0.
\]
where $\Delta_{n, \theta_0} = \sqrt{n}(\hat{\theta}_n - \theta_0)$ for any best-regular estimator-sequence $\hat{\theta}_n$.

**Proof.** For a proof, see theorem 4.5, as well as the misspecified theorems in chapter 5.

Since the total-variational distance $\|N(\mu, \Sigma) - N(\nu, \Sigma)\|$ is bounded by a multiple of $\|\mu - \nu\|$, we find that the assertion of the Bernstein-Von-Mises theorem can also be formulated with $\sqrt{n}(\hat{\theta}_n - \theta_0)$ replacing $\Delta_{n, \theta_0}$. Using the invariance of total-variation under rescaling and shifts, this leads to the conclusion that:

$$\sup_B \left| \Pi(\theta \in B \mid X_1, \ldots, X_n) - N(\hat{\theta}_n, (nI_{\theta_0})^{-1})(B) \right| \to 0,$$

in $P_{\theta_0}$-probability, where $(\hat{\theta}_n)$ denotes any best-regular estimator sequence. In particular, according to theorem 4.1 and equivalence (4.3), the maximum-likelihood estimator is best-regular under stated smoothness conditions on the (log-)likelihood. This serves to motivate the often-heard statement that “Bayesian statistics coincides with the maximum-likelihood estimator asymptotically”. In figure 4.1, Bernstein-von Mises-type of convergence of the posterior is demonstrated with a graphical/numerical example. Also displayed in figure 4.1 are the MAP-estimator of

![Graphical illustration of posterior density convergence](image)

**Fig. 4.1** Convergence of the posterior density. The samples used for calculation of the posterior distributions consist of $n$ observations; the model consists of all normal distributions with mean between $-1$ and 2 and variance 1 and has a polynomial prior, shown in the first ($n=0$) graph. For all sample sizes, the maximum a posteriori and maximum likelihood estimators are indicated by a vertical line and a dashed vertical line respectively. (From Kleijn (2003))
4.2 Le Cam’s Bernstein-von Mises theorem

Here, the MLE is efficient so it forms a possible centring sequence for the limiting sequence of normal distributions in the assertion of the Bernstein-Von Mises theorem. Furthermore it is noted that the posterior concentrates more and more sharply, reflecting the $n^{-1}$-proportionality of the variance of its limiting sequence of normals. It is perhaps a bit surprising in figure 4.1 to see limiting normality obtain already at such relatively low values of the sample size $n$. It cannot be excluded that this is merely a manifestation the normality of the underlying model, but onset of normality of the posterior appears to happen at unexpectedly low values of $n$ also in other smooth, parametric setting. It suggests that asymptotic conclusions based on the Bernstein-Von Mises limit accrue validity fairly rapidly, for $n$ in the order of several hundred to several thousand i.i.d. replications of the observation, at least, in well-behaved simple cases.

The uniformity in the assertion of the Bernstein-Von Mises theorem over model subsets $B$ implies that it holds also for model subsets that are random. In particular, given some $0 < \alpha < 1$, it is noted that the smallest sets $C_\alpha (X_1, \ldots, X_n)$ such that,

$$N_{\hat{\theta}_n (nI_{\theta_0})^{-1}} (C_\alpha (X_1, \ldots, X_n)) \geq 1 - \alpha,$$

are ellipsoids of the form (4.1). Since posterior coverage of $C_\alpha$ converges to the l.h.s. in the above display, in accordance with the Bernstein-Von Mises limit, we see that the $C_\alpha$ are asymptotic credible sets of posterior coverage $1 - \alpha$. Conversely, any sequence $(D_n (X_1, \ldots, X_n))$ of (data-dependent) credible sets of coverage $1 - \alpha$, is a sequence of sets that have asymptotic confidence level (arbitrarily close to) $1 - \alpha$ (where we use that $\hat{\theta}_n$ is best-regular). This gives rise to an identification in smooth, parametric models between inference based on frequentist best-regular point-estimators and inference based on Bayesian posteriors. In a practical sense, it eliminates the need to estimate $\theta$ and the Fisher information $I_\theta$ at $\theta$ to arrive at asymptotic confidence sets, if we have an approximation of the posterior distribution of high enough quality (e.g. from MCMC simulation), provided the Bernstein-Von Mises theorem holds. In parametric models of higher dimension, maximization of the likelihood can be more costly computationally than generation of a suitable MCMC approximation for the posterior [194] (although often the computational burdens are comparable).

Before we continue with the proof of the Bernstein-Von Mises theorem, let is briefly reflect on its conditions: local asymptotic normality and non-singularity of the associated Fisher information are minimal smoothness conditions. They also arise in theorem 4.1 and form the backdrop for any discussion of efficiency. More significant is the required existence of a “consistent” test sequence: what is required is that, asymptotically, we can distinguish $R_0$ from any complement of a $\theta$-neighbourhood around $\theta_0$ uniformly. One should compare this condition with that of consistency of near-maximizers of the likelihood in theorem 4.1. Apparently, if such a global (rather than $n^{-1/2}$-sized local) consistency guarantee can not be given, likelihood-based methods like ML or Bayesian estimation cannot be trusted to give rise to asymptotic normality (in their respective forms). In the next section, we shall divide the Bernstein-Von Mises theorem in two parts, with a requirement of local
4.2.1 Proof of the Bernstein-von Mises theorem

We prove the assertion of the Bernstein-Von Mises theorem using a smoothness property that is slightly stronger than LAN.

**Definition 4.5.** We say that a parametric model $\mathcal{P}$ is stochastically LAN at $\theta_0$, if the LAN property of definition 4.1 is satisfied for every random sequence $(h_n)$ that is bounded in probability, i.e. for all $h_n = O_P(1)$:

$$\log \prod_{i=1}^n \frac{p_{\theta_0} + n^{-1/2}h_n(x_i)}{p_{\theta_0}} - h_n^T I_{\theta_0} h_n - \frac{1}{2} h_n^T I_{\theta_0} h_n = o_P(1), \quad (4.5)$$

for random vectors $\Gamma_{n,\theta_0}$ such that $\Gamma_{n,\theta_0} \overset{h.w.}{\longrightarrow} N_k(0, I_{\theta_0})$.

**Theorem 4.5.** Let the sample $X_1, X_2, \ldots$ be distributed i.i.d. $\cdot P_0$. Let $\Theta \subset \mathbb{R}^k$ be open, let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be stochastically LAN at $\theta_0$ with non-singular Fisher information $I_{\theta_0}$ and let the prior $\Pi$ on $\Theta$ be thick. Furthermore, assume that for every sequence of origin-centred balls $(K_n) \subset \mathbb{R}^d$ with radii $M_n \to \infty$, we have:

$$\Pi_n \left( h \in K_n \mid X_1, \ldots, X_n \right) \overset{P_0}{\longrightarrow} 1. \quad (4.6)$$

Then posteriors converge to normal distributions in total variation:

$$\sup_B \left| \Pi_n \left( h \in B \mid X_1, \ldots, X_n \right) - N_{\Delta_n, \theta_0}^{-1} (B) \right| \overset{P_0}{\longrightarrow} 0, \quad (4.7)$$

where $\Delta_n, \theta_0 = \sqrt{n}(\hat{\theta}_n - \theta_0)$ for any best-regular estimator-sequence $\hat{\theta}_n$.

**Proof.** The proof is split into two parts: in the first part, we prove the assertion conditional on a compact neighbourhood $K$ of $0$ in $\Theta$, and in the second part we diagonalize based on a sequence $(K_n)$ with $\cup_n K_n = \mathbb{R}^k$ to prove (5.6). Throughout the proof we denote the posterior for $h$ given $X_1, X_2, \ldots, X_n$ by $\Pi_n$ and the normal distribution $N_{\Delta_n, \theta_0}$ by $\Phi_n$ (for $\Delta_n, \theta_0$, see lemma 4.2). For $K \subset \mathbb{R}^k$, conditional versions are denoted $\Pi_{n,K}$ and $\Phi_{n,K}$ respectively (assuming that $\Pi_n(K) > 0$ and $\Phi_n(K) > 0$, of course).

Let $K \subset \Theta$ be a ball centered on the origin in $\mathbb{R}^k$. For every open neighbourhood $U \subset \Theta$ of $\theta_0$, $\theta_0 + n^{-1/2} K \subset U$ for large enough $n$. Since $\theta_0$ is an internal point of $\Theta$, we can define, for large enough $n$, the random functions $f_n : K \times K \to \mathbb{R}$ by:

$$f_n(g, h) = \left( 1 - \frac{\phi_n(h) s_n(g)}{\phi_n(g) s_n(h)} \right)^+, \quad (4.8)$$

where $\phi_n, s_n$ are the Jacobian and determinant of the score function, respectively.
4.2 Le Cam’s Bernstein-von Mises theorem

where $\phi_n : K \to \mathbb{R}$ is the Lebesgue density of the (randomly located) distribution
$N_{\Delta_n, \theta_0}I_{\theta_0}^{-1}$, $\pi_n : K \to \mathbb{R}$ is the Lebesgue density of the prior for the centred and
rescaled parameter $h$ and $s_n : K \to \mathbb{R}$ equals the likelihood product:

$$s_n(h) = \prod_{i=1}^{n} \frac{P_{\theta_0 + h/\sqrt{n}}(X_i)}{P_{\theta_0}}.$$

Since the model is stochastically LAN by assumption, we have for every random sequence $(h_n) \subset K$:

$$\log s_n(h_n) = h_n \bar{g}_n \bar{\ell}_{\theta_0} - \frac{1}{2} h_n^T I_{\theta_0} h_n + o_P(1),$$

$$\log \phi_n(h_n) = -\frac{1}{2} (h_n - \Delta_n, \bar{\theta}_0)^T I_{\theta_0}(h_n - \Delta_n, \bar{\theta}_0) + \text{constant}.$$  

For any two sequences $(h_n), (g_n) \subset K$, $\pi_n(g_n)/\pi_n(h_n) \to 1$ as $n \to \infty$. Combining
this with the above display we see that:

$$\log \frac{\phi_n(h_n)}{\phi_n(g_n)} s_n(g_n) \pi_n(g_n) s_n(h_n) \pi_n(h_n)$$

$$= -h_n \bar{g}_n \bar{\ell}_{\theta_0} + \frac{1}{2} h_n^T I_{\theta_0} h_n + g_n \bar{g}_n \bar{\ell}_{\theta_0} - \frac{1}{2} g_n^T I_{\theta_0} g_n + o_P(1)$$

$$- \frac{1}{2} (h_n - \Delta_n, \bar{\theta}_0)^T I_{\theta_0}(h_n - \Delta_n, \bar{\theta}_0) + \frac{1}{2} (g_n - \Delta_n, \bar{\theta}_0)^T I_{\theta_0}(g_n - \Delta_n, \bar{\theta}_0)$$

$$= o_P(1)$$

as $n \to \infty$ by lemma 4.2. Since $x \mapsto (1 - e^x)^+$ is continuous on $\mathbb{R}$, we conclude that
for every pair of random sequences $(g_n, h_n) \subset K \times K$:

$$f_n(g_n, h_n) \overset{P_0}{\to} 0, \quad (n \to \infty).$$

For fixed, large enough $n$, $P_0$-almost-sure continuity of $(g, h) \mapsto \log s_n(g)/s_n(h)$ on
$K \times K$ is guaranteed by the stochastic LAN-condition. Each of the locations $\Delta_n, \theta_0$
for $\Phi_n$ is is tight, so $(g, h) \mapsto \phi_n(g)/\phi_n(h)$ is continuous on all of $K \times K P_0$-almost-
surely. Continuity (in a neighbourhood of $\theta_0$) and positivity of the prior density
guarantee that this holds for $(g, h) \mapsto \pi_n(g)/\pi_n(h)$ as well. We conclude that for
large enough $n$, the random functions $f_n$ are continuous on $K \times K$, $P_0$-almost-surely.
Application of lemma 4.4 then leads to the conclusion that,

$$\sup_{g, h \in K} f_n(g, h) \overset{P_0}{\to} 0, \quad (n \to \infty). \quad (4.8)$$

Since $K$ contains a neighbourhood of 0, $\phi_n(K) > 0$ is guaranteed. Let $\Xi_n$ denote the
event that $\Pi_n(K) > 0$. Let $\eta > 0$ be given and based on that, define the events:

$$\Omega_n = \{ \omega : \sup_{g, h \in K} f_n(g, h) \leq \eta \}.$$
Consider the expression (recall that the total-variation norm is bounded by 2):

$$P_0^p \| \Pi_n^K - \Phi_n^K \|_{L_1} \leq P_0^p \| \Pi_n^K - \Phi_n^K \|_{L_1} \Omega_n \cap \Xi_n + 2P_0^p (\Xi_n \setminus \Omega_n). \quad (4.9)$$

As a result of (4.8) the latter term is $o(1)$ as $n \to \infty$. The remaining term on the r.h.s. can be calculated as follows:

$$\frac{1}{2} P_0^p \| \Pi_n^K - \Phi_n^K \|_{L_1} \Omega_n \cap \Xi_n = P_0^p \int_{\mathcal{K}_n} \left( 1 - \frac{d\Phi^K_n}{d\Pi^K_n} \right) + d\Pi^K_n \Omega_n \cap \Xi_n$$

$$= P_0^p \int_{\mathcal{K}_n} \left( 1 - s_n(g) \frac{\pi_n(g)}{\pi_n(h)} \phi^K_n(h) \phi^K_n(g) d\Phi^K_n \right) + d\Pi^K_n \Omega_n \cap \Xi_n.$$

Note that for all $g, h \in K$ we have $\phi^K_n(h) / \phi^K_n(g) = \phi_n(h) / \phi_n(g)$ since, on $K$, $\phi^K_n$ differs from $\phi_n$ only by a normalisation factor. We use Jensen’s inequality (with respect to the $\Phi_n^K$-expectation) for the (convex) function $x \mapsto (1 - x)_+$ to derive:

$$\frac{1}{2} P_0^p \| \Pi_n^K - \Phi_n^K \|_{L_1} \Omega_n \cap \Xi_n \leq P_0^p \int_{\mathcal{K}_n} \left( 1 - s_n(g) \frac{\pi_n(g) \phi_n(h)}{\pi_n(h) \phi_n(g)} \phi^K_n(h) \phi^K_n(g) d\Phi^K_n \right) + d\Pi^K_n \Omega_n \cap \Xi_n$$

$$\leq P_0^p \int_{g, h \in K} f_n(g, h) \Omega_n \cap \Xi_n d\Phi^K_n(g) + d\Pi^K_n(h) \Omega_n \cap \Xi_n \leq \eta.$$  

Combination with (4.9) shows that for all compact $K \subset \mathbb{R}^d$ containing a neighbourhood of 0,

$$P_0^p \| \Pi_n^K - \Phi_n^K \|_{L_1} \to 0.$$

Now let $(K_m)$ be a sequence of origin-centred balls in $\mathbb{R}^k$ with radii $M_m \to \infty$. For each $m \geq 1$, the above display holds, so if we choose a sequence of balls $(K_n)$ that traverses the sequence $K_m$ slowly enough, convergence to zero can still be guaranteed. Moreover, the corresponding events $\Xi_n = \{ \omega : \Pi_n(K_n) > 0 \}$ satisfy $P_0^p(\Xi_n) \to 1$ as a result of (4.6). We conclude that there exists a sequence of radii $(M_n)$ such that $M_n \to \infty$ and

$$P_0^p \| \Pi_n^K_m - \Phi_n^K_m \| \to 0, \quad (4.10)$$

(where it is understood that the conditional probabilities on the l.h.s. are well-defined on sets of probability growing to one). Combining (4.6) and lemma 4.6, we then use lemma 4.5 to conclude that:

$$P_0^p \| \Pi_n^K_m - \Phi_n^K_m \| \to 0,$$

which implies (4.7).
4.2 Le Cam’s Bernstein-von Mises theorem

Aside from a slightly stronger smoothness property in the form of the stochastic LAN condition, theorem 4.5 appears to require more than theorem 4.4, in the sense that it requires posterior consistency at rate \( n^{-1/2} \) rather than the (fixed) tests for consistency. The following lemma shows that, assuming smoothness, the latter condition is enough to satisfy the former. Its proof is based on the construction of a score test that fills in the “gap” left between the fixed-alternative tests and the growing alternative \( \| \theta - \theta_0 \| \geq n^{-1/2} M_n \). A proof is given in the next chapter (see section 7.5), in the more general, misspecified situation.

**Lemma 4.3.** Assume that \( \Theta \subset \mathbb{R}^k \) is open and that the model \( \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \) is identifiable and dominated. Assume that the model is locally asymptotically normal at \( \theta_0 \) with non-singular Fisher information \( I_{\theta_0} \) and that the prior is thick at \( \theta_0 \). Furthermore, suppose that there exists a test sequence \((\phi_n)\) such that,

\[
P_{0_n} \phi_n \to 0, \quad \sup_{\| \theta - \theta_0 \| > \varepsilon} P_{0_n} (1 - \phi_n) \to 0.
\]

Then the posterior converges at rate \( n^{-1/2} \), i.e. for every sequence \( M_n \to \infty \),

\[
\Pi \left( \| \theta - \theta_0 \| \geq n^{-1/2} M_n \mid X_1, \ldots, X_n \right) \to 0.
\]

### 4.2.2 Three subsidiary lemmas

The proof of theorem 4.5 also makes use of the following three lemmas. It is not necessary that the sample is i.i.d., and we denote the true data-distribution by \( P_{0,n} \).

**Lemma 4.4.** Let \((f_n)\) be a sequence of random functions \( K \to \mathbb{R} \), where \( K \) is compact. Assume that for large enough \( n \geq 1 \), \( f_n \) is continuous \( P_{0,n} \)-almost-surely. Then the following are equivalent:

(i) Uniform convergence in probability:

\[
\sup_{h \in K} |f_n(h)| \xrightarrow{P_{0,n}} 0,
\]

(ii) Convergence along any random sequence \((h_n) \subset K\) in probability:

\[
f_n(h_n) \xrightarrow{P_{0,n}} 0,
\]

as \( n \to \infty \).

**Proof.** \(((ii) \Rightarrow (i), by contradiction.)\) Assume that there exist \( \delta, \varepsilon > 0 \) such that:

\[
\limsup_{n \to \infty} P_{0,n} \left( \sup_{h \in K} |f_n(h)| > \delta \right) = \varepsilon.
\]
Since the functions \( f_n \) are continuous \( P_{0,n} \)-almost-surely, there exists (with \( P_{0,n} \)-probability one) a sequence \((\tilde{h}_n)\) such that for every \( n \geq 1 \), \( \tilde{h}_n \in K \) and

\[
|f_n(\tilde{h}_n)| = \sup_{h \in K} |f_n(h)|.
\]

Consequently, for this particular random sequence in \( K \), we have:

\[
\limsup_{n \to \infty} P_{0,n}\left( |f_n(\tilde{h}_n)| > \delta \right) = \varepsilon > 0.
\]

which contradicts \((ii)\). \((i) \Rightarrow (ii)\).

Conversely, given a random sequence \((h_n) \subset K\), and for every \( \delta > 0 \),

\[
P_{0,n}\left( \sup_{h \in K} |f_n(h)| > \delta \right) \geq P_{0,n}\left( |f_n(h_n)| > \delta \right).
\]

Given \((i)\), the l.h.s. converges to zero and hence so does the r.h.s..

The next lemma shows that given two sequences of probability measures, a sequence of balls that grows fast enough can be used conditionally to calculate the difference in total-variational distance, even when the sequences consist of random measures.

**Lemma 4.5.** Let \((\Pi_n)\) and \((\Phi_n)\) be two sequences of random probability measures on \( \mathbb{R}^k \). Let \((K_n)\) be a sequence of subsets of \( \mathbb{R}^k \) such that

\[
\Pi_n(\mathbb{R}^k \setminus K_n) \xrightarrow{P_{0,n}} 0, \quad \Phi_n(\mathbb{R}^k \setminus K_n) \xrightarrow{P_{0,n}} 0. \tag{4.11}
\]

Then

\[
\|\Pi_n - \Phi_n\| - \|\Pi_n^{K_n} - \Phi_n^{K_n}\| \xrightarrow{P_{0,n}} 0. \tag{4.12}
\]

**Proof.** Let \( K \), a measurable subset of \( \mathbb{R}^k \) and \( n \geq 1 \) be given and assume that \( \Pi_n(K) > 0 \) and \( \Phi_n(K) > 0 \). Then for any measurable \( B \subset \mathbb{R}^k \) we have:

\[
|\Pi_n(B) - \Pi_n^{K_n}(B)| = \left| \Pi_n(B) - \frac{\Pi_n(B \cap K)}{\Pi_n(K)} \right| = \left| \Pi_n(B \cap (\mathbb{R}^k \setminus K)) + (1 - \Pi_n(K)^{-1}) \Pi_n(B \cap K) \right| \leq \Pi_n(B \cap (\mathbb{R}^k \setminus K)) + \Pi_n(\mathbb{R}^k \setminus K)\Pi_n(B) \leq 2\Pi_n(\mathbb{R}^k \setminus K).
\]

and hence also:

\[
\left| (\Pi_n(B) - \Pi_n^{K_n}(B)) - (\Phi_n(B) - \Phi_n^{K_n}(B)) \right| \leq 2\left( \Pi_n(\mathbb{R}^k \setminus K) + \Phi_n(\mathbb{R}^k \setminus K) \right). \tag{4.13}
\]

As a result of the triangle inequality, we then find that the difference in total-variation distances between \( \Pi_n \) and \( \Phi_n \) on the one hand and \( \Pi_n^{K_n} \) and \( \Phi_n^{K_n} \) on the other is bounded above by the expression on the right in the above display (which is independent of \( B \)).
Define $A_n, B_n$ to be the events that $\Pi_n(K_n) > 0$, $\Phi_n(K_n) > 0$ respectively. On $\Xi_n = A_n \cap B_n$, $\Pi_n^{K_n}$ and $\Phi_n^{K_n}$ are well-defined probability measures. Assumption (4.11) guarantees that $P_n^0(\Xi_n)$ converges to 1. Restricting attention to the event $\Xi_n$ in the above upon substitution of the sequence $(K_n)$ and using (4.11) for the limit of (4.13) we find (4.12), where it is understood that the conditional probabilities on the l.h.s. are well-defined with probability growing to 1.

To apply the above lemma in the concluding steps of the proof of theorem 4.5, rate conditions for both posterior and limiting normal sequences are needed. The rate condition (4.6) for the posterior is assumed and the following lemma demonstrates that its analog for the sequence of normals is satisfied when the sequence of centre points $\Delta_n, \theta_0$ is uniformly tight.

**Lemma 4.6.** Let $K_n$ be a sequence of balls centred on the origin with radii $M_n \to \infty$. Let $(\Phi_n)$ be a sequence of normal distributions (with fixed covariance matrix $V$) located at the random points $(\Delta_n) \subset \mathbb{R}^k$. If the sequence $\Delta_n$ is uniformly tight, then:

$$\Phi_n(\mathbb{R}^k \setminus K_n) = N_{\Delta_n, V}(\mathbb{R}^k \setminus K_n) \xrightarrow{P_0, n} 0.$$ 

**Proof.** Let $\delta > 0$ be given. Uniform tightness of the sequence $(\Delta_n)$ implies the existence of a constant $L > 0$ such that:

$$\sup_{n \geq 1} P_0, n(\|\Delta_n\| \geq L) \leq \delta.$$ 

For all $n \geq 1$, call $A_n = \{\|\Delta_n\| \geq L\}$. Let $\mu \in \mathbb{R}^k$ be given. Since $N(\mu, V)$ is tight, for every given $\varepsilon > 0$, there exists a constant $L'$ such that $N_{\mu, V}(B(\mu, L')) \geq 1 - \varepsilon$ (where $B(\mu, L')$ defines a ball of radius $L'$ around the point $\mu$). Assuming that $\mu \leq L$, $B(\mu, L') \subset B(0, L + L')$ so that with $M = L + L'$, $N_{\mu, V}(B(0, M)) \geq 1 - \varepsilon$ for all $\mu$ such that $\|\mu\| \leq L$. Choose $N \geq 1$ such that $M_n \geq M$ for all $n \geq N$. Let $\mu \leq L$ be given. Then:

$$P_0, n(\Phi_n(\mathbb{R}^k \setminus B(0, M_n)) > \varepsilon) \leq \delta + P_0, n\left(\{N_{\Delta_n, V}(B(0, M_n)^c) > \varepsilon\} \cap A_n^c\right) \quad (4.14)$$

Note that on the complement of $A_n$, $\|\Delta_n\| < L$, so:

$$N_{\Delta_n, V}(B(0, M_n)^c) \leq 1 - N_{\Delta_n, V}(B(0, M)) \leq 1 - \inf_{\|\mu\| \leq L} N_{\mu, V}(B(0, M)) \leq \varepsilon,$$

and we conclude that the last term on the r.h.s. of (4.14) equals zero.
4.3 Model selection with the BIC criterion [EMPTY]

4.4 Semi-parametric Bernstein-von Mises theorems [EMPTY]

4.5 Exercises

4.1. Assume that \( n^{1/2}(\hat{\theta}_n - \theta_0) \sim N(0, I_{\theta_0}^{-1}) \). Show that the ellipsoids (4.1) are of minimal Lebesgue measure among all subsets of asymptotic coverage \( 1 - \alpha \).

4.2. Consider Hodges’s estimators \( S_n \) of example 4.1. Show that, for any rate sequence \( (\varepsilon_n) \), \( \varepsilon_n \downarrow 0 \), \( (S_n - 0) \xrightarrow{0-w.} 0 \).

4.3. Let \( \Theta = (0, \infty) \) and let \( \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \) be the model of Poisson distributions \( P_\theta \) with means \( \theta \). Let the data be an i.i.d. sample from \( P_{\theta_0} \) for some \( \theta_0 \in \Theta \). Show that this model is LAN for all \( \theta \).

4.4. Let \( \Theta = \mathbb{R} \) and let \( \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \) be the model of normal distributions \( N(\theta, 1) \) of unit variance with means \( \theta \). Let the data be an i.i.d. sample from \( P_{\theta_0} \) for some \( \theta_0 \in \Theta \). Show that this model is LAN for all \( \theta \).

4.5. Let \( f \) be a Lebesgue density on \( \mathbb{R} \) that is symmetric around the origin. Define the model \( \mathcal{P} = \{ P_{\mu, \sigma} : \mu \in \mathbb{R}, \sigma \in (0, \infty) \} \) by densities \( f_{\mu, \sigma}(x) = \sigma^{-1}f((x - \mu)/\sigma) \). Show that the Fisher information matrix is diagonal.

4.6. Let \( P \) and \( Q \) be probability measures on a measurable space \((\mathcal{X}, \mathcal{B})\),
   a. Show that there exists a \( \sigma \)-finite measure \( \mu \) such that \( P, Q \ll \mu \).
   b. Using Radon-Nikodym derivatives \( p = dP/d\mu \) and \( q = dQ/d\mu \), prove that,
      \[ \sup_{B \in \mathcal{B}} |P(B) - Q(B)| = \int |p - q| \, d\mu. \]
   c. Show that, for any sequence \((Q_n)\) of probability measures on \((\mathcal{X}, \mathcal{B})\), there exists a probability measure \( P \) that dominates all \( Q_n \), \( (n \geq 1) \).
   d. Use the completeness of \( L_1(\mathcal{X}, \mathcal{B}, P) \) to show that the metric space \( M(\mathcal{X}, \mathcal{B}) \) of all probability measures on \((\mathcal{X}, \mathcal{B})\) is complete in the topology of total variation.

4.7. Let \( \Theta = (0, \infty) \) and \( \mathcal{P} = \{ N(0, \theta^2) : \theta \in \Theta \} \). Let \( \Pi \) be a thick prior on \( \Theta \). Show that this model satisfies the conditions of the Bernstein-von Mises theorem 4.4. Find the problematic range of parameter values in this model. (Hint: calculate the Fisher information, find a problematic limit for it and describe the effect on the limiting sequence of normal distributions for parameter values close to the problematic limit.)

4.8. Approximation in measure from within by compact subsets has a deep background in analysis. Central is the notion of a Radon measure. Given a Hausdorff topological space \( \Theta \), a Radon measure \( \Pi \) is a Borel measure that is locally finite (meaning that any \( \theta \in \Theta \) has a neighbourhood \( U \) such that \( \Pi(U) < \infty \)) and inner
regular (meaning that for any measurable subset $S \subset \Theta$ and any $\varepsilon > 0$, there exists a compact $K \subset S$ such that $\mu(S \setminus K) < \varepsilon$). Show that any probability measure on a Polish space is Radon. (*NB: This statement can be generalized to continuous images of Polish spaces, known as Souslin spaces.*)

4.9. Prove the following: for $\theta \in \Theta = \mathbb{R}$, let $F_\theta(x) = (1 - e^{-|x-\theta|}) \vee 0$ be the standard exponential distribution function located at $\theta$. Assume that $X_1, X_2, \ldots$ form an i.i.d. sample from $F_{\theta_0}$, for some $\theta_0$. Let $\Pi$ be a thick prior on $\Theta$. Then the associated posterior distribution satisfies, with $h = n(\theta - \theta_0)$,

$$\sup_A \left| \Pi_n \left( h \in A \mid X_1, \ldots, X_n \right) - \text{Exp}_{n(\hat{\theta}_n - \theta_0)}(A) \right| \xrightarrow{\theta_0} 0,$$

where $\hat{\theta}_n = X_{(1)}$ is the maximum likelihood estimate for $\theta_0$ and $\text{Exp}_{a}$ denotes the standard negative exponential distribution located at $a$. (*NB: This is an example of an irregular estimation problem: clearly the model does not depend on $\Theta$ in a differentiable way. Inspection of the assertion shows that the rate of convergence is $n^{-1}$ rather than $n^{-1/2}$, the rate of convergence in regular situations. In addition, the limiting shape of the posterior is not normal but exponential.*)
Chapter 5
Model misspecification

Generally speaking, statistical analysis requires a choice of a model, which may not include the frequentist true distribution of the data. Throughout most of what has preceded, we have assumed that the model $\mathcal{P}$ is well-specified, c.f. definition 1.6. In asymptotic context, well-specification translates into the assumption that for every $n \geq 1$, the true distribution $P_{0,n}$ of the sample $X^n$ lies in the $n$-th model $\mathcal{P}_n$. In the more specific situation that these models are parametrized with the help of a single parameter space $\Theta$ by maps $\Theta \rightarrow \mathcal{P}_n : \theta \mapsto P_{\theta,n}$, well-specification is expressed through the stronger assumption that there exists a $\theta_0 \in \Theta$ such that $P_{0,n} = P_{\theta_0,n}$ for all $n \geq 1$. Assumptions of this nature, which concern the unknown quantity of interest $\theta_0$ directly, are accepted as an article of faith in most frequentist statistical procedures.

In the proofs of theorems, it is rarely a problem if there is no single $\theta_0 \in \Theta$ to explain all $P_{0,n}$, because often on can prove exactly the same for $n$-dependent $\theta_{0,n}$ such that $P_{0,n} = P_{\theta_{0,n}}$. But what happens to our statistical procedures in the far-worse case when,

$$P_{0,n} \notin \mathcal{P}_n,$$

the true distribution of the data does not even lie in the model? The smaller the models $\mathcal{P}_n$, the more stringent the assumption that the model is well-specified. Especially when we consider a parametric models, when $\Theta \subset \mathbb{R}^k$, chances are that the models we have for the true distribution of the data are misspecified, c.f. (5.1). Commonly ignored in practice, this fact implies that many statistical procedures are carried out with misspecified models. Theorems assuming well-specification are used regardless, seldom leading to significant problems, which raises the question: “Why? What can be said about the reliability of statistical tools in misspecified situations?”

When we dissociate the definition of the model from sufficient assumptions on $P_0$ for our tools to work, we explore the maximal extent of their applicability properly. For example, in the case of an i.i.d. problem: given the model $\mathcal{P}$ and $P_0^n$-distributed sample $X_1, \ldots, X_n$, the assertion holds if $P_0$ satisfies stated requirements. Usually, 'stated requirements’ are satisfied by all $P \in \mathcal{P}$ and in optima forma also by a
large set of other distributions, so that the misspecified theorem generalises its well specified version.

5.1 Misspecification in smooth parametric models

In this chapter, we address the misspecification question in the particular, parametric case of the Bernstein-von Mises theorem of chapter 4. In the asymptotic limit, the posterior distribution of a parameter in misspecified LAN parametric models can still be approximated by a random normal distribution, but Bayesian credible sets cease to be valid as approximate confidence sets if the model is misspecified. We obtain the result under conditions that are comparable to those in the well-specified situation: uniform testability against fixed alternatives and sufficient prior mass in neighbourhoods of the point of convergence. The rate of convergence is considered in detail, with special attention for the existence and construction of suitable test sequences.

5.1.1 Misspecified maximum likelihood estimation

One class of point estimators that generalises easily to the misspecified situation is that of $M$-estimators. Consider the sequence $\hat{\theta}_n$ of (near-)maximisers of the functions $M_n$ over a model $\Theta$, with,

$$M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i),$$

for some $P_0$-integrable $m_\theta(x)$ and an i.i.d. sample $X_1, \ldots, X_n \sim P_0^n$ (e.g. $m_\theta(x) = \log p_\theta(x)$ for the maximum-likelihood estimator). Under certain (rather stringent, e.g. Wald’s) conditions [219], $\hat{\theta}_n$ converges to the maximum $\theta^*$ of the function $\theta \mapsto P_0 m_\theta(X)$. If a unique maximum does not exist, the model is flawed more seriously than just by misspecification. In that case we say the model is not $P_0$-identifiable: on the basis of a sample from $P_0$ it is impossible to distinguish between $P_1, P_2 \in \mathcal{P}$, e.g. when $P_1$ and $P_2$ differ only on a null-set of $P_0$.

We consider more closely maximum-likelihood estimation in a smooth parametric model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ without the assumption that $P_1 \in \mathcal{P}$. Consider $\hat{\theta}_n$ that are (near-)maximisers of $\theta \mapsto \frac{1}{n} \sum \log p_\theta(X_i)$. Under noted conditions for consistent $M$-estimation, the sequence $\hat{\theta}_n$ converges to the point $\theta^* \in \Theta$ that minimises the so-called Kullback-Leibler divergence of $P_0$ with respect to $P_0$:

$$\theta \mapsto -P_0 \log \frac{p_\theta}{p_0},$$
over the model $\Theta$. That $\theta^*$ does not correspond to the true distribution $P_0$ is inconsistent: the maximum-likelihood procedure defines the ‘best’ approximation of $P_0$ within $\mathcal{P}$ to be the point of minimal Kullback-Leibler divergence (note that other choices for $x \mapsto m_\theta(x)$ would lead to different ‘projections’). The asymptotic behaviour of the maximum-likelihood procedure is postponed to lemmas 5.1 and 5.2, but we note here that under regularity conditions, maximum likelihood estimation converges to $\theta^*$ in an asymptotically normal way,

$$
\sqrt{n}(\hat{\theta}_n - \theta^*) \overset{P}{\rightarrow} N(0, V_{\theta^*}^{-1}) \quad (5.2)
$$

where $V_{\theta^*}$ is the second-order coefficient in the Taylor expansion of the Kullback-Leibler divergence (assumed non-singular) and $I_{\theta}$ denotes the Fisher information at $\theta^*$ (see, for instance, Huber (1967) [123]). In the well-specified case, $V_{\theta^*}$ equals $I_{\theta}$ and the asymptotic variance reduces to a single instance of the inverse Fisher information, but that cancellation does not occur in the misspecified case.

### 5.1.2 The misspecified Bernstein-von Mises theorem

Consistency of posterior distributions and asymptotic normality of the posterior mean under misspecification have been considered in Berk (1966, 1970) [21, 22] and Bunke and Milhaud (1998) [49]. The behaviour of the full posterior distribution was studied in Kleijn and van der Vaart (2004) [143]. Here we follow the latter and derive the asymptotic normality of the full posterior distribution in the misspecified situation under conditions comparable to those obtained in the well-specified case of section 4.2. We focus on dominated models for i.i.d. data where the posterior distribution follows (2.12) and we assume that the observations are sampled from a density $p_0$ that is not necessarily of the form $p_{\theta_0}$ for some $\theta_0$. It is shown that the Bernstein-Von Mises can be extended to this situation, in the form,

$$
P_0 \| \Pi_n(B \mid X_1, \ldots, X_n) - N(\hat{\theta}_n(nV_{\theta^*})^{-1}(B)) \| \rightarrow 0,
$$

where $\theta^*$ is the parameter value $\theta$ minimizing the Kullback-Leibler divergence $\theta \mapsto P_0 \log(p_0/p_{\theta})$ (provided it exists and is unique), $V_{\theta^*}$ is minus the second derivative matrix of this map, and $\hat{\theta}_n$ are suitable estimators.

According to (5.2), maximum likelihood estimators in the misspecified model are asymptotically normal with mean zero and covariance matrix given by $\Sigma_{\theta^*} = V_{\theta^*}^{-1}(P_0^T \ell_{\theta}^T)^{-1} V_{\theta^*}$. The corresponding Wald-type confidence sets (c.f.e.g. ellipsoids (4.1)) for the misspecified parameter take the form $\hat{\theta}_n + C/\sqrt{n}$ for $C$ a central set in the Gaussian distribution. Because the covariance matrix $V_{\theta^*}$ appearing in the misspecified Bernstein-Von Mises theorem is not the sandwich covariance matrix, credible sets of posterior probability $1 - \alpha$ do not correspond to the misspecified Wald sets. Although they are correctly centered, they may have the wrong...
width, and are in general not $1 - \alpha$-confidence sets. We show below by example that the credible sets may over- or under-cover, depending on the true distribution of the observations and the model, and to extreme amounts.

This chapter’s presentation is split into two parts: in section 5.2 we derive normality of the posterior given that it shrinks at a $\sqrt{n}$-rate of posterior convergence (theorem 5.1). We actually state this result for the general situation of locally asymptotically normal (LAN) models, and next specify to the i.i.d. case. Next in section 7.5 we discuss results guaranteeing the desired rate of convergence, where we first show sufficiency of existence of certain tests (theorem 5.4), and next construct appropriate tests (theorem 5.6).

### 5.2 Posterior limit distribution

Throughout the presentation of the misspecified Bernstein-von Mises theorem and its consequences, we denote the model parametrizations by $\theta \mapsto P_\theta^{(n)}$ (when possibly non-i.i.d.) and by $\theta \mapsto P_\theta^0$ (when i.i.d.), deviating from the notation $\theta \mapsto P_{\theta,n}$ used elsewhere, for typographic reasons.

#### 5.2.1 Asymptotic normality in LAN models

Let $\Theta$ be an open subset of $\mathbb{R}^d$ parameterising statistical models $\{P_\theta^{(n)} : \theta \in \Theta\}$. For simplicity, we assume that for each $n$ there exists a single measure that dominates all measures $P_\theta^{(n)}$ as well as a “true measure” $P_\theta^0$, and we assume that there exist densities $p_\theta^{(n)}$ and $p_\theta^0$ such that the maps $(\theta, x) \mapsto p_\theta^{(n)}$ are measurable.

Generalizing definition 4.1, we consider models satisfying a stochastic LAN condition at an inner point $\theta^* \in \Theta$ and relative to a given norming rate $\delta_n \to 0$: there exist random vectors $\Delta_n, \theta^*$ and non-singular matrices $V_{\theta^*}$ such that the sequence $\Delta_n, \theta^*$ is bounded in probability and for every compact set $K \subset \mathbb{R}^d$,

$$\sup_{h \in K} \log \frac{p_\theta^{(n)}(X^{(n)}) - \theta^* V_{\theta^*} \Delta_n, \theta^*}{p_\theta^0} \to 0, \quad (5.3)$$

in $P_\theta^0$-probability.

The prior measure $\Pi$ on $\Theta$ is assumed to be a probability measure with Lebesgue-density $\pi$, continuous and positive on a neighbourhood of a given point $\theta^*$. Priors satisfying these criteria assign enough mass to (sufficiently small) balls around $\theta^*$ to allow for optimal rates of convergence of the posterior if certain regularity conditions are met (see section 7.5).
Like before, the posterior based on an observation $X^{(n)}$ is denoted $\Pi_n(\cdot | X^{(n)})$: for every Borel set $A$,

$$
\Pi_n(A \mid X^{(n)}) = \int_A p_\theta^{(n)}(X^{(n)}) \pi(\theta) \, d\theta / \int_{\Theta} p_\theta^{(n)}(X^{(n)}) \pi(\theta) \, d\theta.
$$  

(5.4)

We stress that both definition (5.3) and the assertion of theorem 5.1 below involve convergence in $P_0^{(n)}$-probability, the true distribution of the data.

**Theorem 5.1.** Assume that (5.3) holds for some $\theta^* \in \Theta$ and let the prior $\Pi$ be as indicated above. Furthermore, assume that for every sequence of constants $M_n \to \infty$, we have:

$$
P_0^{(n)} \Pi_n(\| \theta - \theta^* \| > \delta_n M_n \mid X^{(n)}) \to 0.
$$  

(5.5)

Then the sequence of posteriors converges to a sequence of normal distributions in total variation:

$$
\sup_B \left| \Pi_n(\| \theta - \theta^* \| \mid \delta_n \in B \mid X^{(n)}) - N_{\Delta_n \theta^*, V_{\theta^*}}^{-1}(B) \right| \to 0.
$$  

(5.6)

**Proof.** The proof is identical to that of theorem 4.4, with a few small changes: the local parameter $h$ is now defined with the help of the rate $\delta_n$. Throughout the proof we denote the posterior for $H = (\theta - \theta^*)/\delta_n$ given $X^{(n)}$ by $\Pi_n$, which follows from that for $\theta$ by $\Pi_n( H \in B \mid X^{(n)}) = \Pi_n( (\theta - \theta^*)/\delta_n \in B \mid X^{(n)})$ for all Borel sets $B$. Furthermore, we denote the normal distribution $N_{\Delta_n \theta^*, V_{\theta^*}}^{-1}$ by $\Phi_n$. For a compact subset $K \subset \mathbb{R}^d$ such that $\Pi_n( H \in K \mid X^{(n)}) > 0$, we define a conditional version $\Pi^K_n$ of $\Pi_n$ by $\Pi^K_n(B \mid X^{(n)}) = \Pi_n( B \cap K \mid X^{(n)}) / \Pi_n( K \mid X^{(n)})$. Similarly we defined a conditional measure $\Phi^K_n$ corresponding to $\Phi_n$. Then, following the proof of theorem 4.4, it is noted that,

$$
\log \frac{\phi_n(h_n) s_n(g_n) \pi_n(g_n)}{\phi_n(g_n) s_n(h_n) \pi_n(h_n)} = (g_n - h_n)^T V_{\theta^*} \Delta_n \theta^* + \frac{1}{2} h_n^T V_{\theta^*} h_n - \frac{1}{2} g_n^T V_{\theta^*} g_n + o_P(1)
$$

$$
- \frac{1}{2} (h_n - \Delta_n \theta^*)^T V_{\theta^*} (h_n - \Delta_n \theta^*) + \frac{1}{2} (g_n - \Delta_n \theta^*)^T V_{\theta^*} (g_n - \Delta_n \theta^*)
$$

$$
= o_P(1),
$$

as $n \to \infty$. The rest of the proof is identical.

Condition (5.5) fixes the rate of convergence of the posterior distribution to be that occurring in the LAN property. Sufficient conditions to satisfy (5.5) in the case of i.i.d. observations are given in section 7.5.
5.2.2 Asymptotic normality in the i.i.d. case

Consider the situation that the observation is a vector $X^{(n)} = (X_1, \ldots, X_n)$ and the model consists of $n$-fold product measures $P^{(n)}_\theta = P^n_\theta$, where the components $P_\theta$ are given by densities $p_\theta$ such that the maps $(\theta, x) \mapsto p_\theta(x)$ are measurable and $\theta \mapsto p_\theta$ is smooth (in the sense of lemma 5.1). Assume that the observations form an i.i.d. sample from a distribution $P_0$ with density $p_0$ relative to a common dominating measure. Assume that the Kullback-Leibler divergence of the model relative to $P_0$ is finite and minimized at $\theta^* \in \Theta$, i.e.:

$$-P_0 \log \frac{p_{\theta^*}}{p_0} = \inf_{\theta \in \Theta} -P_0 \log \frac{p_\theta}{p_0} < \infty. \quad (5.7)$$

In this situation we set $\delta_n = n^{-1/2}$ and use $\Delta_n, \theta^* = V_n^{-1} G_n \ell_\theta^*$ as the centering sequence (where $\ell_\theta^*$ denotes the score function of the model $\theta \mapsto p_\theta$ at $\theta^*$ and $G_n = \sqrt{n}(P_n - P_0)$ is the empirical process).

Lemmas that establish the LAN expansion (5.3) usually assume a well-specified model, whereas current interest requires local asymptotic normality in misspecified situations. To that end we consider the following lemma which gives sufficient conditions.

**Lemma 5.1.** If the function $\theta \mapsto \log p_\theta(X_1)$ is differentiable at $\theta^*$ in $P_0$-probability with derivative $\ell_\theta^*(X_1)$ and:

(i) there is an open neighbourhood $U$ of $\theta^*$ and a square-integrable function $m_\theta$ such that for all $\theta_1, \theta_2 \in U$:

$$\left| \log \frac{p_{\theta_1}}{p_{\theta_2}} \right| \leq m_{\theta^*} \| \theta_1 - \theta_2 \|, \quad (P_0 - a.s.), \quad (5.8)$$

(ii) the Kullback-Leibler divergence with respect to $P_0$ has a 2nd-order Taylor-expansion around $\theta^*$:

$$-P_0 \log \frac{p_{\theta^*}}{p_0} = \frac{1}{2} (\theta - \theta^*) V_{\theta^*} (\theta - \theta^*) + o(\| \theta - \theta^* \|^2), \quad (\theta \to \theta^*), \quad (5.9)$$

where $V_{\theta^*}$ is a positive-definite $d \times d$-matrix,

then (5.3) holds with $\delta_n = n^{-1/2}$ and $\Delta_n, \theta^* = V_{\theta^*}^{-1} G_n \ell_{\theta^*}$. Furthermore, the score function is bounded as follows:

$$\| \ell_\theta^*(X) \| \leq m_{\theta^*} (X), \quad (P_0 - a.s.). \quad (5.10)$$

Finally, we have:

$$P_0 \ell_{\theta^*} = \frac{\partial}{\partial \theta} [P_0 \log p_\theta]_{\theta = \theta^*} = 0. \quad (5.11)$$

**Proof.** Using lemma 19.31 in Van der Vaart (1998) [219] for $\ell_\theta(X) = \log p_\theta(X)$, the conditions of which are satisfied by assumption, we see that for any sequence
(\langle h_n \rangle)$ that is bounded in \(P_0\)-probability:
\[
\mathbb{G}_n \left( \sqrt{n} \left( \ell_{\theta^* + (h_n/\sqrt{n})} - \ell_{\theta^*} \right) - h_n^T \ell_{\theta^*} \right) \xrightarrow{P_0} 0.
\] (5.12)

Hence, we see that,
\[
n \log p_{\theta^*} + h_n/\sqrt{n} \cdot p_{\theta^*} - o_P(1).
\] (5.13)


As noted earlier, lemma 5.2 implies that for consistent maximum-likelihood estimators the distribution of \(\sqrt{n}(\hat{\theta}_n - \theta^*)\) has a normal limit with mean zero and covariance,
\[
\Sigma = V_{\theta^*}^{-1} P_0[\ell_{\theta^*} \ell_{\theta^*}^T] V_{\theta^*}^{-1}.
\]

More important for present purposes, however, is the fact that according to (5.13), this sequence differs from \(\Delta_n, \theta^*\) only by a term of order \(o_P(1)\). Like before, the assertion of the Bernstein-Von Mises theorem can also be formulated as,
\[
\sup_B \left| \Pi_n (\theta \in B \mid X_1, \ldots, X_n) - N_{\hat{\theta}_n, \Sigma} (B) \right| \xrightarrow{P_0} 0,
\]
which demonstrates the usual interpretation of the Bernstein-Von Mises theorem most clearly: the sequence of posteriors resembles more-and-more closely a sequence of “sharpening” normal distributions centred at the maximum-likelihood estimator.
estimators. More generally, any sequence of estimators satisfying (5.13) (i.e. any **best-regular** estimator sequence) may be used to centre the normal limit sequence. The conditions for lemma 5.2 are close to the conditions of the above Bernstein-Von Mises theorem. In the well-specified situation the Lipschitz condition (5.8) can be relaxed slightly and replaced by the condition of differentiability in quadratic mean.

It was noted in the introduction that the mismatch of the asymptotic covariance matrix $V_{\theta}^{-1} P_0 |\ell_{\theta}^{\prime} \ell_{\theta}^{\prime\prime} | V_{\theta}^{-1}$ and the limiting covariance matrix $V_{\theta}^{-1}$ in the Bernstein-Von Mises theorem causes Bayesian credible sets to differ from confidence sets at the nominal level. The following example shows that both over- and under-covering may occur.

**Example 5.1.** Let $P_\theta$ be the normal distribution with mean $\theta$ and variance 1, and let the true distribution possess mean zero and variance $\sigma^2 > 0$. Then $\theta^* = 0$, $P_0^{\ell_{\theta}^{\prime}} = \sigma^2$ and $V_{\theta} = 1$. It follows that the radius of the $1-\alpha$-Bayesian credible set is $z_\alpha / \sqrt{n}$, whereas a $1-\alpha$-confidence set around the mean has radius $z_\alpha \sigma / \sqrt{n}$. Depending on $\sigma^2 \leq 1$ or $\sigma^2 > 1$, the credible set can have coverage arbitrarily close to 0 or 1.

### 5.2.3 Asymptotic normality of point-estimators

Having discussed the posterior distributional limit, a natural question concerns the asymptotic properties of point-estimators derived from the posterior, like the posterior mean and median.

Based on the Bernstein-Von Mises assertion (5.6) alone, one sees that any functional $f : \mathcal{P} \mapsto \mathbb{R}$, continuous relative to the total-variational norm, when applied to the sequence of posterior laws, converges to $f$ applied to the normal limit distribution. Another general consideration follows from a generic construction of point-estimates from posteriors and demonstrate that posterior consistency at rate $\delta_n$ implies frequentist consistency at rate $\delta_n$.

**Theorem 5.2.** Let $X_1, \ldots, X_n$ be distributed i.i.d.-$P_0$ and let $\Pi_n(\cdot | X_1, \ldots, X_n)$ denote a sequence of posterior distributions on $\Theta$ that satisfies (5.5). Then there exist Bayesian point-estimators $\hat{\theta}_n$ such that:

$$\delta_n^{-1}(\hat{\theta}_n - \theta^*) = O_P(1),$$

(5.14)

i.e. $\hat{\theta}_n$ is consistent and converges to $\theta^*$ at rate $\delta_n$.

**Proof.** Define $\hat{\theta}_n$ to be the centre of a smallest ball that contains posterior mass at least $1/2$ (see remark 2.5). Because the ball around $\theta^*$ of radius $\delta_n M_n$ contains posterior mass tending to 1, the radius of a smallest ball must be bounded by $\delta_n M_n$ and the smallest ball must intersect the ball of radius $\delta_n M_n$ around $\theta^*$ with probability tending to 1. This shows that $||\hat{\theta}_n - \theta^*|| \leq 2 \delta_n M_n$ with probability tending to one.
This general point is more appropriate in non-parametric context and the above existence theorem does not pertain to the most widely-used Bayesian point-estimators. Asymptotic normality of the posterior mean in a misspecified model has been analysed in Bunke and Milhaud (1998) [49]; here, we generalize their discussion and prove asymptotic normality and efficiency for a class of point-estimators defined by a general loss function, which includes the posterior mean and median.

Let \( \ell : \mathbb{R}^k \to [0, \infty) \) be a loss-function with the following properties:

\[
\sup_{\|h\| \leq M} \inf_{\|h\| > 2M} \ell(h),
\]

with strict inequality for some \( M \). Furthermore, we assume that \( \ell \) is subpolynomial, i.e. for some \( p > 0 \),

\[
\ell(h) \leq 1 + \|h\|^p.
\]

(5.15)

Define the estimators \( \hat{\theta}_n \) as the formal Bayes estimators (see definition 2.10) that minimize,

\[
t \mapsto \int \ell(\sqrt{n}(t - \theta)) \, d\Pi_n(\theta|X_1, \ldots, X_n).
\]

**Theorem 5.3.** Assume that the model satisfies (5.3) for some \( \theta^* \in \Theta \) and that the conditions of theorems 5.4 are satisfied. Let \( \ell : \mathbb{R}^k \to [0, \infty) \) be a loss-function with the properties listed and assume that \( \int \|\theta\|^p \, d\Pi(\theta) < \infty \). Then under \( P_0 \), the sequence \( \sqrt{n}(\hat{\theta}_n - \theta^*) \) converges weakly to the minimizer of,

\[
t \mapsto Z(t) = \int \ell(t - h) \, dN_{X,V^{-1}} (h),
\]

where \( X \sim N(0, V^{-1}) \) and \( \ell(\theta) = \ell_{\theta} \ell_{\theta}^T | V^{-1} \). Provided that any two minimizers of this process coincide almost-surely, then \( \sqrt{n}(\hat{\theta}_n - \theta^*) \) converges weakly to \( X \) under \( P_0 \).

**Proof.** The theorem can be proved along the same lines as theorem 10.8 in [219]. The main difference is in proving that, for any \( M_n \to \infty \),

\[
U_n := \int_{\|h\| > M_n} \|h\|^p \, d\Pi_n(h|X_1, \ldots, X_n) \xrightarrow{P_0} 0.
\]

(5.16)

Here, abusing notation, we write \( d\Pi_n(h|X_1, \ldots, X_n) \) to denote integrals relative to the posterior distribution of the local parameter \( h = \sqrt{n}(\theta - \theta^*) \). Under misspecification a new proof is required, for which we extend the proof of theorem 5.4 below.

Once (5.16) is established, the proof continues as follows. The variable \( \hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta) \) is the maximizer of the process \( t \mapsto \int \ell(t - h) \, d\Pi_n(h|X_1, \ldots, X_n) \). Then \( \hat{h}_n = O_{P_0}(1) \). Fix some compact set \( K \) and for given \( M > 0 \) define the processes
These expressions tend to zero. The last assertion of the theorem follows, because \( \ell \) has subpolynomial tails, integrable with respect to \( N_{W,0} \). Be-
cause \( \ell \) is minimized uniquely by \( X \), the assumed unicity of maxima of these sample paths, enables

the argmax theorem (corollary 5.58 in [219]) and we conclude that \( \hat{h} \) is the minimizer of \( Z(t) \).

For the proof of (5.16) we adopt the notation of theorem 5.4. The tests \( \omega_n \) employed there can be taken nonrandomized without loss of generality (otherwise replace them for instance by \( 1_{\omega_n>1/2} \) and then \( U_n \omega_n \) tends to zero in probability by the only fact that \( \omega_n \) does so. Thus (5.16) is proved once it is established that, with \( \varepsilon_n = M_n / \sqrt{n} \),

\[
P_0^n (1 - \omega_n) 1_{\Omega_n} \int_{\varepsilon_n \leq \|\theta - \theta^*\| < \varepsilon} n^{p/2} \|\theta - \theta^*\|^p d\Pi_n (\theta \mid X_1, \ldots, X_n) \to 0,
\]

\[
P_0^n (1 - \omega_n) 1_{\Omega_n} \int_{\|\theta - \theta^*\| \geq \varepsilon} n^{p/2} \|\theta - \theta^*\|^p d\Pi_n (\theta \mid X_1, \ldots, X_n) \to 0.
\]

We can use bounds as in the proof of theorem 5.4, but instead of at (5.21) and (5.21) we arrive at the bounds

\[
\frac{e^p (a_n^2 (1 + C) - D \varepsilon^2)}{\Pi (B(a_n, \theta^*; P_0))} n^{p/2} \int \|\theta - \theta^*\|^p d\Pi (\theta),
\]

\[
K' e^{-\frac{1}{2} n D \varepsilon^2} \sum_{j=1}^{\infty} n^{p/2} (j + 1)^{d + p} e^{p \varepsilon} e^{-nD(j^2 - 1)\varepsilon^2}.
\]

These expressions tend to zero. The last assertion of the theorem follows, because for a subconvex loss function the process \( Z \) is minimized uniquely by \( X \), as a consequence of lemma 4.1).
5.3 Rate of convergence

In a Bayesian context, the rate of convergence is defined as the maximal pace at which balls around the point of convergence can be shrunk to radius zero while still capturing a posterior mass that converges to one asymptotically. Current interest lies in the fact that the Bernstein-Von Mises theorem of the previous section formulates condition (5.5) (with $\delta_n = n^{-1/2}$),

$$\Pi_n \left( \| \theta - \theta^* \| \geq M_n / \sqrt{n} \mid X_1, \ldots, X_n \right) \overset{P}{\rightarrow} 0,$$

for all $M_n \rightarrow \infty$. A convenient way of establishing the above is through the condition that suitable test sequences exist. As has been shown in a well-specified context in Ghosal et al. (2000) [101] and under misspecification in Kleijn and Van der Vaart (2003) [142], the most important requirement for convergence of the posterior at a certain rate is the existence of a test-sequence that separates the point of convergence from the complements of balls shrinking at said rate.

This is also the approach we follow here: we show that the sequence of posterior probabilities in the above display converges to zero in $P_0$-probability if a test sequence exists that is suitable in the sense given above (see the proof of theorem 5.4). However, under the regularity conditions that were formulated to establish local asymptotic normality under misspecification in the previous section, more can be said: not complements of shrinking balls, but fixed alternatives are to be suitably testable against $P_0$, thus relaxing the testing condition considerably. Locally, the construction relies on score-tests to separate the point of convergence from complements of neighbourhoods shrinking at rate $1/\sqrt{n}$, using Bernstein’s inequality to obtain exponential power. The tests for fixed alternatives are used to extend those local tests to the full model.

In this section we prove that a prior mass condition and suitable test sequences suffice to prove convergence at the rate required for the Bernstein-Von Mises theorem formulated in section 5.2. The theorem that begins the next subsection summarizes the conclusion. Throughout the section we consider the i.i.d. case, with notation as in section 5.2.2.

5.3.1 Posterior rate of convergence

With use of theorem 5.6, we formulate a theorem that ensures $\sqrt{n}$-rate of convergence for the posterior distributions of smooth, testable models with sufficient prior mass around the point of convergence. The testability condition is formulated using measures $Q_\theta$, defined by,

$$Q_\theta(A) = P_0 \left( \frac{P_\theta}{P_{\theta^*}} A \right),$$

for all $A \in \mathcal{A}$ and all $\theta \in \Theta$. Note that all $Q_\theta$ are dominated by $P_0$ and that $Q_{\theta^*} = P_0$. Also note that if the model is well-specified, then $P_{\theta^*} = P_0$ and $Q_\theta = P_\theta$ for all $\theta$. 

Therefore the use of \( Q_0 \) instead of \( P_0 \) to formulate the testing condition is relevant only in the misspecified situation (see Kleijn and Van der Vaart (2006) [142] for more on this subject). The proof of theorem 5.4 makes use of Kullback-Leibler neighbourhoods of \( \theta^* \) of the form:

\[
B(\varepsilon, \theta^*: P_0) = \left\{ \theta \in \Theta : -P_0 \log \frac{P_0}{P_{\theta^*}} \leq \varepsilon^2, P_0 \left( \log \frac{P_0}{P_{\theta^*}} \right)^2 \leq \varepsilon^2 \right\},
\]

for some \( \varepsilon > 0 \).

**Theorem 5.4.** Assume that the model satisfies the smoothness conditions of lemma 5.1, where in addition, it is required that \( P_0(p_\theta / p_{\theta^*}) < \infty \) for all \( \theta \) in a neighbourhood of \( \theta^* \) and \( P_0(e^{s \theta^*}) < \infty \) for some \( s > 0 \). Assume that the prior possesses a density that is continuous and positive in a neighbourhood of \( \theta^* \). Furthermore, assume that \( P_0 \ell_{\theta^*} \ell_{\theta^*} \) is invertible and that for every \( \varepsilon > 0 \) there exists a sequence of tests \((\phi_n)\) such that:

\[
P_0^\varepsilon \phi_n \to 0, \quad \sup_{\|\theta - \theta^*\| \geq \varepsilon} Q_0^\varepsilon (1 - \phi_n) \to 0. \tag{5.18}
\]

Then the posterior converges at rate \( 1/\sqrt{n} \), i.e. for every sequence \((M_n)\), \( M_n \to \infty \):

\[
\Pi_n \left\{ \theta \in \Theta : \|\theta - \theta^*\| \geq M_n / \sqrt{n} \mid X_1, X_2, \ldots, X_n \right\} \to 0.
\]

**Proof.** Let \((M_n)\) be given, and define the sequence \((\varepsilon_n)\) by \( \varepsilon_n = M_n / \sqrt{n} \). According to theorem 5.6 there exists a sequence of tests \((\omega_n)\) and constants \( D > 0 \) and \( \varepsilon > 0 \) such that (5.23) holds. We use these tests to split the \( P_0^\varepsilon \)-expectation of the posterior measure as follows:

\[
P_0^\varepsilon \Pi \left\{ \theta : \|\theta - \theta^*\| \geq \varepsilon_n \mid X_1, X_2, \ldots, X_n \right\} \leq P_0^\varepsilon \omega_n + P_0^\varepsilon (1 - \omega_n) \Pi \left\{ \theta : \|\theta - \theta^*\| \geq \varepsilon_n \mid X_1, X_2, \ldots, X_n \right\}.
\]

The first term is of order \( o(1) \) as \( n \to \infty \) by (5.23). Given a constant \( \varepsilon > 0 \) (to be specified later), the second term can be decomposed as:

\[
P_0^\varepsilon (1 - \omega_n) \Pi \left\{ \theta : \|\theta - \theta^*\| \geq \varepsilon_n \mid X_1, X_2, \ldots, X_n \right\}
\]

\[
= P_0^\varepsilon (1 - \omega_n) \Pi \left\{ \theta : \|\theta - \theta^*\| \geq \varepsilon \mid X_1, X_2, \ldots, X_n \right\}
\]

\[
+ P_0^\varepsilon (1 - \omega_n) \Pi \left\{ \theta : \varepsilon_n \leq \|\theta - \theta^*\| < \varepsilon \mid X_1, X_2, \ldots, X_n \right\}. \tag{5.19}
\]

Given two constants \( M, M' > 0 \) (also to be specified at a later stage), we define the sequences \((a_n)\), \( a_n = M / \sqrt{\log n / n} \) and \((b_n)\), \( b_n = M' \varepsilon_n \). Based on \( a_n \) and \( b_n \), we define two sequences of events:

\[
\Xi_n = \left\{ \int_{\Theta} \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(x_i) d\Pi(\theta) \leq \Pi(B(a_n, \theta^*: P_0)) e^{-na^2/(1+C)} \right\},
\]

\[
\Omega_n = \left\{ \int_{\Theta} \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(x_i) d\Pi(\theta) \leq \Pi(B(b_n, \theta^*: P_0)) e^{-nb^2/(1+C)} \right\}.
\]
5.3 Rate of convergence

The sequence \((\mathcal{Z}_n)\) is used to split the first term on the r.h.s. of (5.19) and estimate it as follows:

\[
P_0^n(1 - \omega_n) \Pi(\theta : \|\theta - \theta^*\| \geq \varepsilon \mid X_1, X_2, \ldots, X_n)
\leq P_0^n(\mathcal{Z}_n) + P_0^n(1 - \omega_n) 1_{\Omega, \mathcal{Z}_n} \Pi(\theta : \|\theta - \theta^*\| \geq \varepsilon \mid X_1, X_2, \ldots, X_n).
\]

According to lemma 5.3, the first term is of order \(o(1)\) as \(n \to \infty\). The second term is estimated further with the use of lemmas 5.3, 5.4 and theorem 5.6: for some \(C > 0\),

\[
P_0^n(1 - \omega_n) 1_{\Omega, \mathcal{Z}_n} \Pi(\theta : \|\theta - \theta^*\| \geq \varepsilon \mid X_1, X_2, \ldots, X_n)
\leq \frac{e^{n a_2^2(1 + C)}}{\Pi(B(a_n, \theta^*; P_0))} \left( \int_{\theta : \|\theta - \theta^*\| \geq \varepsilon} Q_0^n(1 - \omega_n) d\Pi(\theta) \right)
\leq \frac{e^{n a_2^2(1 + C) - D \varepsilon^2}}{\Pi(B(a_n, \theta^*; P_0))} \Pi(\theta : \|\theta - \theta^*\| \geq \varepsilon).
\]

Note that \(a_2^2(1 + C) - D \varepsilon^2 \leq -a_2^2(1 + C)\) for large enough \(n\), so that:

\[
\frac{e^{n a_2^2(1 + C) - D \varepsilon^2}}{\Pi(B(a_n, \theta^*; P_0))} \leq K^{-1} e^{-n a_2^2(1 + C)(a_n)^{-d}} \leq \frac{1}{M^{d/2} K} (\log n)^{-d/2} n^{-M^2(1 + C) + \frac{d}{2}},
\]

for large enough \(n\), using (5.22). A large enough choice for the constant \(M\) then ensures that the expression on the l.h.s. in the next-to-last display is of order \(o(1)\) as \(n \to \infty\).

The sequence \((\Omega_n)\) is used to split the second term on the r.h.s. of (5.19) after which we estimate it in a similar manner. Again the term that derives from \(1_{\Omega_n}\) is of order \(o(1)\), and

\[
P_0^n(1 - \omega_n) 1_{\Omega_n} \Pi(\theta : \varepsilon_n \leq \|\theta - \theta^*\| < \varepsilon \mid X_1, X_2, \ldots, X_n)
\leq \frac{e^{n a_2^2(1 + C)}}{\Pi(B(b_n, \theta^*; P_0))} \sum_{j=1}^{J} \int_{A_{n,j}} Q_0^n(1 - \omega_n) d\Pi(\theta),
\]

where we have split the domain of integration into spherical shells \(A_{n,j}\), \(1 \leq j \leq J\), with \(J\) the smallest integer such that \((J + 1)\varepsilon_n > \varepsilon\): \(A_{n,j} = \{ \theta : j \varepsilon_n \leq \|\theta - \theta^*\| \leq ((j + 1)\varepsilon_n) \wedge \varepsilon \}\). Applying theorem 5.6 to each of the shells separately, we obtain:

\[
P_0^n(1 - \omega_n) 1_{\Omega_n} \Pi(\theta : \varepsilon_n \leq \|\theta - \theta^*\| < \varepsilon \mid X_1, X_2, \ldots, X_n)
= \sum_{j=1}^{J} e^{n a_2^2(1 + C) - n D \varepsilon_n^2} \sup_{\theta \in A_{n,j}} \frac{Q_0^n(1 - \omega_n)}{\Pi(B(b_n, \theta^*; P_0))} \Pi(A_{n,j})
\leq \sum_{j=1}^{J} e^{n a_2^2(1 + C) - n D \varepsilon_n^2} \frac{\Pi(\theta : \|\theta - \theta^*\| \leq (j + 1)\varepsilon_n)}{\Pi(B(b_n, \theta^*; P_0))}.
\]
For a small enough $\varepsilon$ and large enough $n$, the sets $\{ \theta : \| \theta - \theta^* \| \leq (j+1)\varepsilon_n \}$ all fall within the neighbourhood $U$ of $\theta^*$ on which the prior density $\pi$ is continuous. Hence $\pi$ is uniformly bounded by a constant $R > 0$ and we see that: $\Pi( \{ \theta : \| \theta - \theta^* \| \leq (j+1)\varepsilon_n \} \leq RV_d(j+1)^d \varepsilon_n^d$, where $V_d$ is the Lebesgue-volume of the $d$-dimensional ball of unit radius. Combining this with (5.22), there exists a constant $K' > 0$ such that, with $M' < \sqrt{D}/2(1+C)$:

$$P^\theta_0(1-\omega_n)1_{\Omega_0} \Pi(\theta : \varepsilon_n \leq \| \theta - \theta^* \| < \varepsilon \mid X_1, \ldots, X_n) \leq K'e^{-\frac{1}{2}nD\varepsilon_n^2} \sum_{j=1}^\infty (j+1)^d e^{-nD(j^2-1)\varepsilon_n^2},$$

(5.21)

for large enough $n$. The series is convergent and we conclude that this term is also of order $o(1)$ as $n \to \infty$.

Consistent testability of the type (5.18) appears to be a weak requirement because the form of the tests is arbitrary. Consistent testability is of course one of Schwarz’ conditions for consistency (see chapters 6 and 7). To exemplify its power we show in the next theorem that the tests exist as soon as the parameter set is compact and the model is suitably continuous in the parameter.

**Theorem 5.5.** Assume that $\Theta$ is compact and that $\theta^*$ is a unique point of minimum of $\theta \mapsto -P_0 \log p_0$. Furthermore assume that $P_0(p_0/p_0^\theta) < \infty$ for all $\theta \in \Theta$ and that the map,

$$\theta \mapsto P_0\left( \frac{p_0}{p_0^\theta \cdot \theta^\theta} \right),$$

is continuous at $\theta_1$ for every $s$ in a left neighbourhood of $I$, for every $\theta_1$. Then there exist tests $\phi_n$ satisfying (5.18). A sufficient condition is that for every $\theta_1 \in \Theta$ the maps $\theta \mapsto p_0/p_0^\theta$ and $\theta \mapsto p_0/p_0^\theta$ are continuous in $L_1(P_0)$ at $\theta = \theta_1$. 

**Proof.** For given $\theta_1 \neq \theta^*$ consider the tests,

$$\phi_{n, \theta_1} = 1\{\mathbb{P}_n \log (p_0/q_{\theta_1}) < 0\}.$$

Because $\mathbb{P}_n \log (p_0/q_{\theta_1}) \to P_0 \log (p_0/q_{\theta_1})$ in $P_0$:probability by the law of large numbers, and $P_0 \log (p_0/q_{\theta_1}) = P_0 \log (p_0^\theta/p_0^\theta) > 0$ for $\theta_1 \neq \theta^*$ by the definition of $\theta^*$ we have that $P_0^n \phi_{n, \theta_1} \to 0$ as $n \to \infty$. By Markov’s inequality we have that,

$$Q^\theta_0(1 - \phi_{n, \theta_1}) = Q^\theta_0\left( e^{\mathbb{P}_n \log (p_0/q_{\theta_1})} > 1 \right) \leq Q^\theta_0 e^{\mathbb{P}_n \log (p_0/q_{\theta_1})^s} = \rho(\theta_1, \theta, s)^n,$$

for $\rho(\theta_1, \theta, s) = \int P_0 q_{\theta_1} d\mu$. It is known [142] that the Hellinger transform (see also [208] and [162]) $s \mapsto \rho(\theta_1, \theta_1, s)$ tends to $P_0(q_{\theta_1} > 0) = P_0(p_0 > 0)$ as $s \uparrow 1$ and has derivative from the left equal to $P_0 \log (q_{\theta_1}/p_0)1_{q_{\theta_1} > 0} = P_0 \log (p_0^\theta/p_0^\theta)1_{p_0^\theta > 0}$ at $s = 1$. We have that either $P_0(p_0^\theta > 0) < 1$ or $P_0(p_0^\theta > 0) = 1$ and $P_0 \log (p_0^\theta/p_0^\theta)1_{p_0^\theta > 0} = P_0 \log (p_0^\theta/p_0^\theta) < 0$ (or both). In all cases there exists $s_{\theta_1} < 1$ arbitrarily close to 1.
5.3 Rate of convergence

such that \( \rho(\theta_1, \theta, s) < 1 \). By assumption the map \( \theta \mapsto \rho(\theta_1, \theta, s) \) is continuous at \( \theta_1 \). Therefore, for every \( \theta_1 \) there exists an open neighbourhood \( G_{\theta_1} \) such that,

\[
 r_{\theta_1} = \sup_{\theta \in G_{\theta_1}} \rho(\theta_1, \theta, s) < 1.
\]

The set \( \{ \theta \in \Theta : \| \theta - \theta^* \| \geq \varepsilon \} \) is compact and hence can be covered with finitely many sets of the type \( G_{\theta_1} \), for \( 1 = 1, \ldots, k \). We now define

\[
 \phi_n = \max_{i=1,\ldots,k} \phi_{n,\theta_i}.
\]

This test satisfies

\[
 P_0^n \phi_n \leq \sum_{i=1}^k P_{\theta_i}^n \phi_{n,\theta_i} \to 0,
\]

\[
 Q_{\theta_i}^n (1 - \phi_n) \leq \max_{i=1}^k Q_{\theta_i}^n (1 - \phi_{n,\theta_i}) \leq \max_{i=1}^k r_{\theta_i}^n \to 0,
\]

uniformly in \( \theta \in \cup_{i=1}^k G_{\theta_i} \). Therefore the tests \( \phi_n \) satisfy the requirements. To prove the last assertion we write \( \rho(\theta_1, \theta, s) = P_0(p_{\theta} / p_{\theta_1})^s (p_{\theta} / p_{\theta^*})^{1-s} \). Continuity of the maps \( \theta \mapsto (p_{\theta} / p_{\theta_1}) \) and \( \theta \mapsto (p_{\theta} / p_{\theta^*}) \) in \( L_1(P_0) \) can be seen to imply the required continuity of the map \( \theta \mapsto \rho(\theta_1, \theta, s) \).

Beyond compactness it appears impossible to give mere qualitative sufficient conditions, like continuity, for consistent testability. For “natural” parameterizations it ought to be true that distant parameters (outside a given compact) are the easy ones to test for (and a test designed for a given compact ought to be consistent even for points outside the compact), but this depends on the structure of the model. Alternatively, many models would allow a suitable compactification to which the preceding result can be applied, but we omit a discussion.

In the proof of theorem 5.4, lower bounds in probability on the denominators of posterior probabilities are needed, as provided by the following lemma.

**Lemma 5.3.** For given \( \varepsilon > 0 \) and \( \theta^* \in \Theta \) such that \( P_0 \log (p_0 / p_{\theta^*}) < \infty \) define \( B(\varepsilon, \theta^*; P_0) \) by (5.17). Then for every \( C > 0 \) and probability measure \( \Pi \) on \( \Theta \):

\[
 P_0^n \left( \int_{\Theta} \prod_{i=1}^n \frac{p_{\theta}}{p_{\theta^*}} (X_i) d\Pi(\theta) \leq \Pi (B(\varepsilon, \theta^*; P_0)) e^{-nC(1+C)} \right) \leq \frac{1}{C^2n\varepsilon^2}.
\]

**Proof.** This lemma can also be found as lemma 7.1 in Kleijn and Van der Vaart (2003) [142].

Moreover, the prior mass of the Kullback-Leibler neighbourhoods \( B(\varepsilon, \theta^*; P_0) \) can be lower-bounded if we make the regularity assumptions for the model used in section 5.2 and the assumption that the prior has a Lebesgue density that is well-behaved at \( \theta^* \).
Lemma 5.4. Under the smoothness conditions of lemma 5.1 and assuming that the prior density $\pi$ is continuous and strictly positive in $\theta^*$, there exists a constant $K > 0$ such that the prior mass of the Kullback-Leibler neighbourhoods $B(\epsilon, \theta^*; P_0)$ satisfies:

$$\Pi(B(\epsilon, \theta^*; P_0)) \geq K\epsilon^d,$$

for small enough $\epsilon > 0$.

Proof. As a result of the smoothness conditions, we have, for some constants $c_1, c_2 > 0$ and small enough $\|\theta - \theta^*\|$, 

$$-P_0\log(p_{\theta}/p_{\theta^*}) \leq c_1\|\theta - \theta^*\|^2, \quad P_0(\log(p_{\theta}/p_{\theta^*}))^2 \leq c_2\|\theta - \theta^*\|^2.$$ 

Defining $c = (1/c_1 \wedge 1/c_2)^{1/2}$, this implies that for small enough $\epsilon > 0$, $\{\theta \in \Theta : \|\theta - \theta^*\| \leq c\epsilon\} \subset B(\epsilon, \theta^*; P_0)$. Since the Lebesgue-density $\pi$ of the prior is continuous and strictly positive in $\theta^*$, we see that there exists a $\delta' > 0$ such that for all $0 < \delta \leq \delta'$: $\Pi(\theta \in \Theta : \|\theta - \theta^*\| \leq \delta) \geq \frac{1}{2}V_d\pi(\theta^*)\delta^d > 0$. Hence, for small enough $\epsilon$, $c\epsilon \leq \delta'$ and we obtain (5.22) upon combination.

5.3.2 Suitable test sequences

In this subsection we prove that the existence of test sequences (under misspecification) of uniform exponential power for complements of shrinking balls around $\theta^*$ versus $P_0$ (as needed in the proof of theorem 5.4), is guaranteed whenever asymptotically consistent test-sequences exist for complements of fixed balls around $\theta^*$ versus $P_0$ and the conditions of lemmas 5.1 and 5.5 are met. The following theorem is inspired by lemma 10.3 in Van der Vaart (1998) [219].

Theorem 5.6. Assume that the conditions of lemma 5.1 are satisfied, where in addition, it is required that $P_0(p_{\theta}/p_{\theta^*}) < \infty$ for all $\theta$ in a neighbourhood of $\theta^*$ and $P_0(e^{m_\theta}) < \infty$ for some $s > 0$. Furthermore, suppose that $P_0^{1/2}, E_0^{1/2}$, is invertible and for every $\epsilon > 0$ there exists a sequence of test functions $(\phi_n)$, such that:

$$P_0^n\phi_n \to 0, \quad \sup_{\{\theta : \|\theta - \theta^*\| \geq \epsilon\}} Q_0^n(1 - \phi_n) \to 0.$$ 

Then for every sequence $(M_n)$ such that $M_n \to \infty$ there exists a sequence of tests $(\omega_n)$ such that for some constants $D > 0, \epsilon > 0$ and large enough $n$:

$$P_0^n\omega_n \to 0, \quad Q_0^n(1 - \omega_n) \leq e^{-nD\|\theta - \theta^*\|^2 + \epsilon^2},$$

for all $\theta \in \Theta$ such that $\|\theta - \theta^*\| \geq M_n/\sqrt{n}$.

Proof. Let $(M_n)$ be given. We construct two sequences of tests: one sequence to test $P_0$ versus $\{Q_0 : \theta \in \Theta_1\}$ with $\Theta_1 = \{\theta \in \Theta : M_n/\sqrt{n} \leq \|\theta - \theta^*\| \leq \epsilon\}$, and the other to test $P_0$ versus $\{Q_0 : \theta \in \Theta_2\}$ with $\Theta_2 = \{\theta : \|\theta - \theta^*\| > \epsilon\}$, both uniformly with
exponential power (for a suitable choice of ε). We combine these sequences to test
P₀ versus \{Q_{θ} : \|θ - θ^\ast\| \geq M_n/\sqrt{n}\} uniformly with exponential power.

For the construction of the first sequence, a constant L > 0 is chosen to truncate the score-function component-wise (i.e. for all 1 ≤ k ≤ d, \((\ell_{θ^*})_k = 0\) if \((\ell_{θ^*})_k \geq L\) and \((\ell_{θ^*})_k = (\ell_{θ^*})_k\) otherwise) and we define:

$$\omega_{1,n} = 1\{\|P_n - P₀\|^{1/2} > \sqrt{M_n/n}\},$$

Because the function \(\ell_{θ^*}\) is square-integrable, we can ensure that the matrices
\(P₀(\ell_{θ^*})_k, P₀(\ell_{θ^*})_k^T\) and \(P₀(\ell_{θ^*})_k, (\ell_{θ^*})_k^T\) are arbitrarily close (for instance in operator norm) by a sufficiently large choice for the constant L. We fix such an L throughout the proof.

By the central limit theorem \(P₀\omega_{1,n} = P₀\left(\|\sqrt{n}(P_n - P₀)\|_{L^2}^2 > M_n\right) \to 0\). Turning to \(Qₜ(1 - \omega_{1,n})\) for \(θ ∈ Θₜ\), we note that for all θ:

$$Qₜ\left(\|P_n - P₀\|^{1/2} ≤ \sqrt{M_n/n}\right) = Qₜ\left(\sup_{v ∈ S}(P_n - P₀)\ell_{θ^*} ≤ \sqrt{M_n/n}\right)
\leq \inf_{v ∈ S}Qₜ\left(v^T(P_n - P₀)\ell_{θ^*} ≤ \sqrt{M_n/n}\right),$$

where S is the sphere of unity in \(R^d\). With the choice \(v = (θ - θ^\ast)/\|θ - θ^\ast\|\) as an upper bound for the r.h.s. in the above display, we note that:

$$Qₜ\left((θ - θ^\ast)^T(P_n - P₀)\ell_{θ^*} ≤ \sqrt{M_n/n}\|θ - θ^\ast\|\right)
= Qₜ\left((θ^\ast - θ)^T(P_n - Q₀)\ell_{θ^*} ≥ (θ - θ^\ast)^T(Q₀ - Q₀^*)\ell_{θ^*} - \sqrt{M_n/n}\|θ - θ^\ast\|\right),$$

where we have used the notation (for all θ ∈ Θₜ with small enough e > 0) \(Q₀ = \|Q₀\|^{-1}Q₀\) and the fact that \(P₀ = Q₀^* = Q₀^\ast\). By straightforward manipulation, we find:

$$(θ - θ^\ast)^T(Q₀ - Q₀^*)\ell_{θ^*}
= \frac{1}{P₀(p_θ/p_θ^\ast)}(θ - θ^\ast)^T(P₀((p_θ/p_θ^\ast) - 1)\ell_{θ^*} + (1 - P₀(p_θ/p_θ^\ast))P₀\ell_{θ^*}).$$

In view of lemma 5.5 and conditions (5.8), (5.9), \((P₀(p_θ/p_θ^\ast) - 1)\) is of order \(O(||θ - θ^\ast||^2)\) as \((θ → θ^\ast)\), which means that if we approximate the above display up to order \(o(||θ - θ^\ast||^2)\), we can limit attention on the r.h.s. to the first term in the last factor and equate the first factor to 1. Furthermore, using the differentiability of \(θ → \log P₀(p_θ/p_θ^\ast)\), condition (5.8) and lemma 5.5, we see that:

$$P₀\left(\|\frac{p_θ}{p_θ^\ast} - 1 - (θ - θ^\ast)^T\ell_{θ^*}\|\right)
\leq P₀\left(\|\frac{p_θ}{p_θ^\ast} - 1 - \log \frac{p_θ}{p_θ^\ast}\|\right) + P₀\left(\|\log \frac{p_θ}{p_θ^\ast} - (θ - θ^\ast)^T\ell_{θ^*}\|\right).$$
which is \(o(\|\theta - \theta^*\|)\). Also note that since \(M_n \to \infty\) and for all \(\theta \in \Theta_1\), \(\|\theta - \theta^*\| \geq M_n / \sqrt{n}, -\|\theta - \theta^*\| \sqrt{M_n/n} \geq -\|\theta - \theta^*\|^2 (M_n)^{-1/2}\). Summarizing the above and combining with the remark made at the beginning of the proof concerning the choice of \(L\), we find that for every \(\delta > 0\), there exist choices of \(\varepsilon > 0\), \(L > 0\) and \(N \geq 1\) such that for all \(n \geq N\) and all \(\theta \in \Theta_1\):

\[
(\theta - \theta^*)^T (\tilde{Q}_\theta - \bar{Q}_\theta) \hat{I}_n - \sqrt{M_n/n} \|\theta - \theta^*\| \\
\geq (\theta - \theta^*)^T P_0(\tilde{\ell}_\theta \hat{I}_n) (\theta - \theta^*) - \delta \|\theta - \theta^*\|^2.
\]

We denote \(\Delta(\theta) = (\theta - \theta^*)^T P_0(\tilde{\ell}_\theta \hat{I}_n) (\theta - \theta^*)\) and since \(P_0(\tilde{\ell}_\theta \hat{I}_n)\) is strictly positive definite by assumption, its smallest eigenvalue \(c\) is greater than zero. Hence, 

\[-\delta \|\theta - \theta^*\|^2 \geq -\delta / c \Delta(\theta)\]

and there exists a constant \(r(\delta)\) (depending only on the matrix \(P_0(\tilde{\ell}_\theta \hat{I}_n)\)) and with the property that \(r(\delta) \to 1\) if \(\delta \to 0\) such that:

\[
Q^\varepsilon(1 - \omega_{1,n}) \leq Q^\varepsilon(\theta^* - \theta)^T (\tilde{Q}_\theta - \bar{Q}_\theta) \hat{I}_n \geq r(\delta) \Delta(\theta),
\]

for small enough \(\varepsilon\), large enough \(L\) and large enough \(n\), demonstrating that the type-II error is bounded above by the (unnormalized) tail probability \(Q^\varepsilon(W_n \geq r(\delta) \Delta(\theta))\) of the mean of the variables \(W_i = (\theta^* - \theta)^T (\tilde{I}_n - \bar{Q}_\theta \hat{I}_n)\), \(1 \leq i \leq n\), so that \(\bar{Q}_\theta W_i = 0\). The random variables \(W_i\) are independent and bounded since:

\[
\|W_i\| \leq \|\theta - \theta^*\| (\|\hat{I}_n\| \theta^* + \|\bar{Q}_\theta \hat{I}_n\|) \leq 2L \sqrt{d} \|\theta - \theta^*\|.
\]

The variance of \(W_i\) under \(\bar{Q}_\theta\) is expressed as follows:

\[
\mathbb{V} \mathbb{a} r_{\bar{Q}_\theta} W_i = (\theta - \theta^*)^T \left[ \tilde{Q}_\theta (\hat{\ell}_\theta \hat{I}_n) P_0(\tilde{\ell}_\theta \hat{I}_n) \right] (\theta - \theta^*).
\]

Using that \(P_0\hat{\ell}_\theta = 0\) (see (5.11)), the above can be estimated like before, with the result that there exists a constant \(s(\delta)\) (depending only on (the largest eigenvalue of) the matrix \(P_0\hat{\ell}_\theta \hat{I}_n\)) and with the property that \(s(\delta) \to 1\) as \(\delta \to 0\) such that:

\[
\mathbb{V} \mathbb{a} r_{\bar{Q}_\theta} (W_i) \leq s(\delta) \Delta(\theta),
\]

for small enough \(\varepsilon\) and large enough \(L\). We apply Bernstein’s inequality (see, for instance, Pollard (1984) [?], pp. 192–193) to obtain:

\[
Q^\varepsilon(1 - \omega_{1,n}) = \|Q^\varepsilon\| n Q^\varepsilon(W_1 + \ldots + W_n \geq nr(\delta) \Delta(\theta)) \\
\leq \|Q^\varepsilon\|^n \exp\left(-\frac{1}{2} \frac{r(\delta)^2 n \Delta(\theta)}{s(\delta) r(\delta) \|\theta - \theta^*\|}ight).
\]

(5.24)

The factor \(r(\delta)^2 (s(\delta) + \frac{1}{2} L \sqrt{d} \|\theta - \theta^*\| \|\theta - \theta^*\| r(\delta))^{-1}\) lies arbitrarily close to 1 for sufficiently small choices of \(\delta\) and \(\varepsilon\). As for the \(n\)-th power of the norm of \(Q^\varepsilon\), we use lemma 5.5, (5.8) and (5.9) to estimate the norm of \(Q^\varepsilon\) as follows:
\[ \| Q_\theta \| = 1 + R_0 \log \frac{P_\theta}{P_{\theta^*}} + \frac{1}{2} P_0 \left( \log \frac{P_\theta}{P_{\theta^*}} \right)^2 + o(\| \theta - \theta^* \|)^2 \]
\[ \leq 1 + R_0 \log \frac{P_\theta}{P_{\theta^*}} + \frac{1}{2} (\theta - \theta^*)^T R_0 \left( \ell_{\theta^*}^T \ell_{\theta^*} \right) (\theta - \theta^*) + o(\| \theta - \theta^* \|)^2 \] (5.25)
\[ \leq 1 - \frac{1}{2} (\theta - \theta^*)^T V_{\theta^*} (\theta - \theta^*) + \frac{1}{4} u(\delta) \Delta(\theta), \]
for some constant \( u(\delta) \) such that \( u(\delta) \to 1 \) if \( \delta \to 0 \). Because \( 1 + x \leq e^x \) for all \( x \in \mathbb{R} \), we obtain, for sufficiently small \( \| \theta - \theta^* \| \):
\[ Q_\theta^*(1 - \omega_{1,n}) \leq \exp \left( -\frac{n}{2} (\theta - \theta^*)^T V_{\theta^*} (\theta - \theta^*) + \frac{n}{2} u(\delta) - t(\delta)) \Delta(\theta) \right). \] (5.26)
Note that \( u(\delta) - t(\delta) \to 0 \) as \( \delta \to 0 \) and \( \Delta(\theta) \) is upper bounded by a multiple of \( \| \theta - \theta^* \|^2 \). Since \( V_{\theta^*} \) is assumed to be invertible, we conclude that there exists a constant \( C > 0 \) such that for large enough \( L \), small enough \( \varepsilon > 0 \) and large enough \( n \):
\[ Q_\theta^*(1 - \omega_{1,n}) \leq e^{-C_n \| \theta - \theta^* \|^2}. \] (5.27)
Concerning the range \( \| \theta - \theta^* \| > \varepsilon \), an asymptotically consistent uniform test-sequence of \( P_0 \) versus \( Q_\theta \) exists by assumption, and it is shown in chapter 8 that this implies the existence of tests \( \omega_{2,n} \) of uniformly exponential testing power. The sequence \( (\psi_n) \) is defined as the maximum of the two sequences defined above:
\[ \psi_n = \omega_{1,n} \lor \omega_{2,n} \text{ for all } n \geq 1, \] in which case \( P_\theta^0 \psi_n \leq P_\theta^0 \omega_{1,n} + P_\theta^0 \omega_{2,n} \to 0 \) and:
\[ \sup_{\theta \in A_n} Q_\theta^*(1 - \psi_n) = \sup_{\theta \in \Theta_1} Q_\theta^*(1 - \psi_n) \lor \sup_{\theta \in \Theta_2} Q_\theta^*(1 - \psi_n) \]
\[ \leq \sup_{\theta \in \Theta_1} Q_\theta^*(1 - \omega_{1,n}) \lor \sup_{\theta \in \Theta_2} Q_\theta^*(1 - \omega_{2,n}). \]
A suitable choice for the constant \( D > 0 \) lead to (5.23).

The following lemma is used in the proof of theorem 5.6 to control the behaviour of \( \| Q_\theta \| \) in neighbourhoods of \( \theta^* \).

**Lemma 5.5.** Assume that \( P_\theta(p_{\theta}/p_{\theta^*}) \) and \( -P_\theta \log(p_{\theta}/p_{\theta}) \) are finite for all \( \theta \) in a neighbourhood \( U' \) of \( \theta^* \). Furthermore, assume that there exist a measurable function \( m \) such that,
\[ \left| \log \frac{p_{\theta}}{p_{\theta^*}} \right| = m(\| \theta - \theta^* \|), \quad (P_0 - a.s.). \] (5.28)
for all \( \theta \in U' \) and such that \( P_\theta(e^{sm}) < \infty \) for some \( s > 0 \). Then,
\[ P_\theta \left[ \frac{p_\theta}{p_{\theta^*}} - 1 - \log \frac{p_\theta}{p_{\theta^*}} - \frac{1}{2} \left( \log \frac{p_\theta}{p_{\theta^*}} \right)^2 \right] = o(\| \theta - \theta^* \|^2). \]

**Proof.** The function \( R(x) \) defined by \( e^x = 1 + x + \frac{1}{2} x^2 + x^2 R(x) \) increases from \( -\frac{1}{2} \) in the limit \( (x \to -\infty) \) to \( \infty \) as \( (x \to \infty) \), with \( R(x) \to R(0) = 0 \) if \( x \to 0 \). We also have \( |R(0)| \leq R(x) \leq e^x / x^2 \) for all \( x > 0 \). The *l.h.s.* of the assertion of the lemma can be written as
\[ p_0 \left( \log \frac{p_\theta}{p_{\theta^*}} \right)^2 R \left( \log \frac{p_\theta}{p_{\theta^*}} \right) \leq \| \theta - \theta^* \|^2 p_0 \left( m^2 R(\epsilon_m) \right). \]

The expectation on the r.h.s. of the above display is bounded by \( p_0 m^2 R(\epsilon m) \) if \( \| \theta - \theta^* \| \leq \epsilon \). The functions \( m^2 R(\epsilon m) \) are dominated by \( e^{sm} \) for sufficiently small \( \epsilon \) and converge pointwise to \( m^2 R(0) = 0 \) as \( \epsilon \downarrow 0 \). The lemma then follows from the dominated convergence theorem.

5.4 Exercises [EMPTY]
Part II

Nonparametric Bayesian statistics
In non-parametric models the most basic mathematical properties relied upon by parametric methods are not present: for example, in finite dimensional parameter spaces the topology on the parameter space is always the subspace topology inherited form the (unique) norm topology on $\mathbb{R}^n$. Consequently questions regarding the way in which estimators converge asymptotically (e.g. consistently or not? With which rate of convergence and which limit distributions?) are unambiguous in parametric context. By contrast, in non-parametric setting no such default choice exists, raising the immediate question which topology suits our statistical purposes the best and most naturally. If one does not answer this question with sufficient circumspection topological problems occur. Examples include estimators that converge weakly but are inconsistent with respect to the total-variational norm. The Le Cam-Schwartz theorem singles out a weak topology stronger than Prokhorov’s but weaker than total-variation. Another curious property of finite-dimensional vector spaces that does not generalize to infinite-dimensions, is the existence of the (unique translation-invariant) Lebesgue measure. In the formulation of Bayesian statistics on finite-dimensional models, Lebesgue measure almost always serves to dominate both prior and posterior and is used to express homogeneity. In infinite-dimensional statistical models no such default choice for a dominating measure exists, giving rise both to mathematical depth and grave misconceptions.

In this second part of the book, we address non-parametric aspects of Bayesian statistics from the viewpoint of the frequentist. The subject has a long history, described with encyclopedic length and precision in Ghosal and van der Vaart (2018) [106] and summarized in chapter 6. In chapter 7, we question the fundamental nature of the material in chapter 6 and generalize it, to find that the differences between Bayesian and frequentist definitions of the first part, like credible versus confidence sets or Bayesian risk versus minimax risk, tend to dissipate when the sample size grows very large. More particularly, if the prior satisfies a property called remote contiguity, Bayesian methods give rise to asymptotic answers with frequentist validity. Central throughout the discussion is the existence of certain sequences of test functions that control posterior asymptotic convergence. That existence question is considered in depth in chapter 8, where we prove equivalent conditions and apply to consistent frequentist model selection. The last two chapters present applications: chapter 9 discusses errors-in-variables regression along the classical lines of [106] and chapter 6, while chapter 10 applies the methods of chapter 7 to the problem of community detection in a sparse stochastic block model.
Chapter 6
Non-parametric Bayesian statistics

In this chapter, we look exclusively at i.i.d. data in the form of samples $X^n = (X_1, X_2, \ldots, X_n)$ drawn from $P_0^n$ and a model $\mathcal{P}$ of single-observation distributions. We consider only metric models $(\mathcal{P}, d)$, mostly with $d$ equal to the total-variational or Hellinger metric. More complex dependence structures in the sample and more general model topologies are discussed in chapter 7.

6.1 Asymptotic statistics

In such i.i.d. situations, an estimation procedure prescribes a sequence of estimates $\hat{P}_n \in \mathcal{P}$, each calculated using only the first $n$ observations. More generally, any statistical procedure can be indexed by the size $n$ of the i.i.d. sample used to calculate it, leading to sequences of (parameter) estimates, tests, confidence regions, etcetera. Properties of such sequences reflect the behaviour of the estimation procedure with growing sample-size. In chapters 4 and 5 we have already seen examples in parametric setting.

An intuitively reasonable requirement of any estimation procedure is a property known as consistency: the sequence $\hat{P}_n$ approaches the true distribution $P_0$ to within arbitrary precision if the sample on which the estimation is based is made large enough. Similarly, samples of arbitrarily large size should enable one to test with power arbitrarily close to one and define arbitrarily small confidence regions. Further analysis of a consistent sequence $\hat{P}_n$ concerns the (suitably rescaled) distribution of the estimator-sequence around its point of convergence, much like the limiting shape of the posterior in the Bernstein-von Mises theorem. The mathematical formulation of these concepts is based on so-called limit theorems, which describe the behaviour of an estimation procedure in the limit that the sample size goes to infinity.

The study of the asymptotic regime of an estimation procedure is interesting for two reasons. Firstly, as was mentioned in the introduction of chapter 4, asymptotic results provide approximations to exact values: while exact, finite-sample results
are often intractable, the analysis of the large-sample limit is often still possible. (A valid objection to the use of asymptotic approximations is the fact that this practice does not provide any relation between the accuracy of the approximation and the size \( n \) of the sample for which answers are approximated. Limit theorems guarantee that approximation errors fall below arbitrarily small bounds for ‘large enough’ \( n \), but do not specify what ‘large enough’ is exactly. It is common practice to ignore this fact and assume that the asymptotic answer is a ‘good’ approximation for sample sizes that are deemed ‘large enough’.) Secondly, if we have several possible estimation procedures available for a certain problem, asymptotic behaviour provides us with the means to compare their performance on (very) large samples. For example, to choose between two consistent estimation procedures, one considers rate of convergence and properties of the limit distribution characterising the degree of concentration (like asymptotic variance or asymptotic risk).

In this section, we give an overview of some aspects of asymptotic point estimation that are important for the following chapters. It should be noted that this discussion is not intended to be comprehensive, nor is it stretched to full generality. For a more comprehensive presentation, the reader is referred to some of the excellent books devoted entirely to asymptotic statistics, like Ibragimov and Has’minskii (1981) [124], Le Cam and Yang (1990) [166] and Van der Vaart (1998) [219].

### 6.1.1 Consistency, rate and limit distribution

A sequence of estimators \( \hat{P}_n \) is said to be (asymptotically) consistent (respectively almost-surely consistent) if the estimator converges to the true distribution \( P_0 \) in probability (respectively almost-surely) as the sample-size goes to infinity.

**Definition 6.1.** A sequence \( \hat{P}_n \) of estimators in a (single-observation) model \( \mathcal{P} \) with metric \( d \) is said to be **consistent** in a point \( P_0 \in \mathcal{P} \), if:

\[
d(\hat{P}_n, P_0) \xrightarrow{P_0} 0.\]

and simply **consistent** if this holds for all points in \( \mathcal{P} \).

For example as in part 1, in the case of an identifiable, parametric model (with \( k \)-dimensional parameter set \( \Theta \) open in \( \mathbb{R}^k \)) defined by \( \mathcal{P} = \{ P_0 : \theta \in \Theta \} \) with metric \( d(P_{\theta_1}, P_{\theta_2}) = \| \theta_1 - \theta_2 \| \), estimation of \( \theta_0 \) (such that \( P_0 = P_{\theta_0} \)) by \( \hat{\theta}_n \) is consistent if \( \hat{\theta}_n \xrightarrow{P_0} \theta_0 \). In chapter 7, we generalize this to non-metric topologies in the obvious way.

An estimator that is consistent in metric setting may be analysed further by appraisal of its rate of convergence and limit distribution. Let us define the rate of convergence first.

**Definition 6.2.** Let \( \hat{P}_n \) be a sequence of estimators in a (single-observation) model \( \mathcal{P} \) with metric \( d \). Given \( P_0 \in \mathcal{P} \), any sequence \( r_n \) such that
6.2 Posterior concentration and model topology

\[ r_n^{-1} d(\hat{P}_n, P_0) = O_P(1), \]  

(6.1)

is an upper bound to the rate of convergence of the estimator sequence \( \hat{P}_n \) with respect to the metric \( d \), at \( P_0 \). Usually, the rate of convergence is the same for all \( P_0 \) (see, however, examples like 4.1).

The rate of convergence thus describes the scaling necessary to have metric differences between \( \hat{P}_n \) and \( P_0 \) that are distributed in a non-degenerate way, yet remain bounded in probability. This definition is very close to that of the rate associated with sequences of confidence sets, the rate with which balls around the point of convergence may be shrunk while still capturing the estimator with high probability. Note that Prohorov’s theorem guarantees weak convergence of a subsequence of the sequence \( r_n d(\hat{P}_n, P_0) \). Heightening the level of detail one step further still, we require that the sequence of estimators, when centred on its point of convergence and rescaled by the rate, converges weakly to a non-degenerate distribution over the (localised) model.

**Definition 6.3.** Let \( \hat{P}_n \) be a sequence of estimators in a (single-observation) model \( \mathcal{P} \) with metric \( d \). Given \( P_0 \in \mathcal{P} \) and rate sequence \( r_n \), we say that \( \hat{P}_n \) has limit distribution \( L_{P_0} \) at \( P_0 \), if:

\[ r_n^{-1} (\hat{P}_n - P_0) \overset{P_0}{\longrightarrow} L_{P_0}, \]

(6.2)

where \( L_{P_0} \) is a non-degenerate limit distribution.

This definition is rather abstract and leaves questions regarding the particulars of localisation of the model around \( P_0 \). In parametrizing spaces such ambiguities disappear: specialising again to the parametric case, we say that \( \hat{\theta}_n \) converges to \( \theta_0 \) at rate \( r_n^{-1} \) with non-degenerate limit distribution \( L_{\theta_0} \) on \( \mathbb{R}^d \) if:

\[ r_n^{-1} (\hat{\theta}_n - \theta_0) \overset{\theta_0}{\longrightarrow} L_{\theta_0}. \]

(6.3)

We have seen quite a few examples of such limiting behaviour in chapters 4 and 5.

**6.2 Posterior concentration and model topology**

Consistency, certainly incontestable as an asymptotic criterion from the frequentist point of view, is not free of controversy in Bayesian statistics. Specifically, the subjectivist Bayesian point of view does not attach value to any special point of convergence \( P_0 \) because no ‘underlying’ or ‘true’ distribution for the sample \( X_1, X_2, \ldots \) is assumed within the subjectivist paradigm. The notion of ‘merging’ is perhaps closer to the subjectivist’s philosophy: given two different priors \( \Pi_1 \) and \( \Pi_2 \) on a model \( \Theta \), merging is said to occur if the total-variation distance between the posterior predictive distributions goes to zero (see Blackwell and Dubins (1962) [37] and, for an overview, Ghosh and Ramamoorthi (2003) [103]). Relations between
merging and posterior consistency as defined below are discussed in Diaconis and Freedman (1986) [66].

### 6.2.1 Posterior concentration

We start by defining consistency, Bayesian style, generalising to sequentially observed (possibly non-i.i.d.) data, non-dominated models and priors or parameter spaces that may depend on the sample size, according to remark A.1.

**Definition 6.4.** The posteriors \( \Pi(\cdot | X^n) \) are consistent at \( \theta \in \Theta \) if for every neighbourhood \( U \) of \( \theta \),

\[
\Pi(U | X^n) \xrightarrow{P_{\theta_0,n}} 1. \tag{6.4}
\]

The posteriors are said to be consistent if this holds for all \( \theta \in \Theta \). A weaker form of posterior convergence is Bayesian consistency, when (6.4) holds for \( \Pi \)-almost-all \( \theta \in \Theta \). We say that the posterior is almost-surely consistent if convergence occurs almost-surely with respect to some coupling for the sequence \( (P_{\theta_0,n}) \).

For example, in the common case of a metric model \((\mathcal{P}, d)\) for i.i.d. data \( X_1, \ldots, X_n \sim P^n_0 \), consistency is equivalent to the condition that for every \( \varepsilon > 0 \):

\[
\Pi(d(P, P_0) \geq \varepsilon | X_1, X_2, \ldots, X_n) \xrightarrow{P_{\theta_0,n}} 0, \tag{6.5}
\]

since the above display is the complement of an open ball and every open neighbourhood of \( P_0 \) contains an open ball centred on \( P_0 \).

**Proposition 6.1.** Assume that \( \Theta \) is a Hausdorff, completely regular space. The posterior is consistent at \( \theta_0 \in \Theta \), if and only if,

\[
\int f(\theta) d\Pi(\theta | X^n) \xrightarrow{P_{\theta_0,n}} f(\theta_0), \tag{6.6}
\]

for every bounded, continuous \( f : \Theta \to \mathbb{R} \).

**Proof.** Assume (6.4). Let \( f : \Theta \to \mathbb{R} \) be bounded and continuous (with \( M > 0 \) such that \( |f| \leq M \)). Let \( \eta > 0 \) be given and let \( U \subset \Theta \) be a neighbourhood of \( \theta_0 \) such that \( |f(\theta) - f(\theta_0)| < \eta \) for all \( \theta \in U \). Integrate \( f \) with respect to the (assumed to be regular, i.e. \( P_{\theta_0,n} \)-almost-surely well-defined) posterior and to \( \delta_{\theta_0} \):

\[
\left| \int f(\theta) d\Pi(\theta | X^n) - f(\theta_0) \right|
\leq \int_{\Theta \setminus U} |f(\theta) - f(\theta_0)| d\Pi(\theta | X^n) + \int_U |f(\theta) - f(\theta_0)| d\Pi(\theta | X^n)
\leq 2M \Pi(\Theta \setminus U | X^n) + \sup_{\theta \in U} |f(\theta) - f(\theta_0)| \Pi(U | X^n) \leq \eta + o_{P_{\theta_0,n}}(1),
\]
as \( n \to \infty \), so that (6.6) holds. Conversely, assume (6.6). Let \( U \) be an open neighbourhood of \( \theta_0 \). Because \( \Theta \) is completely regular, there exists a continuous \( f : \Theta \to [0,1] \) such that \( f = 1 \) at \( \{ \theta_0 \} \) and \( f = 0 \) on \( \Theta \setminus U \). Then,

\[
\Pi(U|X^n) \geq \int f(\theta) d\Pi(\theta|X^n) \xrightarrow{P_{\theta_0^n}} \int f(\theta) d\delta_{\theta_0}(P) = 1.
\]

Consequently, (6.4) holds.

Metrisable spaces are uniform spaces and a space is uniform iff it is completely regular. So the above implies the following corollary immediately, in the common case of a metric model \((\mathcal{P}, d)\) for i.i.d. data.

**Corollary 6.1.** On metric models \((\mathcal{P}, d)\), (6.4), (6.5) and (6.6) are equivalent.

As becomes clear in chapter 8 the most convenient choice here is not the most natural: Prokhorov’s weak or the total-variational topologies are metric and attractive intuitively, but the natural model topology in i.i.d. context (see the Le Cam-Schwartz theorem, thm. 8.1) is a uniform topology \( \mathcal{U}_0 \) stronger than Prokhorov’s and weaker than total-variation. Without further conditions on the model \( \mathcal{P} \), the topology \( \mathcal{U}_0 \) is non-metrizable. (Recall that regular spaces with a countable basis are metrizable by Urysohn’s metrization theorem). We come back to these issues in chapter 8.

### 6.2.2 Consistency of Bayesian point-estimators

Point-estimators derived from a consistent Bayesian procedure are consistent themselves under some mild conditions. We reiterate that the notion of a point-estimator is not an entirely natural extension to the Bayesian framework: for example, if the model is non-convex (and hardly any model is), the posterior predictive distribution of definition 2.4 lies outside the model. Similarly, perfectly well-defined posteriors may lead to ill-defined point-estimators due to integrability issues or non-existence of maximisers, which become more severe as the model becomes more complicated.

Here, we endow a single-observation model \( \mathcal{P} \) (again, with data \( X_1, X_2, \ldots \) that are distributed i.i.d.-\( P_0 \)) with the total-variational topology and corresponding Borel \( \sigma \)-algebra.

**Theorem 6.1.** Assume that the Borel prior \( \Pi \) and underlying distribution \( P_0 \in \mathcal{P} \) are such that the sequence of posteriors is consistent (\( P_0 \)-almost-surely). Then the posterior predictive distributions \( \hat{P}_n \) are (\( P_0 \)-almost-surely) consistent point-estimators for \( P_0 \) with respect to total-variation.

**Proof.** Note that the domain of definition of the map \( P \mapsto \|P - P_0\| \) extends to the convex hull \( \text{co}(\mathcal{P}) \) of \( \mathcal{P} \) (in the collection of all probability distributions on the sample space). Since \( P \mapsto \|P - P_0\| \) is convex by virtue of the triangle inequality, Jensen’s inequality (see, e.g. theorem 10.2.6 in Dudley (1989) [76]) says that the posterior mean \( \hat{P}_n \) satisfies:
\[ \| \hat{P}_n - P_0 \| = \left\| \int_{\mathcal{P}} P \, d\Pi(P|X_1,\ldots,X_n) - P_0 \right\| \leq \int_{\mathcal{P}} \| P - P_0 \| \, d\Pi(P|X_1,\ldots,X_n). \]

Since the posteriors \( \Pi(\cdot|X_1,\ldots,X_n) P_0 \) converge weakly to \( P_0 \) \( (P_0\text{-almost-surely}) \) and the map \( P \mapsto \| P - P_0 \| \) is bounded and continuous in the total-variational topology, we conclude that the r.h.s. in the above display converges to the expectation of \( \| P - P_0 \| \) under the limit law \( \delta_{P_0} \) \( (P_0\text{-almost-surely}) \), which equals zero. Hence \( \hat{P}_n \) converges to \( P_0 \) in total variation \( (P_0\text{-almost-surely}) \).

More generally, given an arbitrary convex metric \( d \) on the model \( \mathcal{P} \), theorem 6.1 can be proved if the metric \( d \) is convex and bounded on \( \mathcal{P} \). Similar arguments demonstrate consistency for other classes of point estimators derived from a consistent sequence of posterior distributions, for example the formal Bayes estimators of subsection 2.2.2.

### 6.3 Bayesian consistency and Doob’s theorem

In this section, we concentrate on a sufficient condition for Bayesian consistency, a form of posterior consistency that holds for all \( P_0 \in \mathcal{P} \) except (perhaps) in a model subset that is a null-set of the prior.

Perhaps the most famous consistency theorem in Bayesian statistics is that given by Doob as early as 1949 [75].

**Theorem 6.2. (Doob (1949))**

For all \( n \geq 1 \), let \( (X_1,X_2,\ldots,X_n) \in \mathcal{X}^n \) be i.i.d. \(-\) \( P_0 \), where \( P_0 \) lies in a single-observation model \( \mathcal{P} \). Suppose \( \mathcal{X} \) and \( \mathcal{P} \) are Polish spaces and that \( P \mapsto P(A) \) is Borel measurable for every Borel set \( A \subset \mathcal{X} \). Then for any Borel prior \( \Pi \) the posterior is consistent at \( P \), \( P\text{-almost-surely, for } \Pi\text{-almost-all } P. \)

**Proof.** The proof of this theorem is an application of Doob’s martingale convergence theorem. We prove this theorem in chapter 8.

The measurability condition is relatively minor, a lower bound on the strength of the model topology: the associated Borel \( \sigma \)-algebra must have enough detail to talk about model-probabilities \( P(A) \) in a way that is compatible with the measure theory on which the posterior is based. So let us consider the only real condition of theorem 6.2, namely that \( (\mathcal{P},d) \) is Polish \( (e.g. \text{, separable and complete with respect to } d) \) a little closer: although many statistical models are not complete with respect to their (natural) metrics, we may replace \( \mathcal{P} \) here by its \( d \)-completion \( \hat{\mathcal{P}} \) without loss of generality, as long as we redefine the prior \( \hat{\Pi} \) on the Borel \( \sigma \)-algebra of the completion as \( \hat{\Pi}(B) = \Pi(B \cap \mathcal{P}) \), with the difference \( \mathcal{P} \setminus \mathcal{P} \) a null-set of \( \hat{\Pi} \). Separability is a different matter. To relate to parametrizing spaces immediately, note that function spaces like \( L^1(\mu) \) and \( L^\infty(\mu) \) relative to the Lebesgue measure on \( \mathbb{R} \), for example, are closely related but the former is separable while the latter is not [77]. Similarly, smoothness classes display diversity: while Sobolev spaces are separable, the closely related Hölder spaces are not [218]. With regard to Hellinger
and total-variational metrics on spaces of single-observation distributions, there is a very concrete equivalence: according to proposition 6.3 (e.g. for real-valued single observations \(X\), \(\mathcal{P}\) is Hellinger separable iff \(\mathcal{P}\) is dominated). The above delineates the realm of applicability of Doob’s theorem more accurately than the mere characterization Polish: in practical terms, we require a metric \(d\) on \(\mathcal{P}\) that is (i.) strong enough to guarantee measurability of \(P \rightarrow P(A)\), while not so strong as to ruin separability. The most natural formulation is in terms of densities with the Hellinger metric or total-variational metric (e.g. for real-valued data).

**Corollary 6.2.** For all \(n \geq 1\), let \((X_1, X_2, \ldots, X_n) \in \mathbb{R}^n\) be i.i.d. \(-\) \(P_0\), where \(P_0\) lies in a dominated single-observation model \(\mathcal{P}\) endowed with the Hellinger or total-variational metric. Then for any Borel prior \(\Pi\), the posterior is consistent at \(P\), \(P\)-almost-surely, for \(\Pi\)-almost-all \(P\).

**Proof.** (For more on Hellinger metric and Borel \(\sigma\)-algebra, see Breiman et al. [47], Le Cam (1986) [162] and Strasser [208].) All functions \(P \rightarrow P(A)\) are continuous with respect to the total-variational metric (which is topologically equivalent to the Hellinger metric), so they are measurable with respect to the Borel \(\sigma\)-algebra for the Hellinger (or total-variational) topology on \(\mathcal{P}\). The corollary is then an application of theorem 6.2 to the Polish (Hellinger or total-variational) completion \(\hat{\mathcal{P}}\) of the model \(\mathcal{P}\), with prior \(\hat{\Pi}\).

For (most) Bayesians Doob’s theorem is more than enough: c.f. the last remarks before example 2.5, the Bayesian views the model as derivative of the prior and thus only really defined up to prior null-sets. To illustrate this point from the parametric perspective: for parametric models \(\Theta \rightarrow \mathcal{P} : \theta \mapsto P_\theta\) with open \(\Theta \subset \mathbb{R}^d\) and a prior that dominates the restriction of the Lebesgue measure to \(\Theta\), the above theorem leaves room for inconsistency only on sets of Lebesgue measure zero (which can be called “intuitively small”, with certain reason). A popular way of stating this, is that consistency theorems like the above show that “the data overrides prior beliefs asymptotically”.

However, this note of optimism relies entirely on intuition based on Lebesgue measure. Consequently, there is absolutely no implied analogous promises outside of any parametric context. Indeed, Doob’s theorem becomes highly problematic in non-parametric models: the theorem stays true exactly as stated, it simply means something else than intuition suggests. Strictly speaking, only frequentists recognize consistency problems: Doob’s proof says nothing about specific points in the model, i.e. given a particular \(P_0 \in \mathcal{P}\) underlying the sample, Doob’s theorem does not give conditions that can be checked to see whether the Bayesian procedure will be consistent at this particular \(P_0\); it is always possible that \(P_0\) belongs to the null-set for which inconsistency occurs. That such null-sets may be large, is clear from example 2.5 and that, indeed, this may lead to grave problems in non-parametric situations, becomes apparent when we consider the counterexamples given by Freedman (1963,1965) [92, 93] and Diaconis and Freedman (1986) [66, 67]. Non-parametric examples of inconsistency in Bayesian regression can be found in Cox (1993) [57] and Diaconis and Freedman (1998) [69]. Basically what is shown is that the null-
6 Non-parametric Bayesian statistics

set on which inconsistency occurs in Doob’s theorem can be rather large in non-parametric situations. Some authors are tempted to present the above as definitive proof of the fact that Bayesian statistics are useless in non-parametric estimation problems. More precise would be the statement that not every choice of prior is suitable. We come back to Freedman’s counterexamples in subsection 6.7.1.

6.4 Dirichlet priors and posteriors

Next we consider the Dirichlet process prior, a probability measure on the full non-parametric model for a large class of Borel spaces \((\mathcal{X}, \mathcal{B})\). Here, we assume that \(\mathcal{X} = \mathbb{R}\), that \(\mathcal{B}\) is the Borel \(\sigma\)-algebra on \(\mathbb{R}\) and denote the collection of all probability measures on \((\mathbb{R}, \mathcal{B})\) by \(M(\mathbb{R}, \mathcal{B})\). We consider the collection \(\{P(A) : A \in \mathcal{B}\}\) as a collection of random quantities and wonder if they can be realised as a stochastic process, that is, whether there exists a probability space \((\Omega, \mathcal{F}, \Pi)\) with random variables \(P(A) : \Omega \to [0, 1]\) for each \(A \in \mathcal{B}\) (sometimes called a coupling for the \(P(A)’s)\).

6.4.1 Existence of the Dirichlet process

To construct \((\Omega, \mathcal{F}, \Pi)\), the Kolmogorov existence theorem (see theorem B.8) formulates straightforward consistency conditions on the distributions for finite subsets of \(\{P(A) : A \in \mathcal{B}\}\). Dirichlet distributions of the form (3.22) for finite partitions satisfy the consistency conditions, and Kolmogorov’s theorem asserts existence of a prior \(\Pi\) on sample-paths \(P : \mathcal{B} \to [0, 1]\). However, the construction implies only finite additivity of \(P\), not \(\sigma\)-additivity, so a proof that \(P\) is a measure with \(\Pi\)-probability one is required before we conclude that \(\Pi\) is a probability measure on \(M(\mathbb{R}, \mathcal{B})\), the Dirichlet process prior.

**Theorem 6.3. (Existence of the Dirichlet process)**

Given a finite measure \(\alpha\) on \((\mathbb{R}, \mathcal{B})\), there exists a probability measure \(D_\alpha\) on \(M(\mathbb{R}, \mathcal{B})\) (called the Dirichlet process prior with parameter \(\alpha\)) such that for \(P \sim D_\alpha\) and every \(\mathcal{B}\)-measurable partition \((B_1, \ldots, B_k)\) of \(\mathbb{R}\),

\[
(P(B_1), \ldots, P(B_k)) \sim D_{(\alpha(B_1), \ldots, \alpha(B_k))}.
\]

**Proof.** Let \(k \geq 1\) and \(A_1, \ldots, A_k \in \mathcal{B}\) be given. Through the indicators \(1_{A_i}\) for these sets, we define \(2^k\) new sets,

\[
1_{B_{\nu_1 \ldots \nu_k}} = \prod_{i=1}^{k} 1_{A_i}^{\nu_i} (1 - 1_{A_i})^{1 - \nu_i},
\]
where $v_1, \ldots, v_k \in \{0,1\}$. Then the collection \( \{B_{v_1, \ldots, v_k} : v_i \in \{0,1\}, 1 \leq i \leq k\} \) forms a partition of \( \mathbb{R} \). For the \( P\)-probabilities corresponding to this partition, we assume finite-dimensional marginals
\[
(P(B_{v_1, \ldots, v_k}) : v_i \in \{0,1\}, 1 \leq i \leq k) \sim \Pi_{B_{v_1, \ldots, v_k}, v_i \in \{0,1\}, 1 \leq i \leq k},
\]
The distribution of the vector \( (P(A_1), \ldots, P(A_k)) \) then follows from:
\[
P(A_i) = \sum_{\{i: v_i = 1\}} P(B_{v_1, \ldots, v_k}),
\]
and (3.21), for all \( 1 \leq i \leq k \). This defines marginal distributions for all finite subsets of \( \mathcal{B} \), as needed in theorem B.8. To define the underlying probability space \( (\Omega, \mathcal{F}, \Pi) \) we now impose two conditions.

(F1) With \( \Pi \)-probability one, the empty set has \( P \)-measure zero:
\[
\Pi(P(\emptyset) = 0) = 1.
\]

(F2) Let \( k, k' \geq 1 \) be given. If \( (B_1, \ldots, B_k) \) is a partition and \( (B'_1, \ldots, B'_{k'}) \) a refinement thereof, with
\[
B_1 = \bigcup_{i=1}^{r_1} B'_i, \ldots, B_k = \bigcup_{i=r_{k-1}+1}^{k'} B'_i,
\]
(for certain \( r_1 < \ldots < r_{k-1} \)), then we have the following equality in distribution:
\[
\mathcal{L} \left( \sum_{i=1}^{r_1} P(B'_i), \ldots, \sum_{i=r_{k-1}+1}^{k'} P(B'_i) \right) = \mathcal{L} \left( P(B_1), \ldots, P(B_k) \right). \tag{6.8}
\]

Condition (F1) ensures that if \( (A_1, \ldots, A_k) \) is itself a partition of \( \mathbb{R} \), the above construction does not lead to a contradiction. Condition (F2) ensures finite additivity of \( P \) with prior probability one, i.e., for any \( A, B, C \in \mathcal{B} \) such that \( A \cap B = \emptyset \) and \( A \cup B = C \),
\[
\Pi \left( P(A) + P(B) = P(C) \right) = 1. \tag{6.9}
\]

Ferguson (1973,1974) [89, 90] has shown that conditions (F1) and (F2) imply that Kolmogorov’s consistency conditions (K1) and (K2) (see section B.4) are satisfied. As we have seen in the first part of this section, if we impose the Dirichlet distribution:
\[
(P(B_{v_1, \ldots, v_k}) : v_i \in \{0,1\}, 1 \leq i \leq k) \sim D_\alpha(B_{v_1, \ldots, v_k}, v_i \in \{0,1\}, 1 \leq i \leq k). \tag{6.10}
\]
and \( \alpha \) is a measure on \( \mathcal{B} \), condition (F2) is satisfied. Combining all of this, we conclude that there exists a probability space \( (\Omega, \mathcal{F}, \Pi) \) on which the stochastic process \( \{P(A) : A \in \mathcal{B}\} \) can be represented with finite dimensional marginals c.f. (6.10). Lemma 6.1 shows that \( \Pi(P \in M(\mathbb{R}, \mathcal{B})) = 1 \), completing the proof.
The last line in the above proof may require some further explanation: \(P\) is merely the sample-path of our stochastic process. The notation \(P(A)\) suggests that \(P\) is a probability measure, but all we have shown up to that point, is that (F1) and (F2) imply that \(P\) is a finitely additive set-function such that:

\[
\Pi(\{P(B) \in [0,1] : B \in \mathcal{B}\}) = 1.
\]

What remains to be demonstrated is \(\Pi\)-almost-sure \(\sigma\)-additivity of \(P\). We follow the historical proof [89, 90], which contains a mistake. So, although the following assertion is true, the proof contains a hard-to-spot mistake that we correct only in subsection 7.4.3. Finding the mistake is left to the reader as exercise 6.1.

**Lemma 6.1.** If \(\Pi\) is a Dirichlet process prior \(D_\alpha\),

\[
\Pi(\{P \in M(\mathcal{X}, \mathcal{B})\}) = 1.
\]

**Proof.** Let \((A_n)_{n \geq 1}\) be a sequence in \(\mathcal{B}\) that decreases to \(\emptyset\). Since \(\alpha\) is \(\sigma\)-additive, \(\alpha(A_n) \to \alpha(\emptyset) = 0\). Therefore, there exists a subsequence \((A_{n_j})_{j \geq 1}\) such that \(\sum_j \alpha(A_{n_j}) < \infty\). For fixed \(\varepsilon > 0\), using Markov’s inequality first,

\[
\sum_{j \geq 1} \Pi(P(A_{n_j}) > \varepsilon) \leq \frac{1}{\varepsilon} \int P(A_{n_j}) d\Pi(P) = \frac{1}{\varepsilon} \sum_{j \geq 1} \frac{\alpha(A_{n_j})}{\alpha(\mathbb{R})} < \infty,
\]

according to lemma 3.4. From the Borel-Cantelli lemma (see lemma B.2), we see that

\[
\Pi(\limsup_{j \to \infty} \{P(A_{n_j}) > \varepsilon\}) = \Pi\left(\bigcap_{j \geq 1} \bigcup_{j \geq J} \{P(A_{n_j}) > \varepsilon\}\right) = 0,
\]

which shows that \(\lim_j P(A_{n_j}) = 0\), \(\Pi\)-almost-surely. Since, by \(\Pi\)-almost-sure finite additivity of \(P\),

\[
\Pi(P(A_n) \geq P(A_{n+1}) \geq \ldots) = 1,
\]

we conclude that \(\lim_n P(A_n) = 0\), \(\Pi\)-almost-surely. By the continuity theorem for measures (see theorem B.1 and the proof in [138], theorem 3.2), \(P\) is \(\sigma\)-additive \(\Pi\)-almost-surely.

### 6.4.2 Support of Dirichlet priors

Once existence of the Dirichlet process prior on \(M(\mathbb{R}, \mathcal{B})\) is established, one wonders about the nature of the constructed prior. As it turns out, the Dirichlet process prior has some remarkable properties.

**Lemma 6.2.** (Dirichlet process prior support in Prokhorov’s weak topology)

Consider \(M(\mathbb{R}, \mathcal{B})\) with Prokhorov’s weak topology. Let \(\alpha\) be a finite measure on \((\mathbb{R}, \mathcal{B})\). The support of \(D_\alpha\) is given by,
supp $\tilde{\mathcal{J}}_1(D_\alpha) = \{ P \in M(\mathbb{R}, \mathcal{B}) : \text{supp}(P) \subset \text{supp}(\alpha) \}.$

But we can be more precise, as shown in the following lemma and its corollary: note that, if $P \ll \alpha$ then $\text{supp}(P) \subset \text{supp}(\alpha)$, but the opposite is not true.

**Lemma 6.3.** Let $\alpha$ be a finite measure on $(\mathbb{R}, \mathcal{B})$ and let $\{ P(A) : A \in \mathcal{B} \}$ be the associated Dirichlet process with distribution $D_\alpha$. Let $Q \in M(\mathbb{R}, \mathcal{B})$ be such that $Q \ll \alpha$. Then, for any $I \geq 1$ and $A_1, \ldots, A_I \in \mathcal{B}$ and $\varepsilon > 0$,

$$D_\alpha(P \in M(\mathbb{R}, \mathcal{B}) : |P(A_i) - Q(A_i)| < \varepsilon, 1 \leq i \leq m) > 0.$$  \hspace{1cm} (6.11)

**Proof.** The proof of this lemma can be found in [103], theorem 3.2.4.

For the topology $\mathcal{T}_1$ used below to formulate the corollary, we refer to appendix section ??.

**Corollary 6.3.** ($\mathcal{T}_1$-support of the Dirichlet process prior)

Let $\alpha$ be a finite measure on $(\mathbb{R}, \mathcal{B})$ and let $\{ P(A) : A \in \mathcal{B} \}$ be the associated Dirichlet process with distribution $D_\alpha$. Then, for the topology $\mathcal{T}_1$ on $M(\mathbb{R}, \mathcal{B})$,

$$\text{supp}_{\mathcal{T}_1}(D_\alpha) = \{ P \in M(\mathbb{R}, \mathcal{B}) : P \ll \alpha \}.$$

**Proof.** If $Q$ is not dominated by $\alpha$, then there exists an $A \in \mathcal{B}$ such that $Q(A) > 0 = \alpha(A)$. Consequently, for small enough $\varepsilon' > 0$, the $\mathcal{T}_1$-open neighbourhood $U = \{ P \in M(\mathbb{R}, \mathcal{B}) : |P(A) - Q(A)| < \varepsilon' \}$ does not meet $\{ P \in M(\mathbb{R}, \mathcal{B}) : P \ll \alpha \}$, so $\{ P \in M(\mathbb{R}, \mathcal{B}) : P \ll \alpha \}$ is closed. In addition, c.f. definition 3.9, $\Pi(P(A) > 0) = 0$, so $U$ receives $D_\alpha$-mass zero, and we conclude that $Q$ does not lie in the support of $D_\alpha$. Conversely, let $Q \in M(\mathbb{R}, \mathcal{B})$, $\varepsilon > 0$, $k \geq 1$ and $\mathcal{B}$-measurable $\phi : \mathbb{R} \rightarrow [0,1]$ $(1 \leq l \leq k)$ be given. There exists an $n \geq 1$ such that for every $l$, there is a $\mathcal{B}$-measurable partition $\{ A_{m,l} \}_{m=1}^n$ of $\mathbb{R}$ and constants $0 \leq f_{m,l} \leq 1$, $1 \leq m \leq n$, $1 \leq l \leq k$, such that the simple functions $f_l(x) = \sum_{m=1}^n f_{m,l} 1\{ x \in A_{m,l} \}$ approximate $\phi_l$ uniformly,

$$\sup_{x \in \mathbb{R}} |\phi_l(x) - f_l(x)| < \varepsilon/4.$$  

If we define the $A_1, \ldots, A_I$ of lemma 6.3 to be the $A_{m,l}$, for all $1 \leq m \leq n$, $1 \leq l \leq k$, then,

$$\{ P \in M(\mathbb{R}, \mathcal{B}) : |P(A_i) - Q(A_i)| < \varepsilon/m, 1 \leq i \leq I \} \subset \{ P \in M(\mathbb{R}, \mathcal{B}) : |P\phi - Q\phi| < \varepsilon, 1 \leq l \leq k \}.$$

We have shown that every basis element for $\mathcal{T}_1$ contains a set of the form in (6.11). For every point $Q'$ in the closure of $\{ P \in M(\mathbb{R}, \mathcal{B}) : P \ll \alpha \}$ and every open neighbourhood $U$ of $Q'$, there exists a point $Q \ll \alpha$ and a neighbourhood $V$ of $\mathcal{T}_1$ with $V \subset U$. According to lemma 6.3, this implies $Q' \in \text{supp}_{\mathcal{T}_1}(D_\alpha)$.

The following property reveals a most remarkable property of Dirichlet process priors: the subset $D(\mathbb{R}, \mathcal{B})$ of all discrete convex combinations of Dirac measures (see example B.3) receives prior mass equal to one.
Lemma 6.4. Let \( \alpha \) be a finite measure on \((\mathbb{R}, \mathcal{B})\) and let \( \{P(A) : A \in \mathcal{B}\} \) be the associated Dirichlet process with distribution \( D_\alpha \). Then,

\[
D_\alpha \left( P \in D(\mathbb{R}, \mathcal{B}) \right) = 1.
\]

Proof. The proof of this lemma can be found in [103], theorem 3.2.3.

The above phenomenon leads to problems with support or convergence in stronger topologies (like total variation or Hellinger topologies) and with regard to the Kullback-Leibler criteria that govern asymptotic concentration of posterior mass. Generalizing somewhat, we conclude from lemma 6.4 that the Dirichlet process prior is not suited for estimation of densities. Although clearly dense enough in \( M(\mathbb{R}, \mathcal{B}) \) in Prokhorov’s weak topology and in \( \mathcal{T}_1 \), the set \( D(\mathbb{R}, \mathcal{B}) \) may be rather sparse in stronger topologies (notwithstanding the fact that mixture models with a Dirichlet process prior for the mixing distribution can be (minimax) optimal for the estimation of mixture densities [101]).

### 6.4.3 Consistency of Dirichlet posteriors

The most important result of this section is the fact that the family of Dirichlet process priors on \( M(\mathbb{R}, \mathcal{B}) \) is a conjugate family for the full, non-parametric model on \((\mathbb{R}, \mathcal{B})\), as stated in the following theorem.

**Theorem 6.4.** Let \( X_1, X_2, \ldots \) be an i.i.d. sample of observations in \( \mathbb{R} \). Let \( \alpha \) be a finite measure on \((\mathbb{R}, \mathcal{B})\) with associated Dirichlet process prior \( \Pi = D_\alpha \). For any measurable \( C \subset M(\mathbb{R}, \mathcal{B}) \),

\[
\Pi \left( P \in C \mid X_1, \ldots, X_n \right) = D_\alpha \left( \sum_{i=1}^{n} \delta_{X_i} \mid C \right).
\]

Proof. Consider cylinders of the form,

\[
\{ P \in M(\mathbb{R}, \mathcal{B}) : (P(A_1), \ldots, P(A_k)) \in B \},
\]

where \( A_1, \ldots, A_k \in \mathcal{B} \) form a partition of \( \mathbb{R} \), and \( B \) lies in the \( k \)-fold product \( \sigma \)-algebra of the Borel \( \sigma \)-algebra on \([0,1]\). Sets of this form generate the \( \sigma \)-algebra that is the domain of the Dirichlet prior. Equality for the cylinder sets depends only on the marginal distribution of \( (P(A_1), \ldots, P(A_k)) \), and consequently, the posterior for \( (P(A_1), \ldots, P(A_k)) \) is given by theorem 3.2. The posterior for \( P \in M(\mathbb{R}, \mathcal{B}) \) is now fixed because equality of two measures on a generating ring implies equality on the whole \( \sigma \)-algebra. Conjugacy of the family of Dirichlet process distributions follows, and the base measure for the posterior is of the form \( \alpha' \) in (3.23).

**Example 6.1.** Let \( X_1, X_2, \ldots \) be an i.i.d. sample of observations in \( \mathbb{R} \). Let \( \alpha \) be a finite measure on \((\mathbb{R}, \mathcal{B})\) with associated Dirichlet process prior \( \Pi = D_\alpha \). Let \( B \in \mathcal{B} \) be given. The expectation \( P_\alpha(B) \) of \( P(B) \) under \( D_\alpha \) equals,
the measure of $B$ under $\alpha$ normalized to be a probability measure $P_\alpha$. The posterior predictive distribution (see definition 2.7), is then given by:

$$
\int P(B) d\Pi(P \mid X_1, \ldots, X_n) = \frac{\alpha(\mathbb{R})}{\alpha(\mathbb{R}) + n P_\alpha(B)} + n \frac{1}{\alpha(\mathbb{R}) + n} \sum_{i=1}^{n} \delta_{X_i}(B),
$$

almost-surely. With reference to decompositions (3.12) and (3.14), we see that the posterior predictive distribution $\hat{P}_n$ can be viewed as a convex combination of the empirical distribution $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ and a bias term in the form of the prior mean distribution $P_\alpha$.

$$
\hat{P}_n = \lambda_n P_\alpha + (1 - \lambda_n) P_n,
$$

almost-surely. As a result, we see that,

$$
\|\hat{P}_n - P_n\|_{TV} = \lambda_n \|P_\alpha - P_n\| \leq \lambda_n,
$$

almost-surely. Since $\lambda_n \to 0$ as $n \to \infty$, the difference between the sequence of posterior predictive distributions $(\hat{P}_n)_{n \geq 1}$ and the sequence of empirical measures $(P_n)_{n \geq 1}$ converges to zero in total variation as we let the samplesize grow to infinity. Note that the estimator $P_n$ for $P_\alpha$ based on i.i.d. data $X_n = (X_1, \ldots, X_n) \sim P^n$, is $\mathcal{F}_1$-consistent. (Namely, for every $\varepsilon > 0$ and measurable $f : \mathbb{R} \to [0, 1]$, $P_n f \to P f$, $P$-almost-surely, by the law of large numbers. As a consequence, for every $P$ and every $\mathcal{F}_1$-neighbourhood $U$ of $P$, $P_n \in U$ for large enough $n \geq 1$, $P$-almost-surely.) To shed some light on the estimator $P_n$ from a generalized ML perspective, Dvoretzky, Kiefer and Wolfowitz (1956) [78] have shown that the empirical distribution $P_n$ can be viewed as a non-parametric maximum-likelihood estimator (NPMLE), an MLE in non-dominated models, and discuss forms of consistency for $P_n$.

### 6.5 Schwartz’s frequentist posterior consistency theorem

Fortunately, a theorem exists that provides sufficient conditions for consistency at a specific point $P_0 \in \mathcal{P}$. As such, Schwartz’s theorem below is a frequentist consistency theorem for the posterior in non-parametric setting.

**Theorem 6.5.** *(Schwartz (1965))*

For all $n \geq 1$, let $(X_1, X_2, \ldots, X_n) \in \mathcal{X}^n$ be i.i.d. $- P_0$, where $P_0$ lies in a model $\mathcal{P}$. Let $U$ denote an open neighbourhood of $P_0$ in $\mathcal{P}$. If,

(i) there exist measurable $\phi_n : \mathcal{X}^n \to [0, 1]$, such that,


P_0^n \phi_n = o(1), \quad \sup_{Q \in U^c} Q^n(1 - \phi_n) = o(1), \quad (6.14)

(ii) and \( \Pi \) is a Kullback-Leibler prior, i.e. for all \( \delta > 0 \),

\[ \Pi \left( P \in \mathcal{P} : -P_0 \log \frac{dP}{dP_0} < \delta \right) > 0, \quad (6.15) \]

then \( \Pi(U|X^n) \xrightarrow{P_0\text{-a.s.}} 1. \)

The condition of domination in the above theorem is strictly speaking redundant: it is possible to give the entire proof in its present form, if we replace \( p/p_0 \) by the Radon-Nikodym derivative \( dP/dP_0 \) throughout (and we change the third equality in (6.18) into less-or-equal). Furthermore, it should be noted that the counterexamples of Diaconis and Freedman, for example, fail the lower bound for prior mass in Kullback-Leibler neighbourhoods of the true distribution in Schwartz's theorem (see Barron et al. (1999) [16]).

Comparing condition (6.14) with the assertion of Urysohn’s lemma or the definition of complete regularity, one notices conceptually similar roles for separating functions and test functions: the sequence of test functions in (6.14) “separates” the singleton \( \{P_0\} \) from the alternative, as a stochastic, uniform limit. The central condition of Schwartz’s theorem, requiring the existence of a suitable test sequence, is revisited in chapters 7 and 8. The work of Le Cam is full of references to this device, so it is difficult to indicate original references; an abridged list should at least contain, Le Cam (1973) [160], Birge (1983,1984) [32, 33], Le Cam (1986) [162] and, in the context of posterior convergence, Ghosal et al. (2000) [101]). Condition 6.15 induces a weakened form of contiguity (see appendix C.3) and is analysed in more generality in chapter 7.

### 6.5.1 Proof of Schwartz’s theorem

**Proof.** Let \( \varepsilon > 0 \) be given. Define \( V \) to be the complement of the open \( \varepsilon \)-ball of radius \( \varepsilon \) around \( P_0 \) in \( \mathcal{P} \):

\[ V = \{ P \in \mathcal{P} : d(P, P_0) \geq \varepsilon \}. \]

We start by splitting the \( n \)-th posterior measure of \( V \) with the test function \( \phi_n \) and taking the limes superior:

\[
\limsup_{n \to \infty} \Pi(V|X_1, \ldots, X_n) \\
\leq \limsup_{n \to \infty} \Pi(V|X_1, \ldots, X_n)(1 - \phi_n) + \limsup_{n \to \infty} \Pi(V|X_1, \ldots, X_n) \phi_n.
\]

(6.16)
For given $\eta > 0$ (to be fixed at a later stage) we consider the subset $K_\eta = \{ P \in \mathcal{P} : -P_0 \log(p/p_0) \leq \eta \}$. For every $P \in K_\eta$, the strong law of large numbers says that:

$$\left| P_n \log \frac{P}{P_0} - P_0 \log \frac{P}{P_0} \right| \to 0, \quad (P_0 - a.s.).$$

Hence for every $\alpha > \eta$ and all $P \in K_\eta$, there exists an $N \geq 1$ such that for all $n \geq N$, $\prod_{i=1}^{n} (p/p_0)(X_i) \geq e^{-\alpha n}$, $P_0^\alpha$-almost-surely. This can be used to lower-bound the denominator in the expression for the posterior $P_0^\alpha$-almost-surely as follows:

$$\liminf_{n \to \infty} e^{\alpha n} \int_{\mathcal{V}} \prod_{i=1}^{n} \frac{P}{P_0}(X_i) d\Pi(P) \geq \liminf_{n \to \infty} e^{\alpha n} \int_{K_\eta} \prod_{i=1}^{n} \frac{P}{P_0}(X_i) d\Pi(P) \geq \int_{K_\eta} \liminf_{n \to \infty} e^{\alpha n} \prod_{i=1}^{n} \frac{P}{P_0}(X_i) d\Pi(P) \geq \Pi(K_\eta),$$

where we use Fatou’s lemma to obtain the second inequality. Since by assumption, $\Pi(K_\eta) > 0$ we see that the first term on the r.h.s. of (6.16) can be estimated as follows:

$$\limsup_{n \to \infty} \Pi(V|X_1, \ldots, X_n) (1 - \phi_n)(X_1, \ldots, X_n)$$

$$= \limsup_{n \to \infty} \frac{\int_{V} \prod_{i=1}^{n} \frac{P}{P_0}(X_i)(1 - \phi_n)(X_1, \ldots, X_n) d\Pi(P)}{\int_{\mathcal{V}} \prod_{i=1}^{n} \frac{P}{P_0}(X_i) d\Pi(P)}$$

$$\leq \frac{\liminf_{n \to \infty} e^{\alpha n} \int_{\mathcal{V}} \prod_{i=1}^{n} (p/p_0)(X_i)(1 - \phi_n)(X_1, \ldots, X_n) d\Pi(P)}{\Pi(K_\eta) \limsup_{n \to \infty} f_n(X_1, \ldots, X_n)},$$

where we use the following, $P_0^\alpha$-almost-surely defined sequence of non-negative random variables $(f_n)_{n \geq 1}, f_n : \mathcal{X}^n \to \mathbb{R}$:

$$f_n(X_1, \ldots, X_n) = e^{\alpha n} \int_{\mathcal{V}} \prod_{i=1}^{n} \frac{P}{P_0}(X_i)(1 - \phi_n)(X_1, \ldots, X_n) d\Pi(P).$$

Fubini’s theorem and the fact that the test-sequence can be assumed to be uniformly exponential (see lemma 6.5) guarantee that there exists a constant $B > 0$ such that for large enough $n$,
\[ P^n_0 f_n = P^n_0 f_n = e^{\alpha \alpha} \int_V P^n_0 \left( \prod_{i=1}^n \frac{P}{P_0}(X_i)(1 - \phi_n)(X_1, \ldots, X_n) \right) d\Pi(P) \leq e^{\alpha \alpha} \int_V P^n(1 - \phi_n) d\Pi(P) \leq e^{-n(\beta - \alpha)}. \]  

(6.18)

We choose \( \eta \) strictly below \( \beta \) and can then choose \( \alpha \) such that \( \eta < \alpha < 1/2(\beta + \eta) \). Markov’s inequality can be used to show that:

\[ P^n_0 \left( f_n > e^{-n/2(\beta - \eta)} \right) \leq e^{\alpha - n} \left( \beta + \eta \right) \]

Hence the series \( \sum_{n=1}^\infty P^n_0 \left( f_n > \exp -n/2(\beta - \eta) \right) \) converges and the first Borel-Cantelli lemma then leads to the conclusion that:

\[ 0 = P^n_0 \left( \bigcap_{N=1}^\infty \bigcup_{n \geq N} \left\{ f_n > e^{-n/2(\beta - \eta)} \right\} \right) \geq P^n_0 \left( \limsup_{n \to \infty} \left( f_n - e^{-n/2(\beta - \eta)} \right) > 0 \right) \]

Since \( f_n \geq 0 \), we see that \( f_n \to 0 \) (\( P_0 \)-a.s.), which we substitute in (6.17).

We estimate the last term on the r.h.s. of (6.16) with an argument similar to that used above for the functions \( f_n \). Note that \( P^n_0 \Pi(V|X_1, \ldots, X_n) \phi_n \leq P^n_0 \phi_n \leq e^{-nC} \) for some positive constant \( C \), according to lemma 6.5. Markov’s inequality and the first Borel-Cantelli lemma suffice to show that:

\[ \phi_n \Pi(V|X_1, \ldots, X_n) \xrightarrow{P_0 \text{-a.s.}} 0. \]  

(6.19)

Combination of (6.17) and (6.19) proves that (6.16) equals zero.

**Lemma 6.5.** Suppose that for given \( \varepsilon > 0 \) there exists a sequence of tests \( (\phi_n)_{n \geq 1} \) such that:

\[ P^n_0 \phi_n \to 0, \quad \sup_{P \in V_\varepsilon} P^n(1 - \phi_n) \to 0 \]

where \( V_\varepsilon = \{ P \in \mathcal{P} : d(P, P_0) \geq \varepsilon \} \). Then there exists a sequence of tests \( (\omega_n)_{n \geq 1} \) and positive constants \( C, D \) such that:

\[ P^n_0 \omega_n \leq e^{-nC}, \quad \sup_{P \in V_\varepsilon} P^n(1 - \omega_n) \leq e^{-nD} \]  

(6.20)

**6.6 Posterior convergence at a rate**

Recalling the formulation of posterior consistency given in (6.5), we define the rate of convergence for a consistent sequence of posteriors as the maximal speed with which we can let the balls \( d(P, P_0) < \varepsilon \) shrink to radius zero, while still capturing a posterior mass that converges to one in the limit \( n \to \infty \). We formalise this as follows.

**Definition 6.5.** Let \( \mathcal{P} \) be a model with metric \( d \) and Borel prior \( \Pi \). Assume that \( X_1, X_2, \ldots \) are i.i.d.-\( P_0 \), for some \( P_0 \in \mathcal{P} \). Let the sequence \( \varepsilon_n \) be such that \( \varepsilon_n >
6.6 Posterior convergence at a rate

0 and $\varepsilon_n \downarrow 0$. We say that the sequence of posterior measures $\Pi(\cdot|X_1, X_2, \ldots, X_n)$ converges (at least) at rate $\varepsilon_n$ at $P_0$, if for all sequences $M_n \to \infty$:

$$
\Pi\left( d(P, P_0) \geq M_n \varepsilon_n \,|\, X_1, X_2, \ldots, X_n \right) \overset{p_0}{\to} 0,
$$

(6.21)

To demonstrate how this definition relates to the rate of convergence for derived point-estimators like the posterior predictive distribution, assume that the \(\sigma\)-algebra on the model contains the Borel \(\sigma\)-algebra corresponding to the metric topology generated by \(d\). We also assume that the sequence of posteriors satisfies (6.21). With the sequence \(\varepsilon_n\), we define moreover the point estimators $\tilde{P}_n$ as (near-)maximisers in the model of the maps:

$$
P \mapsto \Pi\left( B(P, \varepsilon_n) \,|\, X_1, \ldots, X_n \right),
$$

where $B(P, \varepsilon) \subset \mathcal{P}$ is the \(d\)-ball of radius \(\varepsilon\) around \(P\) in the model.

**Proposition 6.2.** For every sequence $M_n \to \infty$, the estimator sequence $\tilde{P}_n$ satisfies

$$
P_0\left( d(\tilde{P}_n, P_0) \leq 2 M_n \varepsilon_n \right) \to 1
$$

(6.22)

As a result, $\varepsilon_n^{-1}$ is a lower bound for the rate at which $\tilde{P}_n$ converges to $P_0$ with respect to $d$.

**Proof.** Let $\tilde{P}_n$ like above be given. By definition of a near-maximiser:

$$
\Pi\left( B(\tilde{P}_n, M_n \varepsilon_n) \,|\, X_1, \ldots, X_n \right) \geq \sup_{P \in \mathcal{P}} \Pi\left( B(P, M_n \varepsilon_n) \,|\, X_1, \ldots, X_n \right) - o_{P_0}(1)
$$

$$
\geq \Pi\left( B(P_0, M_n \varepsilon_n) \,|\, X_1, \ldots, X_n \right) - o_{P_0}(1).
$$

Because the first term on the r.h.s. of the above display converges to one (according to (6.21)) and the second to zero in $P_0$-probability, the l.h.s. converges to one in $P_0$-probability. Since $B(\tilde{P}_n, M_n \varepsilon_n) \cap B(P_0, M_n \varepsilon_n) = \emptyset$ if $d(\tilde{P}_n, P_0) > 2 M_n \varepsilon_n$, the fact that the total posterior mass of the model does not exceed one guarantees that $d(\tilde{P}_n, P_0) \leq 2 M_n \varepsilon_n$ with $P_0$-probability growing to one as $n \to \infty$, demonstrating that $\varepsilon_n^{-1}$ is a lower bound to the rate.

(Proposition 6.2 was noted in Ghosal et al. (2000) [101] but not its immediate implication, that balls of radii $3 M_n \varepsilon_n$ centred on $\tilde{P}_n$ are confidence balls. This observation is explored further in subsection 7.7.1.) A proof that does not differ in an essential way from the above can be given for the centre point of the $d$-ball of minimal radius containing posterior mass $p > 1/2$.

For the posterior predictive distribution we can prove a similar result if we specify the convergence of the posterior measure of complements of balls a little further. Consider a model $\mathcal{P}$ with Hellinger metric $H$ and the corresponding Borel $\sigma$-algebra. By the convexity of $P \mapsto H^2(P, P_0)$, the fact that this map can be extended to the convex hull of $\mathcal{P}$ and Jensen’s inequality (see the proof of theorem 6.1):
\[ H^2(\hat{P}_n, P_0) = H^2\left( \int P d\Pi(P | X_1, \ldots, X_n), P_0 \right) \leq \int_{\{H(P, P_0) > M_n \varepsilon_n\}} H^2(P, P_0) d\Pi(P | X_1, \ldots, X_n) \]
\[ + \int_{\{H(P, P_0) \leq M_n \varepsilon_n\}} H^2(P, P_0) d\Pi(P | X_1, \ldots, X_n) \]
\[ \leq 2\Pi(H(P, P_0) > M_n \varepsilon_n | X_1, \ldots, X_n) \]
\[ + M_n^2 \varepsilon_n^2 \Pi(H(P, P_0) \leq M_n \varepsilon_n | X_1, X_n) \]

If \( \Pi(H^2(P, P_0) > a_n^2 | X_1, \ldots, X_n) = o_{P_0}(a_n^2) \) for all (deterministic) sequences \( a_n \downarrow 0 \), the above display implies that \( H(\hat{P}_n, P_0) \) is bounded in probability by a multiple of \( M_n \varepsilon_n \) (c.f. (6.22)) for all sequences \( M_n \to \infty \), leading to the same conclusion as that of proposition 6.2.

The possibility to construct point estimator sequences from posterior distributions converging at the same rate (e.g. \( \hat{P}_n \) above), implies that limitations on the rate of convergence (arising in particular in non-parametric estimation problems, see (6.31) below, for example) derived for point estimation, apply unabated to Bayesian rates. This argument applies to other asymptotic performance criteria as well.

### 6.6.1 The Ghosal-Ghosh-van der Vaart theorem

With regard to sufficient conditions for the defining property (6.21) of posterior rates of convergence, we note that the number of references on this subject is relatively small when compared to the literature concerning Bayesian consistency. We note first of all, Le Cam (1973) [160] and Ibragimov and Has’minskii (1981) [124], who prove that under regularity conditions, posteriors on parametric models achieve \( \sqrt{n} \)-rate of convergence. Le Cam (1986) [162] considers rates of convergence of formal Bayes estimators, based on unpublished work using what is now known as Le Cam’s inequality (see Le Cam (197X) [161] and subsection 7.1.3). The two main references dealing with Bayesian rates of convergence in non-parametric models are Ghosal, Ghosh and Van der Vaart (2000) [101] and Shen and Wasserman (2001) [204]. We postpone a more complete discussion of the literature to section 7.5.

Again, we assume a (non-parametric) model \( \mathcal{P} \) with metric \( d \) and prior \( \Pi \). To formulate the main theorem of this subsection we define, for every \( \varepsilon > 0 \),

\[ B(\varepsilon; P_0) = \left\{ P \in \mathcal{P} : -P_0 \log \frac{P}{P_0} \leq \varepsilon^2, P_0 \left( \log \frac{P}{P_0} \right)^2 \leq \varepsilon^2 \right\}. \]  

(6.23)

This allows us to formulate a more specific version of Schwartz’s Kullback-Leibler condition (6.15), in the form of (6.24).
6.6 Posterior convergence at a rate

**Theorem 6.6.** Suppose that for a sequence $\varepsilon_n$ with $\varepsilon_n > 0$, $\varepsilon_n \downarrow 0$ and $n\varepsilon_n^2 \to \infty$, some $P_0 \in \mathcal{P}$, and a sequence $(\Pi_n)$ of priors, the following two conditions hold:

(i) There exists a constant $C > 0$ such that:

$$\Pi_n(B(\varepsilon_n; P_0)) \geq e^{-nC\varepsilon_n^2}.$$  \hspace{1cm} (6.24)

(ii) There exists a sequence $\phi_n$ of test-functions $\phi_n$ and a constant $L > 0$ such that:

$$P_0^\phi \phi_n \to 0, \quad \sup_{P: d(P, P_0) \geq \varepsilon_n} P^n(1 - \phi_n) \leq e^{-nL\varepsilon_n^2}.$$  \hspace{1cm} (6.25)

Then for a sufficiently large $M > 0$,

$$P_0^\Xi \Pi_n(d(P, P_0) \geq M\varepsilon_n \mid X_1, \ldots, X_n) \to 0.$$  \hspace{1cm} (6.26)

(Note that the assertion establishes convergence in $P_0$-expectation, which implies convergence in $P_0$-probability because the posterior is bounded.) The rate theorem given here is a variation on theorem 2.1 in Ghosal, Ghosh and Van der Vaart (2000) [101]; their version is different in two respects. First of all they express the testing condition through a sufficient condition based on the model’s entropy numbers. We come back to this point in subsection 6.6.3. Secondly, they restrict attention to a sequence of models $\mathcal{R}_n$ that grows in $\Pi_n$-measure to the full model $\mathcal{P}$ sufficiently fast,

$$\Pi_n(\mathcal{R}_n \setminus \mathcal{P}_n) \leq e^{-nL\varepsilon_n^2}.$$  \hspace{1cm} (6.27)

The submodels $(\mathcal{R}_n)$ are then used to express the entropy condition and referred to as a **sieve** (pronounced with an ending like the word “give”) that approximates $\mathcal{P}$ quickly enough with growing $n$, c.f. (6.27). This separation between submodels of controlled entropy and complements of bounded prior mass is due to Barron (1988) [9] and Barron et al. (1999) [14].

In subsections 6.6.3 and 6.6.4, we analyse conditions (6.25) and (6.24) separately. First, we prove theorem 6.6.

### 6.6.2 Proof of the Ghosal-Ghosh-van der Vaart theorem

**Proof.** Define, for every $\eta > 0$, $A(\eta) = \{ P \in \mathcal{P} : d(P, P_0) \geq \eta \}$. The expectation in (6.26) can be decomposed using the tests $\phi_n$; for every $n \geq 1$ and every $M > 1$, we have:

$$P_0^n \Pi_n(A(M\varepsilon_n) \mid X_1, \ldots, X_n)$$

$$= P_0^n \phi_n(X) \Pi_n(A(M\varepsilon_n) \mid X_1, \ldots, X_n) + P_0^n(1 - \phi_n)(X) \Pi_n(A(M\varepsilon_n) \mid X_1, \ldots, X_n).$$

We estimate the terms on the right-hand side separately. Due to the first inequality in (6.25), the first term converges to zero. To estimate the second term, we (assume
that $\mathcal{D}$ is dominated, in a non-essential way, see remark:dominationcondition below) and substitute (2.13) to obtain,
\[
P_0^0 \Pi \left( A(M\varepsilon_n) \mid X_1, \ldots, X_n \right) (1 - \phi_n)(X)
= P_0^0 \left[ \int_{A(M\varepsilon_n)} \prod_{i=1}^{n} \frac{P}{P_0}(X_i) d\Pi(P) (1 - \phi_n)(X) \right] = \int_{\mathcal{D}} \prod_{i=1}^{n} \frac{P}{P_0}(X_i) d\Pi(P) \tag{6.28}
\]
in which the denominator can be lower-bounded by application of lemma 6.7, since by assumption (6.24), $\Pi(B(\varepsilon_n; P_0)) > 0$. Let $\Omega_n$ be the subset in $\mathcal{D}^n$ for which the inequality between left- and right-hand sides in the following display holds:
\[
\int_{\mathcal{D}} \prod_{i=1}^{n} \frac{P}{P_0}(X_i) d\Pi(P) \geq \int_{B(\varepsilon_n; P_0)} \prod_{i=1}^{n} \frac{P}{P_0}(X_i) d\Pi(P) \geq e^{-(1+K)nt_n^2} \Pi(B(\varepsilon_n; P_0)),
\]
as in (6.35), with $K > 0$ as yet unspecified. Decomposing the $P_0^0$-expectation in (6.28) into separate integrals over $\Omega_n$ and $\mathcal{D}^n \setminus \Omega_n$, we find:
\[
P_0^0 \Pi \left( A(M\varepsilon_n) \mid X_1, \ldots, X_n \right) (1 - \phi_n)
\leq P_0^0 \Pi \left( A(M\varepsilon_n) \mid X_1, \ldots, X_n \right) (1 - \phi_n) 1_{\Omega_n} + P_0^0 (\mathcal{D}^n \setminus \Omega_n).
\]
Note that $P_0^0 (\mathcal{D}^n \setminus \Omega_n) = o(1)$ as $n \to \infty$ according to (6.35). The first term is estimated as follows:
\[
P_0^0 \Pi \left( A(M\varepsilon_n) \mid X_1, \ldots, X_n \right) (1 - \phi_n) 1_{\Omega_n}
\leq e^{(1+K)nt_n^2} P_0 \left( 1 - \phi_n \right) \int_{A(M\varepsilon_n)} \prod_{i=1}^{n} \frac{P}{P_0}(X_i) d\Pi(P)
\leq e^{(1+K)nt_n^2} \frac{\Pi(A(M\varepsilon_n))}{\Pi(B(\varepsilon_n; P_0))} \sup_{\varepsilon \in A(M\varepsilon_n)} P\left( 1 - \phi_n \right),
\]
where we have substituted (6.29) and used the positivity of the integrand, applied Fubini’s theorem and bounded the integrand by its supremum over $A(M\varepsilon_n)$. Application of the second inequality in (6.25) gives:
\[
P_0^0 \Pi \left( A(M\varepsilon_n) \mid X_1, \ldots, X_n \right) (1 - \phi_n) \leq e^{(1+K+C-M^2)t_n^2} \Pi(B(\varepsilon_n; P_0)) + o(1).
\]
Hence, for all $K > 0$ there exists a constant $M > 0$ such that the above expression converges to zero. This leads us to conclude that:
\[
P_0^0 \Pi \left( A(M\varepsilon_n) \mid X_1, \ldots, X_n \right) \to 0, \quad (n \to \infty),
\]
for sufficiently large \( M > 0 \).

### 6.6.3 Entropy numbers and the existence of test sequences

Instead of condition (6.25), Ghosal et al. impose the following alternative condition. (Recall that the packing number \( D(\eta, \mathcal{X}, \rho) \) of a space \( \mathcal{X} \) with metric \( \rho \) is defined as the maximal number of points in \( \mathcal{X} \) such that the \( \rho \)-distance between all pairs is at least \( \eta \). This number is related to the so-called covering number \( N(\eta, \mathcal{X}, \rho) \) which is defined as the minimal number of \( \rho \)-balls of radius \( \eta \) needed to cover \( \mathcal{X} \), by the following inequalities: \( N(\eta, \mathcal{X}, \rho) \leq D(\eta, \mathcal{X}, \rho) \leq N(\eta/2, \mathcal{X}, \rho) \).)

(ii) The \( \epsilon \)-packing numbers \( D(\epsilon, \mathcal{P}_n, d) \) for the models \( \mathcal{P}_n \) satisfy:

\[
D(\epsilon_n, \mathcal{P}_n, d) \leq e^{n\epsilon^2}. \tag{6.31}
\]

If \( d \) is the Hellinger metric \( H \), entropy condition (6.31) implies the existence of a suitable sequence of test functions (see Birg"e (1983,1984) [32, 33] and Le Cam (1973,1986) [160, 162]). Indeed, the vast majority of non-parametric applications of Bayesian methods in the literature is based on this point, which is an application of the minimax theorem, theorem 2.4. We follow Le Cam (1986) [162], section 16.4, specialized to i.i.d.-context.

**Proposition 6.3. (Minimax Hellinger tests)**

Consider a model \( \mathcal{P} \) of single-observation distributions \( P \) for i.i.d. data. Let \( B, V \subset \mathcal{P} \) be convex with \( H(B, V) > 0 \). For every \( n \geq 1 \), there exists a test \( (\hat{\phi}_n) \) such that,

\[
\sup_{P \in B} P^n \phi_n + \sup_{Q \in V} Q^n (1 - \phi_n) \leq e^{-nH^2(B, V)}.
\]

**Proof.** Fix \( n \geq 1 \). We consider the product-space \( \mathcal{X}^n \) in which the \( n \)-th sample \((X_1, \ldots, X_n)\) takes its values: denote the space of all signed measures on \( \mathcal{X}^n \) by \( L_n \) (with convex subspace \( \Pi_n \) of all sub-probability measures \( P \in L_n, 0 \leq P \leq 1 \)) and the space of all bounded, measurable maps \( \mathcal{X}^n \rightarrow \mathbb{R} \) by \( M_n \) (with convex subspace \( \Phi_n \) of all \( \phi \in M_n, 0 \leq \phi \leq 1 \)). The bi-linear form \( L_n \times M_n \rightarrow \mathbb{R} : (\mu, f) \mapsto \int f \, d\mu \) places \( L_n \) and \( M_n \) in dual correspondence, with corresponding locally convex topologies \( \sigma(L_n, M_n) \) and \( \sigma(M_n, L_n) \). We then define \( B^n = \{ P^n : P \in B \} \) and \( V^n = \{ Q^n : Q \in V \} \) and write \( \text{co}(B^n), \text{co}(V^n) \subset \Pi_n \) for their convex hulls, which form the domain of the risk-function \( R_n : \text{co}(B^n) \times \text{co}(V^n) \times M_n \rightarrow [0, 1] \) that adds up error probabilities:

\[
R_n(P, Q, \phi) = P \phi + Q (1 - \phi).
\]

The dependence of \( R_n \) on both \((P, Q)\) and \( \phi \) is linear and \( \phi \rightarrow R_n(P, Q, \phi) \) is \( \sigma(M_n, L_n) \)-continuous, for all \((P, Q) \in \text{co}(B^n) \times \text{co}(V^n) \). Because \( \Phi_n \) is weakly bounded, it is pre-compact in \( \sigma(M_n, L_n) \); by the bi-polar theorem \( \Phi_n = \Phi_n^\circ \), so \( \Phi_n \) is closed, hence complete in (the complete space) \( M_n \). Conclude that \( \Phi_n \) is a compact
subspace of $M_n$. It now follows from the minimax theorem, theorem 2.4, that there exists a test function $\phi_n \in \Phi_n$ such that,

$$\sup \{ P\phi_n + Q(1 - \phi_n) : P \in \text{co}(B^n), Q \in \text{co}(V^n) \}$$

$$= \inf_{\phi \in \Phi} \sup \{ P\phi + Q(1 - \phi) : P \in \text{co}(B^n), Q \in \text{co}(V^n) \}$$

$$= \sup \{ \inf_{\phi \in \Phi} (P\phi + Q(1 - \phi)) : P \in \text{co}(B^n), Q \in \text{co}(V^n) \}$$

$$\leq \sup \left\{ \int_{\mathcal{X}} \sqrt{p(x^n)q(x^n)} \, d\mu(x^n) : P \in \text{co}(B^n), Q \in \text{co}(V^n) \right\},$$

where the last inequality follows as in example 2.15. As is shown in lemma 2 of section 16.4 of [162], the expression on the r.h.s. above factorizes appropriately,

$$\sup \left\{ \int_{\mathcal{X}} \sqrt{p(x^n)q(x^n)} \, d\mu(x^n) : P \in \text{co}(B^n), Q \in \text{co}(V^n) \right\} \leq \sup \left\{ \int_{\mathcal{X}} \sqrt{p(x)q(x)} \, d\mu(x) : P \in B, Q \in V \right\}^n,$$

The observation that the r.h.s. above is dominated by $1 - H^2(P, Q) \leq e^{-H^2(P, Q)}$ then concludes the proof.

The above proposition is applied particularly, with either or both $B$ and $V$ that are Hellinger balls at non-zero Hellinger distance from one another, using the following step that combines individual uniform tests for $B$ versus several (possibly overlapping) alternatives into one uniform test for $B$ versus their union.

**Lemma 6.6.** Fix $n, N \geq 1$ and $B, V_1, \ldots, V_N \subset \mathcal{P}$. If there exist test functions $\phi_i$, $(1 \leq i \leq N)$, such that,

$$\sup_{P \in B} P^n \phi_i + \sup_{Q \in V_i} Q^n (1 - \phi_i) \leq e^{-nH^2(B, V_i)}, \tag{6.32}$$

then there exists a test function $\psi$ such that,

$$\sup_{P \in B} P^n \psi + \max_{1 \leq i \leq N} \sup_{Q \in V_i} Q^n (1 - \psi) \leq N e^{-n \min_{1 \leq i \leq N} [H^2(B, V_i)]}.$$ 

**Proof.** Define $\psi = \max \{ \phi_i : 1 \leq i \leq N \}$, then for any $P \in B$,

$$P^n \psi \leq \sum_{i=1}^{N} P^n \phi_i \leq N \max \{ P^n \phi_i : 1 \leq i \leq N \}$$

and for every $Q \in V$,
\[ Q^\alpha(1 - \psi) = Q^\alpha\{\min(1 - \phi_i) : 1 \leq i \leq N\} \]
\[ \leq \sum_{i=1}^{N} Q^\alpha(1 - \phi_i) \leq N \max\{Q^\alpha(1 - \phi_i) : 1 \leq i \leq N\}. \]

Combination leads to (6.32).

**Corollary 6.4.** Let \( n, N \geq 1 \), a convex \( B \subset \mathcal{P} \) and \( V \) with convex \( V_1, \ldots, V_N \) be given, such that,
\[ V \subset \bigcup_{i=1}^{N} V_i, \quad \varepsilon = \min_{1 \leq i \leq N} H(B, V_i) > 0, \]
then there exists a test function \( \phi \) such that,
\[ \sup_{P \in B} P^\alpha \psi + \sup_{Q \in V} Q^\alpha(1 - \psi) \leq N e^{-n\varepsilon^2}. \]

Now it is clear how the existence of test sequences \( (\phi_n) \) as in (6.25) follows from the entropy condition (6.31). For some \( M > 2 \), define \( B_n = \{P_0\}, V_n = \{P \in \mathcal{P} : H(P, P_0) \geq M\varepsilon_n\} \) and \( N = N(\varepsilon_n, V_n, H) \leq N(\varepsilon_n, \mathcal{P}, H) \leq D(\varepsilon_n, \mathcal{P}, H) \leq e^{n\varepsilon^2} \). The convex cover \( V_1, \ldots, V_N \) of \( V_n \) consists of Hellinger balls of radius \( \varepsilon_n \) (Hellinger balls are convex, see lemma 3 of section 16.4 of [162]) centred in \( V_n \), each of which satisfies \( H(B, V_i) \geq (M - 1)\varepsilon_n \) by virtue of the triangle inequality. Corollary 6.4 then says there exists a test sequence \( (\phi_n) \) such that, for every \( n \geq 1 \),
\[ \sup_{P \in B} P^\alpha \psi + \sup_{Q \in V_n} Q^\alpha(1 - \psi) \leq N e^{-n(M - 1)\varepsilon_n^2} \leq e^{-n(M - 2)\varepsilon_n^2}, \]
which is enough for (6.25).

The sufficient condition for existence of suitable tests that Ghosal *et al.* [101] employ, the entropy condition (6.31), has to be proved in individual cases, however, and that specifies the range of applicability quite somewhat further than some like to admit.

**Example 6.2.** In certain, well-chosen infinite-dimensional spaces (like Sobolev balls, VC-classes or classes of monotone functions, see [218, 106]), entropy numbers are known. Suppose that we have a parametrized model \( \Theta \rightarrow \mathcal{P} : \theta \rightarrow P_\theta \) for single observations and \( \Theta \) is one of these examples, a subspace of a normed space with norm \( \| \cdot \| \), such that, for some sequence \( \eta_n \),
\[ \log N(\eta_n, \Theta, \| \cdot \|) \leq n\eta_n^2, \]
and also that there exist two constants \( K > 0, \alpha > 0 \) such that for all \( \theta_1, \theta_2 \in \Theta \) that are close enough, the Hellinger metric \( H \) is related to \( \| \cdot \| \) through,
\[ H(P_{\theta_1}, P_{\theta_2}) \leq K\| \theta_1 - \theta_2 \|. \]
Then \( \| \cdot \| \)-balls of radius \( \eta \) in \( \Theta \) are mapped into \( H \)-balls of radius \( K\eta \), so that \( \mathcal{P} \) has a cover of \( H \)-balls of radius \( K\eta_n \) of order \( N(\eta_n, \Theta, \| \cdot \|) \). Therefore, condition (6.31)
for the rate $\epsilon_n$ is determined by the entropy bound (6.33) for the parametrizing space. So if the model has a parameter space of the special kind for which entropy bounds can be calculated, and a relation of the Lipschitz-type (6.34) (or slightly more involved, see (9.26) in chapter 9, for example) applies, then the existence of uniform test sequences can be guaranteed. Unfortunately, the number of examples of parameter spaces with known entropy bounds is somewhat limited [218]. We come back to this point and alternatives in chapters 7 and 8.

6.6.4 Lower bounds on prior mass

To conclude this section we give the lemma needed in the proof of theorem 6.6 to lower-bound the denominator of the posterior in probability, leading to lower bounds on prior mass locally around the true distribution of the data. This lemma, presently more a technical afterthought than an integral part of the theory, will be drawn to the foreground when we discuss remote contiguity in chapter 7.

**Lemma 6.7.** Let $\epsilon > 0$ and $p_0 \in \mathcal{P}$ be given and let $B(\epsilon; p_0)$ be defined as in (6.23). If $\Pi(B(\epsilon; p_0)) > 0$, then for every $K > 0$

$$P_0^\epsilon \left( \int_{B(\epsilon; p_0)} \prod_{i=1}^n \frac{p_i}{p_0}(X_i) d\Pi(P) \leq e^{-n\epsilon^2(1+K)} \Pi(B(\epsilon; p_0)) \right) \leq \frac{1}{nK^2 \epsilon^2}. \quad (6.35)$$

**Proof.** Write $\Pi'$ for $\Pi(\cdot | B(\epsilon; p_0))$, the prior conditioned on $B(\epsilon; p_0)$. By Jensen’s inequality,

$$\log \int \prod_{i=1}^n \frac{p_i}{p_0}(X_i) d\Pi'(P) \geq \sum_{i=1}^n \int \log \frac{p_i}{p_0}(X_i) d\Pi'(P).$$

Therefore,

$$P_0^\epsilon \left( \int \prod_{i=1}^n \frac{p_i}{p_0}(X_i) d\Pi'(P) \leq e^{-n\epsilon^2(1+K)} \right) \leq P_0^\epsilon \left( \sqrt{n}(p_n - p_0) \int \log \frac{p_i}{p_0} d\Pi'(P) \right) \leq -\sqrt{n}(1+C)\epsilon^2 - \sqrt{n}p_0 \int \log \frac{p_i}{p_0} d\Pi'(P). \quad (6.36)$$

With the help of Fubini’s theorem and the definition of $B(\epsilon)$, we see that the r.h.s. above is bounded by $-\sqrt{n}C\epsilon^2$. Note that the variance of the integrated log-likelihood is bounded with Jensen’s inequality,

$$\text{Var} \left( \int \log \frac{p_i}{p_0} d\Pi'(P) \right) \leq p_0 \int \left( \log \frac{p_i}{p_0} \right)^2 d\Pi'(P), \quad (6.37)$$
6.7 Frequentist counterexamples

To demonstrate that the assertion of Doob’s Bayesian consistency theorem can be much weaker than expected in non-parametric setting, first Schwartz and then Freedman constructed counterexamples in the early 1960’s, illustrated in sub-section 6.7.1. Freedman’s counterexamples and subsequent examples of posterior inconsistency established a widespread conviction that Bayesian methods were wholly unfit for frequentist purposes, at least in non-parametric context. By contrast Schwartz’s theorem, theorem 6.5 asserts posterior consistency with frequentist rigour.

Over the decades examples of problematic posterior behaviour in non-parametric setting continued to captivate [66, 67, 57, 68, 69, 95, 96], while Schwartz’s theorem received initially limited but steadily growing amounts of attention [100]; subsequent frequentist theorems (e.g. by Barron [9], Barron-Schervish-Wasserman [14], Ghosal-Ghosh-van der Vaart [101], Shen-Wasserman [204], Walker [221] and Walker-Lijoi-Prünster [224], Kleijn-Zhao [145] and many others) have extended the applicability of theorem 6.5 but not its essence, condition (6.15) for the prior.

Regarding the material presented in sections 6.5 and 6.6 the matter is more complicated. Certainly, the work of Ghosal, Ghosh and van der Vaart [101] has been very influential and, together with Schwartz’s classical theorem, forms the frequentist, theoretical backbone for the vast majority of the literature on non-parametric Bayesian statistics since 2000, as described with encyclopaedic completeness in the book by Ghosal and van der Vaart (2018) [106]. But as demonstrated in sub-section 6.7.2, there are very simple parametric models in which the Kullback-Leibler priors of Schwartz or their more specific Ghosal-Ghosh-van der Vaart variations do not exist (while posterior consistency obtains without problems, as we shall see in chapter 7).

6.7.1 Freedman’s counterexamples

The first examples of unexpected posterior inconsistency are due to Schwartz (1961) [200], but it was Freedman (1963) [92] who made the point famous with a simple non-parametric counterexample, discussed in detail as example 6.3 below. In Freedman (1965) [93] it was even shown that inconsistency is generic in a topological sense (see theorem 6.7).

Example 6.3. (Freedman (1963) [92])
Consider a sample \(X_1, X_2, \ldots\) of random positive integers. Denote the space of all
probability distributions on \( \mathbb{N} \) by \( \Lambda \) and assume that the sample is i.i.d.-\( P_0 \), for some \( P_0 \in \Lambda \). For any \( P \in \Lambda \), write \( p(i) = P(X = i) \) for all \( i \geq 1 \). The total-variational and weak topologies on \( \Lambda \) are equivalent (defined, \( P \to Q \) if \( p(i) \to q(i) \) for all \( i \geq 1 \)). Let \( Q \in \Lambda \setminus \{ P_0 \} \) be given. To arrive at a prior with \( P_0 \) in its support, leading to a posterior that concentrates on \( Q \), we consider sequences \( (P_m) \) and \( (Q_m) \) such that \( Q_m \to Q \) and \( P_m \to P_0 \) as \( m \to \infty \). The prior \( \Pi \) places masses \( \alpha_m > 0 \) at \( P_m \) and \( \beta_m > 0 \) at \( Q_m \) \((m \geq 1)\), so that \( P_0 \) lies in the support of \( \Pi \). A careful construction of the distributions \( Q_m \) that involves \( P_0 \), guarantees that the posterior satisfies,

\[
\frac{\Pi(\{Q_m\}|X^n)}{\Pi(\{Q_{m+1}\}|X^n)} \xrightarrow{P_0\text{-a.s.}} 0,
\]

that is, posterior mass is shifted further out into the tail as \( n \) grows to infinity, forcing all posterior mass that resides in \( \{Q_m : m \geq 1\} \) into arbitrarily small neighbourhoods of \( Q \). In a second step, the distributions \( P_m \) and prior weights \( \alpha_m \) are chosen such that the likelihood at \( P_m \) grows large for high values of \( m \) and small for lower values as \( n \) increases, so that the posterior mass in \( \{P_m : m \geq 1\} \) also accumulates in the tail. However, the prior weights \( \alpha_m \) may be chosen to decrease very fast with \( m \), in such a way that,

\[
\frac{\Pi(\{P_m : m \geq 1\}|X^n)}{\Pi(\{Q_m : m \geq 1\}|X^n)} \xrightarrow{P_0\text{-a.s.}} 0,
\]

thus forcing all posterior mass into \( \{Q_m : m \geq 1\} \) as \( n \) grows. Combination of the previous two displays leads to the conclusion that for every neighbourhood \( U_Q \) of \( Q \),

\[
\Pi(U_Q|X^n) \xrightarrow{P_0\text{-a.s.}} 1,
\]

so the posterior is inconsistent. Other choices of the weights \( \alpha_m \) that place more prior mass in the tail do lead to consistent posterior distributions.

Some objected to Freedman’s counterexample, because knowledge of \( P_0 \) is required to construct the prior that causes inconsistency. So it was possible to argue that Freedman’s counterexample amounted to nothing more than a demonstration that unfortunate circumstances could be created, probably not a fact of great concern in any generic sense.

To strengthen Freedman’s point one would need to construct a prior of full support without explicit knowledge of \( P_0 \). In the setting of example 6.3, denote the space of all distributions on \( \Lambda \) by \( \pi(\Lambda) \). Note that since \( \Lambda \) is Polish, so is \( \pi(\Lambda) \) and so is the product \( \Lambda \times \pi(\Lambda) \).

**Theorem 6.7.** (Freedman (1965) [93])

Let \( X_1, X_2, \ldots \) form an sample of i.i.d.-\( P_0 \) random integers, let \( \Lambda \) denote the space of all distributions on \( \mathbb{N} \) and let \( \pi(\Lambda) \) denote the space of all Borel probability measures on \( \Lambda \), both in Prohorov’s weak topology. The set of pairs \( (P_0, \Pi) \in \Lambda \times \pi(\Lambda) \) such that for all open \( U \subset \Lambda \),

\[
\limsup_{n \to \infty} P_0^n \Pi(U|X^n) = 1,
\]
And so, the set of pairs \((P_0, \Pi) \in \Lambda \times \pi(\Lambda)\) for which the limiting behaviour of the posterior is acceptable to the frequentist, is meagre in \(\Lambda \times \pi(\Lambda)\). The proof is based on example 2.5. The question arises, what is the conclusion we draw from Freedman’s objections of inconsistency? (Which he was famous for raising, not only in Bayesian context \([66, 67, 69, 95, 96]\) (do not miss Le Cam’s comment \([163]\) on \([66]\)), but also in other fields \([94]\).) Leaving constructions with intentional pathology aside, it is theorem 6.7 that poses the real challenge to non-parametric Bayesian statistics. However, its message is quite encouraging when read correctly: meagreness in the sense of theorem 6.7 means that there is a condition missing: not all priors are fit for frequentist purpose, indeed a (topologically) large subset of priors are not. The remaining priors, those that are useful to the frequentist, form a (topologically) small subset characterized by a property. Freedman failed to recognize that his result was indicative of the next step in the theoretical development: Schwartz’s Kullback-Leibler condition \((6.15)\) provides exactly such a property (which may explain the tone of \([163]\)).

### 6.7.2 Counterexamples: Schwartz and GGV conditions

In this subsection, it is shown that there exist very simple parametric models in which no prior satisfies Schwartz’s Kullback-Leibler condition \((6.15)\), and similarly, that there are very simple parametric models in which Schwartz’s Kullback-Leibler condition may be satisfied, but no prior satisfies the Ghosal-Ghosh-van der Vaart condition \((6.24)\).

**Example 6.4.** Consider \(X_1, X_2, \ldots\) that are i.i.d. \(P_0\) with Lebesgue density \(p_0: \mathbb{R} \to \mathbb{R}\) supported on an interval of known width (say, 1) but unknown location. Parametrize in terms of a continuous density \(\eta\) on \([0, 1]\) with \(\eta(x) > 0\) for all \(x \in [0, 1]\) and a location \(\theta \in \mathbb{R}\): \(p_{\theta, \eta}(x) = \eta(x - \theta) \mathbb{1}_{[\theta, \theta+1]}(x)\). A moment’s thought makes clear that if \(\theta \neq \theta'\),

\[-P_{\theta, \eta} \log \frac{p_{\theta', \eta'}}{p_{\theta, \eta}} = \infty,\]

for all \(\eta, \eta'\). Therefore Kullback-Leibler neighbourhoods do not have any extent in the \(\theta\)-direction and no prior is a Kullback-Leibler prior in this model. Nonetheless the posterior is consistent (see examples ?? and ??).

**Example 6.5.** Consider an i.i.d. sample of integers \(X_1, X_2, \ldots\) from a distribution \(P_a, (a \geq 1)\), defined by,

\[p_a(k) = P_a(X = k) = \frac{1}{Z_a} k^{-a} (\log k)^{3} \]  \hspace{1cm} (6.38)

for all \(k \geq 2\), with \(Z_a = \sum_{k \geq 2} k^{-a} (\log k)^{-3} < \infty\). As it turns out, for \(a = 1, b > 1\),
Therefore, Schwartz’s KL-condition (6.15) for the prior for the parameter $a$ can be satisfied but there exists no prior such that (6.24) is satisfied for all $P_0$ in the model. (See example ?? for more.) In fact, if we change the third power of the log-factor in the denominator of (6.38) to a square, Schwartz’s KL-priors also do not exist.

The above constructions are indicative of a more general problem: for any $P_0$ it is possible to find distributions $P$ with densities $p$ that are ‘wild enough’ to cause log-likelihood ratios $\log p / p_0$ to loose integrability or square-integrability. Originating in early analyses of posterior inconsistency [14, 66], the phenomenon of data-tracking [223] sketches a similar qualitative picture of situations where posterior consistency fails.

6.8 Exercises [EMPTY]

6.1. Find the mistake in the proof of lemma 6.1 and speculate on possible solutions, to conclude that there is no easy way out. [Hint: Recall that conditional probabilities are defined almost-surely for every $A$ separately, c.f. definition (B.8), but not automatically also almost-surely for all $A$ simultaneously, as a regular conditional distribution. The present problem is similar.]
Chapter 7
Frequentist validity of Bayesian limits

In this chapter, we redevelop the Bayesian theory of non-parametric statistics of the previous chapter from the ground up. The calculations presented in chapter 6 are more than adequate if one is willing to restrict attention to the “traditional” setting for examples in non-parametric statistics, where the data forms an i.i.d. sample from a distribution in a model of Hellinger entropy with known upper-bound, following the path set out in subsections 6.6.3 and 6.6.4.

The setting for present-day non-parametric statistical challenges is more general, though: data in the computer age is not only of very large scale, it is often of much more complex structure than that of an i.i.d. data set. Dependence among data points not only occurs in stochastic processes, like the time-series that typify data in financial markets, but also in random walks and branching processes that occur on graphs, like the widely used world-wide-web sampling technique of webcrawling. Non-parametric models for complex data like those that arise from questions in machine learning and network science are usually not compatible with the technical formulation of subsections 6.6.3 and 6.6.4. (The iterated maps of so-called deep neural networks and the highly dependent preferential attachment model for random graphs are two examples, and so is the community-detection problem of chapter 10.)

Below we re-examine for which priors Bayesian limits are limits valid in the frequentist sense: is Schwartz’s Kullback-Leibler condition perhaps a manifestation of a more general notion? The argument leads to other questions for which insightful answers have been elusive: why is Doob’s theorem completely different from Schwartz’s? Why does weak consistency in the full non-parametric model (e.g. with the Dirichlet process prior [89], or more modern variations [60]) reside in a corner of its own (with tailfreeness [93, 87] as sufficient property of the prior), apparently unrelated to posterior consistency in either Doob’s or Schwartz’s views? And to extend the scope further, what can be said about hypothesis testing, classification, model selection, etcetera? Given that the Bernstein-von Mises theorem cannot be expected to hold in any generality outside parametric setting [57, 96], what relationship exists between credible sets and confidence sets?
The central property to enable frequentist interpretation of posterior asymptotics is defined as \textit{remote contiguity} in section 7.2. It expresses a weakened form of Le Cam’s contiguity, relating the true distribution of the data to localized prior predictive distributions. Where Schwartz’s Kullback-Leibler neighbourhoods represent a choice for the localization appropriate when the sample is \textit{i.i.d.}, remote contiguity generalises the notion to include non-\textit{i.i.d.} samples and sample-size-dependent model/prior pairs. We then see how Doob’s prior-almost-sure consistency is strengthened to reach Schwartz’s frequentist conclusion, or how a test that is consistent prior-almost-surely becomes a test that is consistent in \textit{all} points of the model, or how a Bayesian credible set can be “enlarged” to serve as a frequentist confidence set asymptotically. The latter point extends the main implication of the Bernstein-von Mises theorem, theorem 4.4, to non-parametric models provided the prior induces remote contiguity.

In section 7.1 we concentrate on an inequality that relates testing to posterior concentration and indicates the relation with Le Cam’s inequality. Section 7.2 introduces remote contiguity and the analogue of Le Cam’s First Lemma. In section 7.4, frequentist theorems on the asymptotic behaviour of posterior distributions are proved, on posterior consistency, on rates of convergence, on model selection with posterior odds and on the conversion of credible sets to confidence sets. Section 7.8 formulates the conclusions. The central condition of testability is analysed further in chapter 8. Application to community detection in random graphs follows in chapter 10.

\section*{7.1 Posterior concentration and asymptotic tests}

First we consider a lemma that relates concentration of posterior mass in certain model subsets to the existence of test sequences that distinguish between those subsets. More precisely, it is shown that the expected posterior mass outside a model subset \( V \) with respect to the local prior predictive distribution over a model subset \( B \), is upper bounded (roughly) by the testing power of \textit{any} statistical test for the hypotheses \( B \) versus \( V \): if a test sequence exists, the posterior will concentrate its mass appropriately.

\subsection*{7.1.1 Bayesian test sequences}

We follow Schwartz and consider asymptotic testing; however, we define test sequences immediately in Bayesian context by involving priors from the outset. We consider sequentially observed, (possibly non-\textit{i.i.d.}) samples \( X^n \), distributed according to \( P_{\theta_0,n} \) for some \( \theta_0 \in \Theta \), within the model \( \theta \to P_{\theta,n} \). (More generally, we refer to appendix A for the notation and conventions assumed through this chapter.)
Definition 7.1. Given priors \((\Pi_n)\) on the measurable space \((\Theta, \mathcal{G})\), model subsets \((B_n), (V_n) \subset \mathcal{G}\) and \(\alpha_n \downarrow 0\), a sequence of \(\mathcal{G}_n\)-measurable maps \(\phi_n : \mathcal{G}_n \to [0, 1]\) is called a Bayesian test sequence for \(B_n\) versus \(V_n\) (under \(\Pi_n\)) of power \(\alpha_n\), if,
\[
\int_{B_n} P_{\theta,n} \phi_n d\Pi_n(\theta) + \int_{V_n} P_{\theta,n} (1 - \phi_n) d\Pi_n(\theta) = o(\alpha_n). \tag{7.1}
\]
We say that \((\phi_n)\) is a Bayesian test sequence for \(B_n\) versus \(V_n\) (under \(\Pi_n\)) if (7.1) holds for some \(\alpha_n \downarrow 0\).

Note that if we have sequences \((C_n)\) and \((W_n)\) such that \(C_n \subset B_n\) and \(W_n \subset V_n\) for all \(n \geq 1\), then a Bayesian test sequence for \((B_n)\) versus \((V_n)\) of power \(\alpha_n\) is a Bayesian test sequence for \((C_n)\) versus \((W_n)\) of power (at least) \(\alpha_n\).

Lemma 7.1. For any \(B, V \in \mathcal{G}\) and any measurable \(\phi : \mathcal{G} \to [0, 1]\),
\[
\int_B P_\theta \Pi(V|X) d\Pi(\theta) \leq \int_B P_\theta \phi d\Pi(\theta) + \int_V P_\theta (1 - \phi) d\Pi(\theta). \tag{7.2}
\]
Proof. Due to Bayes’s Rule \((\text{A.4})\) and monotone convergence,
\[
\int (1 - \phi(X)) \Pi(V|X) dP \Pi = \int_V P_\theta (1 - \phi(X)) d\Pi(\theta).
\]
Accordingly,
\[
\int_B P_\theta (1 - \phi) \Pi(V|X) d\Pi(\theta) \leq \int (1 - \phi) \Pi(V|X) dP \Pi = \int_V P_\theta (1 - \phi) d\Pi(\theta).
\]
Inequality (7.2) follows from the fact that \(\Pi(V|X) \leq 1\).

So the mere existence of a test sequence is enough to guarantee posterior concentration, a fact expressed in \(n\)-dependent form through the following proposition.

Proposition 7.1. Let \((\mathcal{G}_n), (\Theta, \mathcal{G}), (\mathcal{P}_n)\) and \((\Pi_n)\) be given. Given sequences \((B_n), (V_n) \subset \mathcal{G}\) and \(\alpha_n, b_n \downarrow 0\) such that \(\alpha_n = o(b_n)\) and \(\Pi_n(B_n) \geq b_n > 0\), we assume there exists a Bayesian test sequence for \(B_n\) versus \(V_n\) of power \(\alpha_n\). Then,
\[
P_n^{\Pi_n/B_n} \Pi(V_n|X^n) = o(\alpha_n b_n^{-1}). \tag{7.3}
\]
for all \(n \geq 1\).

To see how this leads to posterior consistency, consider the following: if the model subsets \(V_n = V\) are all equal to the complement of a neighbourhood \(U\) of \(P_0\), and the \(B_n\) are chosen such that the expectations of the random variables \(X^n \to \Pi(V|X^n)\) under \(P_0^{\Pi_n/B_n}\) “dominate” their expectations under \(P_{0,n}\) in a suitable way, sufficiency of prior mass \(b_n\) given testing power \(\alpha_n \downarrow 0\), is enough to assert that \(P_{0,n} \Pi(V|X^n) \to 0\), so an arbitrarily large fraction of posterior mass is found in \(U\) with high probability for \(n\) large enough.
7.1.2 Existence of Bayesian test sequences

Lemma 7.1 and proposition 7.1 require the existence of test sequences of the Bayesian type. That question is unfamiliar, frequentists are used to test sequences for pointwise or uniform testing, e.g. those of subsection 6.6.3. Another example is formed by complements of weak neighbourhoods, which are testable uniformly as we shall see in chapter 8.

Requiring the existence of a Bayesian test sequence c.f. (7.1) is quite different. We shall illustrate this point in various ways below. First of all the existence of a Bayesian test sequence is linked directly to behaviour of the posterior itself.

Theorem 7.1. Let \((\Theta, \mathcal{G}, \Pi)\) be given and assume that there is a coupling \(X \in \mathcal{X}^\infty\) with distribution \(P_\theta\) and marginals \(X^n \sim P_{\theta,n}\) for every \(\theta \in \Theta\) and \(n \geq 1\). For any \(B, V \in \mathcal{G}\) with \(\Pi(B) > 0, \Pi(V) > 0\), the following are equivalent:

(i) there are \(\mathcal{B}_n\)-msb. \(\phi_n : \mathcal{X}_n \to [0,1]\) such that for \(\Pi\)-almost-all \(\theta \in B, \theta' \in V,\)

\[
\phi_n(X^n) \xrightarrow{P_{\theta,n}-a.s.} 0, \quad \phi_n(X^n) \xrightarrow{P_{\theta',n}-a.s.} 1,
\]

(ii) there are \(\mathcal{B}_n\)-msb. \(\phi_n : \mathcal{X}_n \to [0,1]\) such that for \(\Pi\)-almost-all \(\theta \in B, \theta' \in V,\)

\[P_{\theta,n}\phi_n \to 0, \quad P_{\theta',n}(1 - \phi_n) \to 0,\]

(iii) there are \(\mathcal{B}_n\)-msb. \(\phi_n : \mathcal{X}_n \to [0,1]\) such that,

\[
\int_B P_{\theta,n} \phi_n d\Pi(\theta) + \int_V P_{\theta,n}(1 - \phi_n) d\Pi(\theta) \to 0,
\]

(iv) for \(\Pi\)-almost-all \(\theta \in B, \theta' \in V,\)

\[
\Pi(V | X^n) \xrightarrow{P_{\theta,n}-a.s.} 0, \quad \Pi(B | X^n) \xrightarrow{P_{\theta',n}-a.s.} 0.
\]

Proof. Condition (i) implies (ii) trivially and (ii) implies (iii) by dominated convergence. Assume (iii) and note that by lemma 7.1,

\[
\int P_{\theta,n} \Pi(V | X^n) d\Pi(\theta | B) \to 0.
\]

With the coupling \(X\) of the observations \(X^n\), martingale convergence in \(L^1(\mathcal{X}^\infty \times \Theta)\) (relative to the probability measure \(\Pi^*\) defined by \(\Pi^*(A \times B) = \int_B P_\theta(A) d\Pi(\theta)\) for measurable \(A \subset \mathcal{X}^\infty\) and \(B \subset \Theta\)), shows there is a measurable \(g : \mathcal{X}_n \to [0,1]\) such that,

\[
\int P_\theta |V | X^n - g(X)| d\Pi(\theta | B) \to 0.
\]

So \(\int P_\theta g(X) d\Pi(\theta | B) = 0\), implying that \(g = 0, P_\theta\)-almost-surely for \(\Pi\)-almost-all \(\theta \in B\). Using martingale convergence again (now in \(L^\infty(\mathcal{X}^\infty \times \Theta)\)), conclude
\(\Pi(V|X^n) \to 0, P_\theta\)-almost-surely for \(\Pi\)-almost-all \(\theta \in B\), from which (iv) follows.

Choose \(\phi(X^n) = \Pi(V|X^n)\) to conclude that (i) follows from (iv).

The interpretation of this theorem is gratifying to supporters of the likelihood principle and pure Bayesians: distinctions between model subsets are Bayesian testable, if and only if, they are picked up by the posterior asymptotically, if and only if, there exists a pointwise test for \(B\) versus \(V\) that is \(\Pi\)-almost-surely consistent.

For a second, more frequentist way to illustrate how basic the existence of a Bayesian test sequences is, consider a parameter space \((\Theta, d)\) which is a metric space with fixed Borel prior \(\Pi\) and \(d\)-consistent estimators \(\hat{\theta}_n: \mathcal{X}_n \to \Theta\) for \(\theta\). Then for every \(\theta_0 \in \Theta\) and \(\varepsilon > 0\), there exists a pointwise test sequence (and hence, by dominated convergence, also a Bayesian test sequence) for \(B = \{ \theta \in \Theta : d(\theta, \theta_0) < \frac{1}{2}\varepsilon\}\) versus \(V = \{ \theta \in \Theta : d(\theta, \theta_0) > \varepsilon\}\). This approach is followed in example 7.19 on random walks, see the definition of the test following inequality (7.27).

A third perspective on the existence of Bayesian tests arises from Doob’s argument. From our present perspective, we note that theorem 8.7 implies a proof of Doob’s consistency theorem through the following existence result on Bayesian test sequences. (Note: here and elsewhere in \(i.i.d\). setting, the parameter space \(\Theta\) is the single-observation model \(\mathcal{P}\), \(\theta\) is the single-observation distribution \(P\) and \(\theta \mapsto P_{\theta,n}\) is \(P \mapsto P^n\).)

**Proposition 7.2.** Consider a model \(\mathcal{P}\) of single-observation distributions \(P\) for \(i.i.d\). data \((X_1, X_2, \ldots, X_n) \sim P^n, (n \geq 1)\). Assume that \(\mathcal{P}\) is a Polish space with Borel prior \(\Pi\). For any Borel set \(V\) there is a Bayesian test sequence for \(V\) versus \(\mathcal{P} \setminus V\) under \(\Pi\).

**Proof.** We prove this theorem in chapter 8. (See [162], section 17.1, proposition 1 with the indicator for \(V\); see also [47].)

Doob’s theorem is recovered when we let \(V\) be the complement of any open neighbourhood \(U\) of \(P_0\). Comparing with conditions for the existence of uniform tests, Bayesian tests are quite abundant: whereas uniform testing relies on the minimax theorem (forcing convexity, compactness and continuity requirements into the picture), Bayesian tests exist quite generally (at least, for Polish parameters with \(i.i.d\). data).

The fourth perspective on the existence of Bayesian tests concerns a direct way to construct a Bayesian test sequence of optimal power, based on the fact that we are really only testing barycentres against each other: let priors \((\Pi_n)\) and \(\mathcal{P}\)-measurable model subsets \(B_n, V_n\) be given. For given tests \((\phi_n)\) and power sequence \(a_n\), write (7.1) as follows:

\[
\Pi_n(B_n) P_n^{\Pi_n|B_n} \phi_n(X^n) + \Pi_n(V_n) P_n^{\Pi_n|V_n}(1 - \phi_n(X^n)) = o(a_n),
\]

and note that what is required here, is a (weighted) test of \((P_n^{\Pi_n|B_n})\) versus \((P_n^{\Pi_n|V_n})\). The likelihood-ratio test of example 2.15 (denote the density for \(P_n^{\Pi_n|B_n}\) with respect to \(\mu_n = P_n^{\Pi_n|B_n} + P_n^{\Pi_n|V_n}\) by \(p_{B_n,n}\), and similar for \(P_n^{\Pi_n|V_n}\),)
\[ \phi_n(X^n) = \{ (\Pi_n(V_n) p_{\Pi_n}(X^n))^\alpha \geq \Pi_n(B_n) p_{\Pi_n}(X^n) \} \]

is optimal and has power \([|\Pi_n(B_n)| p_{\Pi_n}^B_n \wedge |\Pi_n(V_n)| p_{\Pi_n}^V_n|\]. This proves the following useful proposition that re-expresses power in terms of the relevant Hellinger transform (see Remark 1 of section 16.4 in Le Cam (1986) [162]).

**Proposition 7.3.** Fix \( n \geq 1 \) and let a prior \( (\Pi_n) \) and measurable model subsets \( B_n, V_n \) be given. There exists a test function \( \phi_n : \mathcal{F}_n \to [0, 1] \) such that,

\[
\int_{B_n} P_{\theta, n} \phi_n d\Pi_n(\theta) + \int_{V_n} P_{\theta, n}(1 - \phi_n) d\Pi_n(\theta) 
\leq \int \left( \Pi_n(B_n) p_{\Pi_n}(x) \right)^\alpha \left( \Pi_n(V_n) p_{\Pi_n}(x) \right)^{1-\alpha} d\mu_n(x), \tag{7.4}
\]

for any \( 0 \leq \alpha \leq 1 \).

Proposition 7.3 generalises proposition 7.2 and makes Bayesian tests available with a sharp bound on the power under fully general conditions. For the connection with minimax tests, we note the following. If \( \{P_{\theta, n} : \theta \in B_n\} \) and \( \{P_{\theta, n} : \theta \in V_n\} \) are convex sets (and the \( \Pi_n \) are Radon measures, e.g. in Polish parameter spaces), then,

\[
H(P_{\Pi_n}^B_n, P_{\Pi_n}^V_n) \geq \inf \{ H(P_{\theta, n}, P_{\theta', n}) : \theta \in B_n, \theta' \in V_n \}. 
\]

Combination with (7.4) for \( \alpha = 1/2 \), implies that the minimax upper bound in \( i.i.d. \) cases, c.f. proposition 6.3, remains valid:

\[
\int_{B_n} P_{\theta, n} \phi_n d\Pi_n(P) + \int_{V_n} Q_{\theta, n}(1 - \phi_n) d\Pi_n(Q) \leq \sqrt{|\Pi_n(B_n)| \Pi_n(V_n)} e^{-\frac{\alpha_2}{2}}, \tag{7.5}
\]

where \( \alpha_n = \inf \{ H(P, Q) : P \in B_n, Q \in V_n \} \). Given \( a_n \downarrow 0 \), any pointwise test \( \phi_n \) that satisfies (7.1) for all probability measures \( \Pi_n \) on \( \Theta \), is a (weighted) minimax test for \( B_n \) versus \( V_n \) of power \( a_n \).

Note that the above enhances the role that the prior plays in the frequentist discussion of the asymptotic behaviour of the posterior: the prior is not only important in requirements like (6.15), but can also be of influence in the testing condition: where testing power is relatively weak, prior mass should be scarce to compensate and where testing power is strong, prior mass can be plentiful. To make use of this, one imposes *upper bounds on prior mass* in certain hard-to-test subsets of the model (as opposed to *lower bounds* like (6.15)). See example 7.19 on random-walk data.

In the Hellinger-geometric view, the prior determines whether the local prior predictive distributions \( \pi_{\Pi_n}^B_n \) and \( \pi_{\Pi_n}^V_n \) lie close together or not in Hellinger distance, and thus to the r.h.s. of (7.4) for \( \alpha = 1/2 \).
7.1.3 Le Cam’s inequality

Referring to the argument following proposition 7.1, one way of guaranteeing that the expectations of $X_n \mapsto \Pi(V|X^n)$ under $P_{\Pi|B_n}$ approximate those under $P_{0,n}$ is to choose $B_n = \{ \theta \in \Theta : \| P_{\theta,n} - P_{0,n} \| \leq \delta_n \}$, for some sequence $\delta_n \to 0$, because in that case, $| P_{0,n} \psi - P_{\Pi|B_n} \psi | \leq \| P_{0,n} - P_{\Pi|B_n} \| \leq \delta_n$, for any random variable $\psi : \mathcal{X} \to [0,1]$. Without fixing the definition of the sets $B_n$, one may use this step to specify inequality (7.2) further:

$$P_{0,n} \Pi(V_n|X) \leq \left\| P_{0,n} - P_{\Pi|B_n} \right\| + \int P_{0,n} \phi_n d \Pi_n(\theta|B_n) + \frac{\Pi_n(V_n)}{\Pi_n(B_n)} \int P_{0,n} (1 - \phi_n) d \Pi_n(\theta|V_n), \quad (7.6)$$

for $B_n$ and $V_n$ such that $\Pi_n(B_n) > 0$ and $\Pi_n(V_n) > 0$. Le Cam’s inequality (7.6) is used, for example, in the proof of the Bernstein-von Mises theorem, see lemma 2 in section 8.4 of [166]. A less successful application pertains to non-parametric posterior rates of convergence for i.i.d. data, in an unpublished paper [161]. Rates of convergence obtained in this way are suboptimal: Le Cam qualifies the first term on the right-hand side of (7.6) as a “considerable nuisance” and concludes that “it is unclear at the time of this writing what general features, besides the metric structure, could be used to refine the results”, (see [162], end of section 16.6). In [228], Le Cam relates the posterior question to dimensionality restrictions [160, 204, 101] and reiterates, “And for Bayes risk, I know that just the metric structure does not catch everything, but I don’t know what else to look at, except calculations.”

7.2 Remote contiguity

Le Cam’s notion of contiguity (see Le Cam (1960) [157]) describes an asymptotic version of absolute continuity, applicable to sequences of probability measures in a limiting sense. A condensed overview of the most basic characterizations of contiguity and some essential references are found in appendix C.3. In this section we weaken the property of contiguity in a way that is suitable to promote $\Pi$-almost-everywhere Bayesian limits to frequentist limits that hold everywhere.

7.2.1 Definition and criteria for remote contiguity

The notion of “domination” left undefined in the argument following proposition 7.1 is made rigorous here.

**Definition 7.2.** Given measurable spaces $(\mathcal{X}_n, \mathcal{F}_n)$, $n \geq 1$ with two sequences $(P_n)$ and $(Q_n)$ of probability measures and a sequence $\rho_n \downarrow 0$, we say that $Q_n$ is $\rho_n$-
remotely contiguous with respect to \( P_n \), notation \( Q_n \ll P_n^{-1} P_0 \), if,

\[
P_n \phi_n(X^n) = o(P_n) \quad \Rightarrow \quad Q_n \phi_n(X^n) = o(1),
\]

for every sequence of \( \mathcal{B}_n \)-measurable \( \phi_n : \mathcal{X}_n \to [0,1] \).

Note that for a sequence \( \{Q_n\} \) that is \( a_n \)-remotely contiguous with respect to \( \{P_n\} \), there exists no test sequence that distinguishes between \( P_n \) and \( Q_n \) with power of order \( o(a_n) \). Note also that given two sequences \( \{P_n\} \) and \( \{Q_n\} \), contiguity \( P_n \ll Q_n \) is equivalent to remote contiguity \( P_n \ll a_n^{-1} Q_n \) for all \( a_n \downarrow 0 \). Given sequences \( a_n, b_n \downarrow 0 \) with \( a_n = O(b_n) \), \( b_n \)-remote contiguity implies \( a_n \)-remote contiguity of \( \{P_n\} \) with respect to \( \{Q_n\} \).

**Example 7.1.** Let \( \mathcal{P} \) be a model for the distribution of a single observation in i.i.d. samples \( X^n = (X_1, \ldots, X_n) \). Let \( P_0, P \) and \( \varepsilon > 0 \) be such that \( -P_0 \log(dP/dP_0) < \varepsilon^2 \).

The law of large numbers implies that for large enough \( n \),

\[
\frac{dP^n}{dP_0}(X^n) \geq e^{-\frac{\varepsilon^2}{2n}},
\]

with \( P_0^n \)-probability one. Consequently, for large enough \( n \) and for any \( \mathcal{B}_n \)-measurable sequence \( \psi_n : \mathcal{X}_n \to [0,1] \),

\[
P^n \psi_n \geq e^{-\frac{\varepsilon^2}{2n} P_0^n \psi_n}.
\]

Therefore, if \( P^n \phi_n = o(\exp\left(-\frac{1}{2}n\varepsilon^2\right)) \) then \( P_0^n \phi_n = o(1) \). Conclude that for every \( \varepsilon > 0 \), the Kullback-Leibler neighbourhood \( \{P : -P_0 \log(dP/dP_0) < \varepsilon^2 \} \) consists of model distributions for which the sequence \( \{P_0^n\} \) of product distributions are \( \exp\left(-\frac{1}{2}n\varepsilon^2\right) \)-remotely contiguous with respect to \( \{P^n\} \).

Criteria for remote contiguity are given in the lemma below; note that, here, we give sufficient conditions, rather than necessary and sufficient, as in Le Cam’s First Lemma, lemma C.4 or . (For the precise, \( Q_n \)-almost-sure definition of \( (dP^n/dQ_n)^{-1} \), see appendix A.)

**Lemma 7.2.** Given probability measures \( \{P_n\}, \{Q_n\} \) on measurable spaces \( (\mathcal{X}_n, \mathcal{B}_n) \) and \( a_n \downarrow 0 \), \( Q_n \ll a_n^{-1} P_n \), if any of the following hold:

(i) for any bounded, \( \mathcal{B}_n \)-measurable \( T_n : \mathcal{X}_n \to [0,1] \), \( a_n^{-1} T_n \xrightarrow{P_n} 0 \) implies \( T_n \xrightarrow{Q_n} 0 \),

(ii) for any \( \varepsilon > 0 \), there is a \( \delta > 0 \) such that \( Q_n(dP_n/dQ_n < \delta a_n) < \varepsilon \), for large enough \( n \),

(iii) there is a \( b > 0 \) such that \( \lim\inf_n b a_n^{-1} P_n(dQ_n/dP_n > b a_n^{-1}) = 1 \),

(iv) for any \( \varepsilon > 0 \), there is a constant \( c > 0 \) such that \( \|Q_n - Q_n \wedge c a_n^{-1} P_n\| < \varepsilon \), for large enough \( n \),

(v) under \( Q_n \) every subsequence of \( \{a_n(dP_n/dQ_n)^{-1}\} \) has a weakly convergent subsequence.

**Remark 7.1.** The proof of this lemma actually shows that (i) or (iv) implies remote contiguity; that (ii) or (iii) implies (iv) and that (v) is equivalent to (ii).
7.2 Remote contiguity

Proof. Assume (i). Let \( \phi_n: \mathcal{X}_n \rightarrow [0, 1] \) be given and assume that \( P_n \phi_n = o(a_n) \). By Markov’s inequality, for every \( \varepsilon > 0 \), \( P_n (a_n^{-1} \phi_n > \varepsilon) = o(1) \). By assumption, it now follows that \( \phi_n \overset{P_n}{\rightarrow} 0 \). Because 0 \( \leq \phi_n \leq 1 \) the latter conclusion is equivalent to \( Q_n \phi_n = o(1) \). Conclude that \( Q_n < a_n^{-1} P_n \). Next, assume (iv). Let \( \varepsilon > 0 \) and \( \phi_n: \mathcal{X}_n \rightarrow [0, 1] \) be given. There exist \( c > 0 \) and \( N \geq 1 \) such that for all \( n \geq N \),

\[
Q_n \phi_n < c a_n^{-1} P_n \phi_n + \frac{\varepsilon}{2}.
\]

If we assume that \( P_n \phi_n = o(a_n) \) then there is a \( N' \geq N \) such that \( c a_n^{-1} P_n \phi_n < \varepsilon / 2 \) for all \( n \geq N' \). Consequently, for every \( \varepsilon > 0 \), there exists an \( N' \geq 1 \) such that \( Q_n \phi_n < \varepsilon \) for all \( n \geq N' \). Conclude that \( Q_n < a_n^{-1} P_n \). To show that (iii) \( \Rightarrow \) (iv), let \( \mu_n = P_n + Q_n \) and denote \( \mu_n \)-densities for \( P_n, Q_n \) by \( p_n, q_n: \mathcal{X}_n \rightarrow \mathbb{R} \). Then, for any \( n \geq 1, c > 0 \),

\[
\| Q_n - Q_n \wedge c a_n^{-1} P_n \| = \sup_{A \in \mathcal{B}_n} \left( \int_A q_n \, d\mu_n - \int_A q_n \, d\mu_n \wedge \int_A c a_n^{-1} p_n \, d\mu_n \right)
\]

\[
\leq \sup_{A \in \mathcal{B}_n} \int_A (q_n - q_n \wedge c a_n^{-1} p_n) \, d\mu_n
\]

\[
= \int 1\{q_n > c a_n^{-1} p_n\} (q_n - c a_n^{-1} p_n) \, d\mu_n.
\]

(7.10)

Note that the right-hand side of (7.10) is bounded above by \( Q_n (dP_n / dQ_n < c^{-1} a_n) \),

To show that (iii) \( \Rightarrow \) (iv), it is noted that, for all \( c > 0 \) and \( n \geq 1 \),

\[
0 \leq \int c a_n^{-1} P_n (q_n > c a_n^{-1} p_n) \leq Q_n (q_n > c a_n^{-1} p_n) \leq 1,
\]

so (7.10) goes to zero if \( \liminf_{n \to \infty} c a_n^{-1} P_n (dQ_n / dP_n > c a_n^{-1}) = 1 \). To prove that (v) \( \Leftrightarrow \) (ii), note that Prohorov’s theorem says that weak convergence of a subsequence within any subsequence of \( a_n (dP_n / dQ_n)^{-1} \) under \( Q_n \) (see appendix A) is equivalent to the asymptotic tightness of \( (a_n (dP_n / dQ_n)^{-1}) : n \geq 1 \) under \( Q_n \), i.e. for every \( \varepsilon > 0 \) there exists an \( M > 0 \) such that \( Q_n (a_n (dP_n / dQ_n)^{-1} > M) < \varepsilon \) for all \( n \geq 1 \). This is equivalent to (ii).

To conclude this section, we specify the definition of remote contiguity slightly further.

**Definition 7.3.** Given measurable spaces \( (\mathcal{X}_n, \mathcal{B}_n), (n \geq 1) \) with two sequences \( (P_n) \) and \( (Q_n) \) of probability measures and sequences \( \rho_n, \sigma_n > 0, \rho_n, \sigma_n \to 0 \), we say that \( Q_n \) is \( \rho_n \)-to-\( \sigma_n \) remotely contiguous with respect to \( P_n \), notation \( \sigma_n^{-1} Q_n < \rho_n^{-1} P_n \),

\[
P_n \phi_n (X^n) = o(\rho_n) \quad \Leftrightarrow \quad Q_n \phi_n (X^n) = o(\sigma_n),
\]

for every sequence of \( \mathcal{B}_n \)-measurable \( \phi_n: \mathcal{X}_n \to [0, 1] \).

Like definition 7.2, definition 7.3 allows for reformulation similar to lemma 7.2, e.g. if for some sequences \( \rho_n, \sigma_n \) like in definition 7.3,

\[
\| Q_n - Q_n \wedge \sigma_n \rho_n^{-1} P_n \| = o(\sigma_n),
\]
then $\sigma_n^{-1}Q_n \prec P_n^{-1}P_n$. We leave the formulation of other sufficient conditions to the reader.

Example 7.2. Inequality (7.9) in example 7.1 implies that $b_n^{-1}P_0^\alpha \prec a_n^{-1}P_n$, for any $a_n \leq \exp(-n\alpha^2)$ with $\alpha^2 > \frac{1}{2}e^2$ and $b_n = \exp(-n(\alpha^2 - \frac{1}{2}e^2))$. It is noted that this implies that $\phi_n(X^n) Q_n \overset{Q_n \text{s.a.}}{\longrightarrow} 0$ for any $\phi_n : \mathcal{X}_n \rightarrow [0, 1]$ such that $P_n \phi_n(X^n) = o(\rho_n)$ (more generally, this holds whenever $\sum_n \sigma_n < \infty$, as a consequence of the first Borel-Cantelli lemma).

7.3 Remote contiguity for Bayesian limits

The relevant applications in the context of Bayesian limit theorems concern remote contiguity of the sequence of true distributions $P_{\theta_0,n}$ with respect to local prior predictive distributions $P^{\Pi_n|B_n}$, where the sets $B_n \subset \Theta$ are such that,

$$P_{\theta_0,n} \prec a_n^{-1}P^{\Pi_n|B_n}, \tag{7.11}$$

for some rate $a_n \downarrow 0$. In the case of i.i.d. data, Barron [9] introduces strong and weak notions of merging of $P_{\theta_0,n}$ with (non-local) prior predictive distributions $P^{\Pi_n}$. The weak version imposes condition (ii) of lemma 7.2 for all exponential rates simultaneously. Strong merging (or matching [8]) coincides with Schwartz’s almost-sure limit, while Weak merging (and weak matching) are viewed as limits in probability.

By contrast, if we have a specific rate $a_n$ in mind, the relevant stochastic mode of convergence for remote contiguity is not almost-sure convergence or even convergence in probability, but convergence with respect to Prokhorov’s weak topology: namely, according to lemma 7.2-(v), (7.11) holds if inverse likelihood ratios $Z_n$ have a weak limit $Z$ when re-scaled by $a_n$,

$$Z_n = \left(\frac{dP^{\Pi_n|B_n}}{dP_{\theta_0,n}}\right)^{-1}(X^n), \quad a_n Z_n P_{\theta_0,n} \overset{w.}{\longrightarrow} Z.$$

But condition (7.11) can also be written out, for example to the requirement that for some constant $\delta > 0$,

$$P_{\theta_0,n} \left( \int \frac{dP_{\theta,n}}{dP_{\theta_0,n}}(X^n) \ d\Pi_n(\theta|B_n) < \delta a_n \right) \rightarrow 0,$$

with the help of lemma 7.2-(ii). This allows us to reformulate 6.7 as follows.

Proposition 7.4. Consider a model $\mathcal{P}$ of single-observation distributions $P$ for i.i.d. data $(X_1, X_2, \ldots, X_n) \sim P^n$, $(n \geq 1)$, with priors $(\Pi_n)$. Let $\varepsilon_n > 0$, $\varepsilon_n \downarrow 0$ and $P_0 \in \mathcal{P}$ be given and let $B_n = B(\varepsilon_n; P_0)$ be defined as in (6.23). Assuming $\Pi_n(B_n) > 0$, we have,

$$P_0^\alpha \prec e^{-n\varepsilon_n^2(1+\delta)} P^{\Pi_n|B_n}$$
The next proposition should be viewed in light of Le Cam and Yang (1988) [164], which considers properties like contiguity, convergence of experiments and local asymptotic normality in situations of statistical information loss. To make the present case compatible, we think of (remote) contiguity for probability measures that arise as marginals for the data \( X^n \) when information concerning the (Bayesian random) parameter \( \theta \) is unavailable.

**Proposition 7.5.** Let \( \theta_0 \in \Theta \) and a prior \( \Pi : \mathcal{G} \to [0,1] \) be given. Let \( B \) be a measurable subset of \( \Theta \) such that \( \Pi(B) > 0 \). Assume that for some \( a_n \downarrow 0 \), the family, 

\[
\left\{ a_n \left( \frac{dP_{\theta,n}}{dP_{\theta_0,n}} \right)^{-1}(X^n) : \theta \in B, n \geq 1 \right\},
\]

is uniformly tight under \( P_{\theta_0,n} \). Then \( P_{\theta_0,n} \ll a_n^{-1} P_{\theta,n}^{\Pi|B} \).

**Proof.** For every \( \varepsilon > 0 \), there exists a constant \( \delta > 0 \) such that, 

\[
P_{\theta_0,n} \left( a_n \left( \frac{dP_{\theta,n}}{dP_{\theta_0,n}} \right)^{-1}(X^n) > \frac{1}{\delta} \right) < \varepsilon,
\]

for all \( \theta \in B, n \geq 1 \). For this choice of \( \delta \), condition (ii) of lemma 7.2 is satisfied for all \( \theta \in B \) simultaneously, and c.f. the proof of said lemma, for given \( \varepsilon > 0 \), there exists a \( c > 0 \) such that, 

\[
\|P_{\theta_0,n} - P_{\theta_0,n} \wedge c a_n^{-1} P_{\theta,n}\| < \varepsilon,
\]

(7.12) for all \( \theta \in B, n \geq 1 \). Now note that for any \( A \in \mathcal{B}_n \), 

\[
0 \leq P_{\theta_0,n}(A) - P_{\theta_0,n}(A) \wedge c a_n^{-1} P_{\theta,n}^{\Pi|B}(A)
\]

\[
\leq \int (P_{\theta_0,n}(A) - P_{\theta_0,n}(A) \wedge c a_n^{-1} P_{\theta,n}(A)) d\Pi(\theta|B).
\]

Taking the supremum with respect to \( A \), we find the following inequality in terms of total variational norms, 

\[
\left\|P_{\theta_0,n} - P_{\theta_0,n} \wedge c a_n^{-1} P_{\theta,n}^{\Pi|B}\right\| \leq \int \left\|P_{\theta_0,n} - P_{\theta_0,n} \wedge c a_n^{-1} P_{\theta,n}\right\| d\Pi(\theta|B).
\]

Since the total-variational norm is bounded and \( \Pi(\cdot|B) \) is a probability measure, Fatou’s lemma says that, 

\[
\limsup_{n \to \infty} \left\|P_{\theta_0,n} - P_{\theta_0,n} \wedge c a_n^{-1} P_{\theta,n}^{\Pi|B}\right\| \leq \int \limsup_{n \to \infty} \left\|P_{\theta_0,n} - P_{\theta_0,n} \wedge c a_n^{-1} P_{\theta,n}\right\| d\Pi(\theta|B),
\]
and the r.h.s. equals zero c.f. (7.12). According to condition (iv) of lemma 7.2 this implies the assertion.

To re-establish contact with the notion of merging, note the following. If remote contiguity of the type (7.11) can be achieved for a sequence of subsets \((B_n)\), then it also holds for any sequence of sets (e.g. all equal to \(\Theta\), in Barron’s case) that contain the \(B_n\) but at a rate that differs proportionally to the fraction of prior masses.

**Lemma 7.3.** For all \(n \geq 1\), let \(B_n \subset \Theta\) be such that \(\Pi_n(B_n) > 0\) and \(C_n\) such that \(B_n \subset C_n\) with \(c_n = \Pi_n(B_n)/\Pi_n(C_n) \downarrow 0\), then,

\[
P_n^{\Pi_n|B_n} \prec c_n^{-1} P_n^{\Pi_n|C_n}.
\]

Also, if for some sequence \((P_n)\), \(P_n \prec a_n^{-1} P_n^{\Pi_n|B_n}\), then \(P_n \prec a_n^{-1} c_n^{-1} P_n^{\Pi_n|C_n}\).

**Proof.** Fix \(n \geq 1\). Because \(B_n \subset C_n\), for every \(A \in \mathcal{B}_n\), we have,

\[
\int_{B_n} P_{\theta_n}(A) d\Pi(\theta) \leq \int_{C_n} P_{\theta_n}(A) d\Pi(\theta),
\]

so \(P_n^{\Pi_n|B_n}(A) \leq \Pi_n(C_n)/\Pi_n(B_n) P_n^{\Pi_n|C_n}(A)\). So if for some sequence \(\phi_n : \mathcal{X}_n \to [0,1]\), we have \(P_n^{\Pi_n|C_n} \phi_n(X^n) = o(\Pi_n(B_n)/\Pi_n(C_n))\), then the \(P_n^{\Pi_n|B_n}\)-expectations of \(\phi_n(X^n)\) are \(o(1)\), proving the first claim. If \(P_n^{\Pi_n|C_n} \phi_n(X^n) = o(a_n \Pi_n(B_n)/\Pi_n(C_n))\), then \(P_n^{\Pi_n|B_n} \phi_n(X^n) = o(a_n)\) and, hence, \(P_n \phi_n(X^n) = o(1)\).

So when considering possible choices for the sequence \((B_n)\), smaller choices lead to rates \(a_n\) that go to zero more slowly, rendering (7.7) applicable to more sequences of test functions. This advantage is to be balanced against later requirements that \(\Pi_n(B_n)\) may not decrease too fast.

### 7.3.1 Remote contiguity, examples in regression

To better understand the counterexamples of subsection 6.7.1, notice the high sensitivity of remote contiguity to the existence of subsets of the sample spaces assigned probability zero under some model distributions, while the true probability is non-zero. More generally, remote contiguity is sensitive to subsets \(E_n\) assigned fast-decreasing probabilities under local prior predictive distributions \(P_n^{\Pi_n|E_n}(E_n)\), while the probabilities \(P_{\theta_n}(E_n)\) remain high, which is what definition 7.2 expresses. The rate \(a_n \downarrow 0\) helps to control the likelihood ratio (compare to the unscaled limits of likelihood ratios that play a central role in the theory of convergence of statistical experiments [162, 166, 214]), conceivably enough to force uniform tightness in many non-parametric situations.

To compare contiguity and its remote analogue in the context of (parametric and non-parametric) Bayesian regression, consider the following example.
Example 7.3. Let $\mathcal{F}$ denote a class of functions $\mathbb{R} \to \mathbb{R}$. We consider samples $X^n = ((X_1, Y_1), \ldots, (X_n, Y_n))$, $(n \geq 1)$ of points in $\mathbb{R}^2$, assumed to be related through,

$$Y_i = f_0(X_i) + e_i,$$

for some unknown $f_0 \in \mathcal{F}$, where the errors are i.i.d. standard normal $e_1, \ldots, e_n \sim N(0, 1)^n$ and independent of the i.i.d. covariates $X_1, \ldots, X_n \sim P^n$, for some ancillary distribution $P$ on $\mathbb{R}$. Assume that $\mathcal{F} \subset L^2(P)$ and that $P f_0(X) = 0$ for all $f \in \mathcal{F}$. We use the $L^2$-norm $\|f\|_{L^2}^2 = \int f^2 \, dP$ to define a metric $d$ on $\mathcal{F}$, $d(f, g) = \|f - g\|_{L^2}$. Given a parameter $f \in \mathcal{F}$, denote the sample distributions as $P_{f, n}$. We distinguish two cases: (a) the case of linear regression, where $\mathcal{F} = \{f_0 : \mathbb{R} \to \mathbb{R} : \theta \in \Theta\}$, where $\theta = (a, b) \in \Theta = \mathbb{R}^2$ and $f_0(x) = ax + b$; (b) the case of non-parametric regression, where we do not restrict $\mathcal{F}$ beforehand; and (c) a case where we replace the parameter $f$ by non-parametric point-estimators $f_n : \mathcal{X}_n \to L^2(P)$, like replacing a nuisance parameter by an estimate to obtain an approximate distribution (e.g. profile likelihood, see Murphy and van der Vaart (2000) [180]).

Let $\Pi$ be a Borel prior $\Pi$ on $\mathcal{F}$ and place remote contiguity in context by assuming, for the moment, that for some $\rho > 0$, there exist $0 < r < \rho$ and $\tau > 0$, as well as Bayesian tests $\phi_n$ for $B = \{f \in \mathcal{F} : \|f - f_0\|_{L^2} < r\}$ versus $V = \{f \in \mathcal{F} : \|f - f_0\|_{L^2} \geq \rho\}$ under $\Pi$ of power $a_n = \exp(-\frac{1}{2} n \tau^2)$. If this is the case, we may assume that $r < \frac{1}{2} \tau$ without loss of generality. Suppose also that $\Pi$ has a support in $L^2(P)$ that contains all of $\mathcal{F}$.

Example 7.4. Let us concentrate on case (b) first: a bit of manipulation casts the $a_n$-rescaled likelihood ratio for $f \in \mathcal{F}$ in the following form,

$$a_n^{-1} \frac{dP_{f, n}}{dP_{f_0, n}}(X^n) = a_n^{-1} \prod_{i=1}^n e^{-\frac{1}{2}(Y_i - f(X_i))^2} = e^{-\frac{1}{2} \sum_{i=1}^n (2e_i (f - f_0)(X_i) + (f - f_0)^2(X_i) - \tau^2)},$$

under $X^n \sim P_{f_0, n}$. The exponent is controlled by the law of large numbers,

$$\frac{1}{n} \sum_{i=1}^n (2e_i (f - f_0)(X_i) + (f - f_0)^2(X_i) - \tau^2) \overset{P_{f_0, n}}{\to} -\|f - f_0\|_{L^2}^2 - \tau^2.$$

Hence, for every $\epsilon > 0$ there exists an $N(f, \epsilon) \geq 1$ such that the exponent in (7.13) satisfies the upper bound,

$$\sum_{i=1}^n (2e_i (f - f_0)(X_i) + (f - f_0)^2(X_i) - \tau^2) \leq n(||f - f_0||_{L^2}^2 - \tau^2 + \epsilon^2),$$

for all $n \geq N(f, \epsilon)$. Since $\Pi(B) > 0$, we may condition $\Pi$ on $B$, choose $\epsilon = \frac{1}{2} \tau$ and use Fatou’s inequality to find that,

$$\liminf_{n \to \infty} e^{\frac{1}{2} \tau^2} \frac{dP_{f, n}}{dP_{f_0, n}}(X^n) \geq \liminf_{n \to \infty} e^{\frac{1}{2} \tau^2} = \infty,$$
\(P^\infty_{f_0}\)-almost-surely. Consequently, for any choice of \(\delta\),
\[
P_{f_0,n}\left(\frac{dP_n^{\|B}|B}}{dP_{f_0,n}}(X^n) < \delta e^{-\frac{1}{2}n\tau^2}\right) \to 0,
\]
and we conclude that \(P_{f_0,n} \ll e^{\frac{1}{2}n\tau^2} P_n^{\|B}|B}\).

Example 7.5. To analyse case (c) next, we consider rates \(a_n = \exp(-\frac{1}{2}n\tau_n^2)\) in remote-contiguity statement (7.13) with \(f = \hat{f}_n(X^n)\). Assume that the estimators \(\hat{f}_n\) are \(L^2(P)\)-consistent at rate \(\varepsilon_n\), i.e. \(\varepsilon_n^{-2}\|\hat{f}_n - f_0\|_{L_2} = O_P(1)\). Also assume that,
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\varepsilon_i(\hat{f}_n - f_0)(X_i)) = O_{P_{f_0,n}}(1),
\]
and,
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left((\hat{f}_n - f_0)^2(X_i) - \|\hat{f}_n - f_0\|_{L_2}^2\right) = O_{P_{f_0,n}}(1),
\]
See subsection 3.4.3 (and particularly 3.4.3.2) in [217]; through the use of maximal inequalities, the latter discusses in great detail how smoothness assumptions on the space \(\mathcal{F}\) allow verification of uniform versions of these convergence statements, which imply the versions above. For example, if \(\mathcal{F}\) is a bounded subset of \(C^\alpha([0,1]^d)\), the space of all \(\alpha\)-times differentiable functions on \([0,1]^d\) (with the Hölder norm), then the least-squares estimator \(\hat{f}_n\) converges to \(f_0\) at \(L^2\)-rate,
\[
\varepsilon_n = n^{-\frac{\alpha}{2d+\alpha}}.
\]
For any \(\tau\) such that \(\varepsilon_n = o(\tau_n)\), we then find that,
\[
P_{f_0,n} \ll e^{\frac{1}{2}n\tau^2} P_{\theta_0,n},
\]
demonstrating that remote contiguity also applies where the approximation of one sequence by another is required, with possible application outside of Bayesian limits.

Example 7.6. As for case (a), one has the choice of using a prior like above, but also to proceed differently: expression (7.13) can be written in terms of a local parameter \(h \in \mathbb{R}^k \) which, for given \(\theta_0\) and \(n \geq 1\), is related to \(\theta\) by \(\theta = \theta_0 + n^{-1/2}h\). For \(h \in \mathbb{R}^2\), we write \(P_{h,n} = P_{\theta_0 + n^{-1/2}h,n} = P_{\theta_0,n}\) and rewrite the likelihood ratio (7.13) as follows,
\[
\frac{dP_{h,n}}{dP_{\theta,n}}(X^n) = e^{\frac{1}{2} \sum_{i=1}^{n} \ell_{\theta_0}(X_i) - \frac{1}{2} h^T I_{\theta_0} h + R_n},
\]
where \(\ell_{\theta_0} : \mathbb{R}^2 \to \mathbb{R}^2 : (x,y) \to (y - a_0x - b_0)(x,1)\) is the score function for \(\theta, I_{\theta_0} = P_{\theta_0,1} \ell_{\theta_0} \ell_{\theta_0}^T\) is the Fisher information matrix and \(R_n \longrightarrow 0\). Assume that \(I_{\theta_0}\) is non-singular and note the central limit,
7.3 Remote contiguity for Bayesian limits

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell_{\theta_0}(X_i, Y_i) \xrightarrow{p_{\theta_0,n}} \mathcal{N}_2(0, I_{\theta_0}),
\]

which expresses local asymptotic normality of the model, c.f. definition 4.1, and implies that for any fixed \( h \in \mathbb{R}^2 \), \( p_{\theta,n} \prec p_{0,n} \), proving remote contiguity at any rate.

Clearly a proof of contiguity puts requirements on the likelihood of a relatively stringent nature compared to the requirements posed by remote contiguity. The LAN example relies on quite subtle argumentation that is natural in parametric context, but cannot be expected to generalise to the same powerful extent in non-parametric setting (notwithstanding successes in semi-parametric statistics). In non-parametric cases, a less delicate argument is required and remote contiguity appears to provide it in an accessible way (through relatively straightforward analysis of weak limits of \( a_n \)-re-scaled inverse likelihood ratios).

Remote contiguity also applies in more irregular situations: example 6.4 does not admit KL priors, but satisfies the requirement of remote contiguity. (Choose \( \eta \) equal to the uniform density for simplicity.)

**Example 7.7.** Consider again the model of example 6.4, where we observe an i.i.d. sample from the uniform distribution on \( [\theta, \theta + 1] \), for unknown \( \theta \in \mathbb{R} \). The model is parametrized in terms of distributions \( P_{\theta} \) with Lebesgue densities of the form \( p_{\theta}(x) = 1_{[\theta, \theta+1]}(x) \), for \( \theta \in \Theta = \mathbb{R} \). Pick a prior \( \Pi \) on \( \Theta \) with a continuous and strictly positive Lebesgue density \( \pi : \mathbb{R} \to \mathbb{R} \) and, for some rate \( \delta_n \downarrow 0 \), choose \( B_n = (\theta_0, \theta_0 + \delta_n) \). Note that for any \( \alpha > 0 \), there exists an \( N \geq 1 \) such that for all \( n \geq N \),

\[
(1 - \alpha) \pi(\theta_0) \delta_n \leq \Pi(B_n) \leq (1 + \alpha) \pi(\theta_0) \delta_n.
\]

Note that for any \( \theta \in B_n \) and \( X^n \sim P_{\theta_0}^n \),

\[
dP_{\theta_0}^n / dP_{\theta_0}^n(X^n) = 1 \{ X(1) > \theta \},
\]

and correspondingly,

\[
\frac{dP_{\Pi|B_n}^n(X^n)}{dP_{\theta_0}^n}(X^n) = \Pi_n(B_n)^{-1} \int_{\theta_0}^{\theta_0 + \delta_n} 1 \{ X(1) > \theta \} d\Pi(\theta)
\]

\[
\geq \frac{1 - \alpha}{1 + \alpha} \delta_n \wedge (X(1) - \theta_0),
\]

for large enough \( n \). As a consequence, for every \( \delta > 0 \) and all \( a_n \downarrow 0 \),

\[
P_{\theta_0}^n \left( \frac{dP_{\Pi|B_n}^n}{dP_{\theta_0}^n}(X^n) < \delta a_n \right) \leq P_{\theta_0}^n \left( \frac{\delta^{-1}_n(X(1) - \theta_0) < (1 + \alpha) \delta a_n}{\delta_n} \right),
\]

for large enough \( n \geq 1 \). Since \( n(X(1) - \theta_0) \) has an exponential weak limit under \( P_{\theta_0}^n \), we choose \( \delta_n = n^{-1} \), so that the r.h.s. in the above display goes to zero. So \( P_{\theta,n} \prec a_n^{-1} P_{\Pi|B_0}^n \), for any \( a_n \downarrow 0 \). Conclude that with these choices for \( \Pi \) and \( B_n \), (7.11) holds, for any \( a_n \).
7.4 Posterior concentration

In this section new frequentist theorems are formulated involving the convergence of posterior distributions. First we give a basic proof for posterior consistency assuming existence of suitable test sequences and remote contiguity of true distributions \( (P_{\theta_0,n}) \) with respect to local prior predictive distributions \( (\Pi_{\theta_0} \mid B_n) \). Then it is not difficult to extend the proof to the case of posterior rates of convergence in metric topologies. With the same methodology it is possible to address questions in Bayesian hypothesis testing and model selection: if a Bayesian test to distinguish between two hypotheses exists and remote contiguity applies, frequentist consistency of the Bayes Factor can be guaranteed. We conclude with a theorem that uses remote contiguity to describe a general relation that exists between credible sets and confidence sets, provided the prior induces remotely-contiguous local prior predictive distributions.

We start with posterior consistency, c.f. definition 6.4 and proposition 6.1. The formulation is has the generality of remark A.1 and the theorem applies to non-i.i.d. data, and with \( n \)-dependent models and priors.

**Theorem 7.2.** Assume that for all \( n \geq 1 \), the data \( X^n \sim P_{\theta_0,n} \) for some \( \theta_0 \in \Theta \). Fix a prior \( \Pi : \mathcal{G} \to [0,1] \) and assume that for given \( B, V \in \mathcal{G} \) with \( \Pi(B) > 0 \) and \( a_n \downarrow 0 \),

(i) there exist Bayesian tests \( \phi_n \) for \( B \) versus \( V \),

\[
\int_B P_{\theta_0,n} \phi_n d\Pi(\theta) + \int_V P_{\theta',n}(1 - \phi_n) d\Pi(\theta') = o(a_n), \tag{7.15}
\]

(ii) the sequence \( P_{\theta_0,n} \) satisfies \( P_{\theta_0,n} \sqsubset a_n^{-1} P_n^{\Pi \mid B} \).

Then \( \Pi(V \mid X^n) \xrightarrow{P_{\theta_0,n}} 0 \).

**Proof.** Choose \( B_n = B, V_n = V \) and use proposition 7.1 to see that \( P_n^{\Pi \mid B} \Pi(V \mid X^n) \) is upper bounded by \( \Pi(B)^{-1} \) times the l.h.s. of (7.15) and, hence, is of order \( o(a_n) \). Condition (ii) then implies that \( P_{\theta_0,n} \Pi(V \mid X^n) = o(1) \), which is equivalent to \( \Pi(V \mid X^n) \xrightarrow{P_{\theta_0,n}} 0 \) since \( 0 \leq \Pi(V \mid X^n) \leq 1, P_{\theta_0,n} \)-almost-surely, for all \( n \geq 1 \).

These conditions may be interpreted as follows: theorem 8.7 lends condition (i) a distinctly Bayesian interpretation: it requires a Bayesian test to set \( V \) apart from \( B \) with testing power \( a_n \). Lemma 7.1 translates this into the (still Bayesian) statement that the posteriors for \( V \) go to zero in \( P_n^{\Pi \mid B} \)-expectation. Condition (ii) is there to promote this Bayesian point to a frequentist one through (7.7).

One of the first questions we have, is how Freedman’s inconsistent posteriors relate to the above. Since test sequences of exponential power exist to separate complements of weak neighbourhoods, c.f. proposition A.1, Freedman’s inconsistencies must violate the requirement of remote contiguity in theorem 7.2.

**Example 7.8.** As noted already, the space \( \Lambda \) of examples 1.1, 2.5 and 6.3 is a Polish space; in particular \( \Lambda \) is metric and second countable, so the subspace \( \mathcal{N} \) contains...
a countable dense subset $D$. For $Q \in D$, let $V$ be the set of all prior probability measures on $\Lambda$ with finite support, of which one point is $Q$ and the remaining points lie in $\Lambda_0$. The proof of the theorem in [93] that asserts that the set of consistent pairs $(\Pi_0, \Pi)$ is of the first category in $\Lambda \times \pi(\Lambda)$ departs from the observation that if $\Pi_0$ lies in $\mathcal{N}$ and we use a prior from $V$, then,

$$\Pi(\{Q\} | X^n) \overset{P_0\text{-a.s.}}{\to} 1,$$

(in fact, as is shown below, with $P^n_{0\alpha}$-probability one there exists an $N \geq 1$ such that $\Pi(\{Q\} | X^n) = 1$ for all $n \geq N$). The proof continues to assert that $V$ lies dense in $\pi(\Lambda)$ and, through sequences of continuous extensions involving $D$, that posterior inconsistency for elements of $V$ implies posterior inconsistency for all $\Pi$ in $\pi(\Lambda)$ with the possible exception of a set of the first category.

From the present perspective it is interesting to view the inconsistency of elements of $V$ in light of the conditions of theorem 7.2. Define, for some bounded $f : \mathbb{N} \to \mathbb{R}$ and $\varepsilon > 0$, two subsets of $\Lambda$,

$$B = \{P : \|P f - P_0 f\| < \frac{1}{2} \varepsilon\}, \quad V = \{P : \|P f - P_0 f\| \geq \varepsilon\}.$$  

Proposition A.1 asserts the existence of a uniform test sequence for $B$ versus $V$ of exponential power. With regard to remote contiguity, for an element $\Pi$ of $V$ with support of order $M + 1$, write,

$$\Pi = \beta \delta_Q + \sum_{m=1}^{M} \alpha_m \delta_{P_m},$$

where $\beta + \sum_{m} \alpha_m = 1$ and $P_m \in \Lambda_0$ ($1 \leq m \leq M$). Without loss of generality, assume that $\varepsilon$ and $f$ are such that $Q$ does not lie in $B$. Consider,

$$\frac{dP_n^{\Pi|B}}{dP_0^n}(X^n) = \frac{1}{\Pi(B)} \int_B \frac{dP^n}{dP_0^n}(X^n) d\Pi(P) \leq \frac{1}{\Pi(B)} \sum_{m=1}^{M} \alpha_m \frac{dP_m^n}{dP_0^n}(X^n).$$

For every $1 \leq m \leq M$, there exists a $k(m)$ such that $P_m(X = k(m)) = 0$, and the probability of the event $E_m$ that none of the $X_1, \ldots, X_n$ equal $k(m)$ is $(1 - P_0(X = k(m)))^n$. Note that $E_m$ is also the event that $dP_m^n / dP_0^n(X^n) > 0$.

Hence for every $1 \leq m \leq M$ and all $X$ in an event of $P_0^n$-probability one, there exists an $N_m \geq 1$ such that $dP_m^n / dP_0^n(X^n) = 0$ for all $n \geq N_m$. Consequently, for all $X$ in an event of $P_0^n$-probability one, there exists an $N \geq 1$ such that $dP_n^{\Pi|B} / dP_0^n(X^n) = 0$ for all $n \geq N$. Therefore, condition (ii) of lemma 7.2 is not satisfied for any sequence $\alpha_n \downarrow 0$. A direct proof that (7.7) does not hold for any $\alpha_n$ is also possible: given the prior $\Pi \in V$, define,

$$\phi_n(X^n) = \prod_{m=1}^{M} \prod_{1 \leq i \leq k} 1_{\{X_i = k(m)\}}.$$
Then the expectation of \( \phi_n \) with respect to the local prior predictive distribution equals zero, so \( \mu_{\Pi|B}^n \phi_n = o(a_n) \) for any \( a_n \downarrow 0 \). However, \( P_0^\Pi \phi_n(X^n) \rightarrow 1 \), so the prior \( \Pi \) does not give rise to a sequence of prior predictive distributions \( (P_\Pi^n) \) with respect to which \( (P_0^\Pi) \) is remotely contiguous, for any \( a_n \downarrow 0 \).

A proof of a theorem very close to Schwartz’s theorem is now possible. Consider condition (i) of theorem 6.5: a well-known argument based on Hoeffding’s inequality guarantees the existence of a uniform test sequence of exponential power whenever a uniform test sequence exists, so Schwartz equivalently assumes that there exists a \( D > 0 \) such that,

\[
P_0^n \phi_n + \sup_{Q \in \mathcal{P} \setminus U} Q^n(1 - \phi_n) = o(e^{-nD}).
\]

(7.16)

We vary slightly and assume the existence of a Bayesian test sequence of exponential power. In the following theorem, let \( \mathcal{P} \) denote a Hausdorff space of single-observation distributions on \( (X, B) \) with Borel prior \( \Pi \).

**Corollary 7.1.** For all \( n \geq 1 \), let \( (X_1, X_2, \ldots, X_n) \sim P_0^n \) for some \( P_0 \in \mathcal{P} \). Let \( U \) denote an open neighbourhood of \( P_0 \) and define \( K(\varepsilon) = \{ \theta \in \mathcal{P} : -P_0 \log(dP_\theta/dP_0) < \varepsilon^2 \} \).

If,

(i) there exist \( \varepsilon > 0, D > 0 \) and a sequence of measurable \( \psi_n : \mathcal{X}^n \rightarrow [0,1] \), such that,

\[
\int_{K(\varepsilon)} P^n \psi_n d\Pi(\theta) + \int_{\mathcal{P} \setminus U} Q^n(1 - \psi_n) d\Pi(\theta) = o(e^{-nD}),
\]

(ii) and \( \Pi(K(\varepsilon)) > 0 \) for all \( \varepsilon > 0 \),

then \( \Pi(U|X^n) \overset{P_0}{\rightarrow} 1 \).

**Proof.** A prior \( \Pi \) satisfying condition (ii) guarantees that \( P_0^n \ll P_\Pi^n \) for all \( n \geq 1 \), c.f. the remark preceding proposition A.2. Choose \( \varepsilon \) such that \( \varepsilon^2 < D \). Recall that for every \( P \in B(\varepsilon) \), the exponential lower bound (7.8) for likelihood ratios of \( dP^n_\theta / dP^n_0 \) exists. Hence the limes inferior of \( \exp(2n\varepsilon^2) \) is greater than or equal to one with \( P_0^n \)-probability one. Then, with the use of Fatou’s lemma and the assumption that \( \Pi(B(\varepsilon)) > 0 \),

\[
\liminf_{n \rightarrow \infty} \frac{e^{nD}}{\Pi(B)} \int_B \frac{dP^n_0(X^n)}{dP^n_{\theta_0}} d\Pi(\theta) \geq 1,
\]

with \( P_0^n \)-probability one, showing that sufficient condition (ii) of lemma 7.2 holds. Conclude that,

\[
P_0^n \ll e^{nD} P_{n|B},
\]

and use theorem 7.2 to see that \( \Pi(U|X^n) \overset{P_0}{\rightarrow} 1 \).

**Example 7.9.** As an example of the tests required under condition (i) of corollary 7.1, consider \( \mathcal{P} \) in the Hellinger topology, assuming totally-boundedness.
Let $U$ be the Hellinger-ball of radius $4\varepsilon$ around $P_{\theta_0}$ of example 7.11 and let $V$ be its complement. The Hellinger ball $B_H(\varepsilon)$ in equation (7.18) contains the set $K(\varepsilon)$. Alternatively we may consider the model in any of the weak topologies $\mathcal{T}_n$: let $\varepsilon > 0$ be given and let $U$ denote a weak neighbourhood of the form 

\[ \{P \in \mathcal{P} : |(P^n - P_0^n)f| \geq 2\varepsilon\} \]

for some bounded measurable $f : \mathcal{X}_n \to [0, 1]$, as in proposition A.1. The set $B$ of proposition A.1 contains a set $K(\delta)$, for some $\delta > 0$.

Both these applications were noted by Schwartz in [201].

### 7.4.1 Consistency of Bayesian point estimators

As we know from proposition 6.1, if $\Theta$ is a Hausdorff, completely regular space, the posterior is consistent at $\theta_0 \in \Theta$, if and only if,

\[ \int f(\theta) d\Pi(\theta | X^n) \xrightarrow{P_{\theta_0}} f(\theta_0), \]

for every bounded, continuous $f : \Theta \to \mathbb{R}$. Proposition 6.1 is used to prove consistency of frequentist point-estimators derived from the posterior, more generally than before in subsection 6.2.2.

**Example 7.10.** Consider a model $\mathcal{P}$ of single-observation distributions $P$ on $(\mathcal{X}, \mathcal{B})$ for i.i.d. data $(X_1, X_2, \ldots, X_n) \sim P^n$, $(n \geq 1)$. Assume that the true distribution of the data is $P_0 \in \mathcal{P}$ and that the model topology is Prohorov’s weak topology or stronger. Then for any bounded, continuous $g : \mathcal{X} \to \mathbb{R}$, the map,

\[ f : \mathcal{P} \to \mathbb{R} : P \mapsto |(P - P_0)g(X)|, \]

is continuous. Assuming that the posterior is weakly consistent at $P_0$,

\[ |P^n_1 g - P_0 g| \leq \int |(P - P_0)g| d\Pi(P | X^n) \xrightarrow{P_{\theta_0}} 0, \quad (7.17) \]

so posterior predictive distributions are consistent point estimators in Prohorov’s weak topology. Replacing the maps $g$ by bounded, measurable maps $\mathcal{X} \to \mathbb{R}$ and assuming posterior consistency in $\mathcal{T}_1$, one proves consistency of posterior predictive distributions in $\mathcal{T}_1$ in exactly the same way. Taking the supremum over measurable $g : \mathcal{X} \to [0, 1]$ in (7.17) and assuming that the posterior is consistent in the total variational topology, posterior predictive distributions are consistent in total variation as frequentist point estimators.
7.4.2 Posterior concentration and Hellinger entropy

Referring to the convexity requirement in proposition 6.3 on minimax tests, it is noted that questions concerning consistency require the existence of tests in which at least one of the two hypotheses is a non-convex set, typically the complement of a neighbourhood. Imposing the model \( \mathcal{P} \) to be of bounded entropy with respect to the Hellinger metric allows construction of such tests, based on the uniform tests of proposition 6.3. Below, we apply well-known constructions for the uniform tests in Schwartz’s theorem from the frequentist literature [160, 32, 33, 101] to the construction of Bayesian tests. Due to relations that exist between metrics for model parameters and the Hellinger metric in many examples and applications, the material covered here is widely applicable in (non-parametric) models for i.i.d. data.

Example 7.11. Consider a model \( \mathcal{P} \) of distributions \( P \) for i.i.d. data \( X^n \sim P^n \), \( n \geq 1 \) and, in addition, suppose that \( \mathcal{P} \) is totally bounded with respect to the Hellinger distance. Let \( P_0 \in \mathcal{P} \) and \( \varepsilon > 0 \) be given, denote \( V(\varepsilon) = \{ P \in \mathcal{P} : H(P_0, P) \geq 4\varepsilon \} \), \( B_\varepsilon = \{ P \in \mathcal{P} : H(P_0, P) < \varepsilon \} \). There exists an \( N(\varepsilon) \geq 1 \) and a cover of \( V(\varepsilon) \) by \( H \)-balls \( V_1, \ldots, V_{N(\varepsilon)} \) of radius \( \varepsilon \) and for any point \( Q \) in any \( V_i \) and any \( P \in B_\varepsilon \), \( H(Q, P) > 2\varepsilon \). According to proposition 7.3 with \( \alpha = 1/2 \) and (7.5), for each \( 1 \leq i \leq N(\varepsilon) \) there exists a Bayesian test sequence \( (\phi_{i,n}) \) for \( B_\varepsilon \) versus \( V_i \) of power (upper bounded by) \( \exp(-2n\varepsilon^2) \). Then, for any subset \( B' \subset B_\varepsilon \),

\[
P_n^{1/B'} \Pi(\cdot|X^n) \leq \sum_{i=1}^{N(\varepsilon)} P_n^{1/B'} \Pi(V_i|X^n)
\leq \frac{1}{\Pi(B')} \sum_{i=1}^{N(\varepsilon)} \left( \int_{B'} P^n \phi_n d\Pi(P) + \int_{V_i} P^n (1 - \phi_n) d\Pi(P) \right)
\leq \sum_{i=1}^{N(\varepsilon)} \frac{\Pi(V_i)}{\Pi(B')} \exp(-2n\varepsilon^2),
\tag{7.18}
\]

which is smaller than or equal to \( e^{-n\varepsilon^2} \) for large enough \( n \).

To balance entropy and prior mass differently in Hellinger separable models, Barron (1988) [9] and Barron et al. (1999) [14] formulate an alternative condition that is based on the Radon property that any prior on a Polish space has.

Example 7.12. Consider a model \( \mathcal{P} \) of distributions \( P \) for i.i.d. data \( X^n \sim P^n \), \( n \geq 1 \), with priors \( (\Pi_n) \). Assume that the model \( \mathcal{P} \) is Polish in the Hellinger topology. Let \( P_0 \in \mathcal{P} \) and \( \varepsilon > 0 \) be given; for a fixed \( M > 1 \), define \( V = \{ P \in \mathcal{P} : H(P_0, P) \geq M\varepsilon \} \), \( B_\varepsilon = \{ P \in \mathcal{P} : H(P_0, P) < \varepsilon \} \). For any sequence \( \delta_m \downarrow 0 \), there exist compacta \( K_m \subset \mathcal{P} \) such that \( \Pi(K_m) \geq 1 - \delta_m \) for all \( m \geq 1 \). For each \( m \geq 1 \), \( K_m \) is Hellinger totally bounded so there exists a Bayesian test sequence \( \phi_{m,n} \) for \( B_\varepsilon \cap K_m \) versus \( V(\varepsilon) \cap K_m \). Since,
7.4 Posterior concentration

\[
\int_{B_H} P^\rho \phi_n d\Pi(P) + \int_V Q^\rho (1 - \phi_n) d\Pi(Q) \\
\leq \int_{B_H \cap K_m} P^\rho \phi_{m,n} d\Pi(P) + \int_{V \cap K_m} Q^\rho (1 - \phi_{m,n}) d\Pi(Q) + \delta_m,
\]

and all three terms go to zero, a diagonalization argument confirms the existence of a Bayesian test for \(B_H \) versus \(V \). To control the power of this test and to generalise to the case where \(\epsilon = \epsilon_n \) is \(n\)-dependent, more is required: as we increase \(m \) with \(n \), the prior mass \(\delta_{m(n)} \) outside of \(K_n = K_{m(n)} \) must decrease fast enough, while the order of the cover must be bounded: if \(\Pi_n(K_n) \geq 1 - \exp(-L_1 n_1 \epsilon_n^2) \) and the Hellinger entropy of \(K_n \) satisfies \(\log N(\epsilon_n, K_n, H) \leq L_2 \epsilon_n^2 \) for some \(L_1, L_2 > 0 \), there exist \(M > 1, L > 0 \), and a sequence of tests \((\phi_n) \) such that,

\[
\int_{B_H(\epsilon_n)} P^\rho \phi_n d\Pi(P) + \int_{V(\epsilon_n)} Q^\rho (1 - \phi_n) d\Pi(Q) \leq e^{-Ln_1^2},
\]

for large enough \(n \). (For related constructions, see Barron (1988) [9], Barron et al. (1999) [14] and Ghosal, Ghosh and van der Vaart (2000) [101].)

### 7.4.3 Finite sample spaces and the tailfree case

Example 7.14 below demonstrates posterior consistency in total variation for \(i.i.d.\) data from a finite sample space, for priors of full support. Extending this, example 7.15 concerns consistency of posteriors for priors that have Freedman’s tailfreeness property [93, 87], like the Dirichlet process prior.

**Example 7.13.** Consider the situation where we observe an \(i.i.d.\) sample of random variables \(X_1, X_2, \ldots \) taking values in a space \(\mathcal{X}_N \) of finite order \(N \). Writing \(\mathcal{X}_N \) as the set of integers \(\{1, \ldots, N\} \), we note that the space \(M \) of all probability measures \(P \) on \((\mathcal{X}_N, 2^{\mathcal{X}_N})\) with the total-variation metric \((P, Q) \mapsto \|P - Q\|\) is in isometric correspondence with the simplex,

\[ S_N = \{p = (p(1), \ldots, p(N)) : \min_k p(k) \geq 0, \Sigma_i p(i) = 1\}, \]

with the metric \((p, q) \mapsto \|p - q\| = \Sigma_k |p(k) - q(k)|\) it inherits from \(\mathbb{R}^N\) with the \(L^1\)-norm, when \(k \mapsto p(k)\) is the density of \(P \in M\) with respect to the counting measure.

We also define \(M' = \{P \in M : P(\{k\}) > 0, 1 \leq k \leq N\} \subset M\) (and \(R_N = \{p \in S_N : p(k) > 0, 1 \leq k \leq N\} \subset S_N\)).

**Proposition 7.6.** If the data is an \(i.i.d.\) sample of \(\mathcal{X}_N\)-valued random variables, then for any \(n \geq 1 \), any Borel prior \(\Pi : \mathcal{F} \rightarrow [0, 1]\) of full support on \(M \), any \(P_0 \in M \) and any ball \(B\) around \(P_0\), there exists an \(\epsilon' > 0\) such that,

\[
P_n^\rho < e^{\frac{1}{2} \epsilon'^2} P_0^{\Pi^{|B|}}, \tag{7.19}
\]
for all $0 < \varepsilon < \varepsilon'$.

Proof. By the inequality $\|P - Q\| \leq -P \log(dQ/dP)$, the ball $B$ around $P_0$ contains all sets of the form $K(\varepsilon) = \{P \in M' : -P_0 \log(dP/dP_0) < \varepsilon\}$, for some $\varepsilon' > 0$ and all $0 < \varepsilon < \varepsilon'$. Fix such an $\varepsilon$. Because the mapping $P \mapsto -P_0 \log(dP/dP_0)$ is continuous on $M'$ and $M'$ is dense in $M$, there exists an open neighbourhood $U$ of $P_0$ in $M$ such that $U \cap M' \subset K(\varepsilon)$. Since both $M'$ and $U$ are open and $\Pi$ has full support, $\Pi(K(\varepsilon)) \geq \Pi(U \cap M') > 0$. With the help of example 7.1, we see that for every $P \in K(\varepsilon)$,

$$e^{\frac{1}{2} n e^2} \frac{dP^n}{dP_0}(X^n) \geq 1,$$

for large enough $n$, $P_0$-almost-surely. Fatou's lemma again confirms condition (ii) of lemma 7.2 is satisfied. Conclude that assertion (7.19) holds.

Example 7.14. We continue with the situation where we observe an i.i.d. sample of random variables $X_1, X_2, \ldots$ taking values in a space $\mathcal{X}_N$ of finite order $N$. For given $\delta > 0$, consider the hypotheses,

$$B = \{P \in M : \|P - P_0\| < \delta\}, \quad V = \{Q \in M : \|Q - P_0\| > 2\delta\}.$$

Noting that $M$ is compact (or with the help of the simplex representation $S_N$) one sees that entropy numbers of $M$ are bounded, so the construction of example 7.11 shows that uniform tests of exponential power $e^{-nD}$ (for some $D > 0$) exist for $B$ versus $V$. Application of proposition 7.6 shows that the choice for an $0 < \varepsilon < \varepsilon'$ small enough, guarantees that $\Pi(V|X^n)$ goes to zero in $P^n_0$-probability. Conclude that the posterior resulting from a prior $\tilde{\Pi}$ of full support on $M$ is consistent in total variation.

Example 7.15. Recall the construction of the Dirichlet process in subsection 6.4; we repeat the basic steps below. Consider Borel measurable partitions of a Polish sample space, say $\mathcal{X} = \mathbb{R}$, to define a “random distribution” $P$ on $\mathcal{X}$, we specify for each partition $\alpha = \{A_1, \ldots, A_N\}$, a Borel prior $\Pi_\alpha$ on $S_N$, identifying $(p_1, \ldots, p_N)$ with the “random variables” $(P(A_1), \ldots, P(A_N))$. Kolmogorov existence of the stochastic process describing all $P(A)$ in a coupled way subjects these $\Pi_\alpha$ to consistency requirements expressing that if $A_1, A_2$ partition $A$, then $P(A_1) + P(A_2)$ must have the same distribution as $P(A)$. If the partitions refine appropriately, the resulting process describes a probability measure $\Pi$ on the space of Borel probability measures on $\mathcal{X}$, i.e. a random distribution on $\mathcal{X}$. Examples are the Dirichlet process prior of subsection 6.4 and the so-called Polya-tree prior.

A special class of priors constructed in this way are the so-called tailfree priors [93, 87]. The process prior associated with a family of $\Pi_\alpha$ like above is said to be tailfree, if for all $\alpha, \beta$ such that $\beta = \{B_1, \ldots, B_M\}$ refines $\alpha = \{A_1, \ldots, A_N\}$, the following holds: for all $1 \leq k \leq N$, $(P(B_{1|k}|A_k), \ldots, P(B_{M|k}|A_k))$ (where the sets $B_{i|k}, \ldots, B_{k|k} \in \beta$ partition $A_k$) is independent of $(P(A_1), \ldots, P(A_N))$. Although somewhat technical, explicit control of the choice for the $\Pi_\alpha$ render the property quite feasible in examples.
Fix a finite, measurable partition $\alpha = \{A_1, \ldots, A_N\}$. For every $n \geq 1$, denote by $\mathcal{S}_{\alpha,n}$ the $\sigma$-algebra $\mathcal{S}(\alpha^n) \subset \mathcal{B}^n$, generated by products of the form $A_{i_1} \times \cdots \times A_{i_n} \subset \mathcal{B}^n$, with $1 \leq i_1, \ldots, i_n \leq N$. Identify $\mathcal{F}_N$ with the collection of unit-vectors $\{e_1, \ldots, e_N\} \subset \mathcal{B}^N$ and define the projection $\varphi_\alpha : \mathcal{F} \to \mathcal{F}_N$ by,

$$\varphi_\alpha(x) = \{1\{x \in A_1\}, \ldots, 1\{x \in A_N\}\}.$$  

We view $\mathcal{F}_N$ (respectively $\mathcal{F}^n_N$) as a probability space, with $\sigma$-algebra $\mathcal{S}_N$ equal to the power set (respectively $\mathcal{S}_{\alpha,n}$, the power set of $\mathcal{F}^n_N$) and probability measures denoted $P_\alpha : \mathcal{S}_N \to [0,1]$ that we identify with elements of $S_N$. Denoting the space of all Borel probability measures on $\mathcal{F}$ by $M^1(\mathcal{F})$, we also define $\varphi_\alpha : M^1(\mathcal{F}) \to S_N,$

$$\varphi_\alpha(P) = (P(A_1), \ldots, P(A_N)),$$

which maps $P$ to its restriction to $\mathcal{S}_{\alpha,1}$, a probability measure on $\mathcal{F}_N$. Under the projection $\varphi_\alpha$, any Borel-measurable random variable $X$ taking values in $\mathcal{F}$ distributed $P \in M^1(\mathcal{F})$ is mapped to a random variable $Z_\alpha = \varphi_\alpha(X)$ that takes values in $\mathcal{F}_N$ (distributed $P_\alpha = \varphi_\alpha(P)$). We also define $Z^n_\alpha = (\varphi_\alpha(x_1), \ldots, \varphi_\alpha(x_n))$, for all $n \geq 1$.

Let $\Pi_\alpha$ denote a Borel prior on $S_N$. The posterior on $S_N$ is then a Borel measure denoted $\Pi_\alpha(\cdot|Z^n_\alpha)$, which satisfies, for all $A \in \mathcal{S}_{\alpha,n}$ and any Borel set $V$ in $S_N$,

$$\int_A \Pi_\alpha(V|Z^n_\alpha) \, d\Pi^n_A = \int_V P^n_\alpha(A) \, d\Pi_\alpha(P_\alpha),$$

by definition of the posterior. In the model for the original $i.i.d.$ sample $X^n$, Bayes’s rule takes the form, for all $A' \in \mathcal{B}_n$ and all Borel sets $V'$ in $M^1(\mathcal{F})$,

$$\int_{A'} \Pi(V'|X^n) \, d\Pi^n_{A'} = \int_{V'} P^n(A') \, d\Pi(P),$$

defining the posterior for $P$. Now specify that $V'$ is the pre-image $\varphi^{-1}_\alpha(V)$ of a Borel measurable $V$ in $S_N$: as a consequence of tailfreeness, the data-dependence of the posterior for such a $V'$, $X^n \to \Pi(V'|X^n)$, is measurable with respect to $\mathcal{S}_{\alpha,n}$ (see Freedman (1965) [93] or Ghosh (2003) [103]). So there exists a function $g_n : \mathcal{F}_N \to [0,1]$ such that,

$$\Pi(V'|X^n = x^n) = g_n(\varphi_\alpha(x_1), \ldots, \varphi_\alpha(x_n)),$$

for $P^n_\alpha$-almost-all $x^n \in \mathcal{B}^n$. Then, for given $A' \in \mathcal{S}_{\alpha,n}$ (with corresponding $A \in \mathcal{S}_{\alpha,n}$),

$$\int_{A'} \Pi(V'|X^n) \, d\Pi^n_{A'} = \int P^n(A')(1_{A'}(X^n) \Pi(V'|X^n)) \, d\Pi(P)$$

$$= \int P^n_\alpha(1_A(Z^n_\alpha) g_n(Z^n_\alpha)) \, d\Pi_\alpha(P_\alpha) = \int g_n(Z^n_\alpha) \, d\Pi^n_\alpha,$$

while also,

$$\int_{V'} P^n(A') \, d\Pi(P) = \int V' P^n_\alpha(A) \, d\Pi_\alpha(P_\alpha).$$
This shows that $Z_n^a \rightarrow g_\alpha(Z_n^a)$ is a version of the posterior $\Pi_\alpha(\cdot | Z_n^a)$. In other words, we can write $\Pi(V'|X^n) = \Pi_\alpha(V|\phi_\alpha(X^n)) = \Pi_\alpha(V|Z_n^a), P_n^\Pi$-almost-surely.

Denote the true distribution of a single observation from $X^n$ by $P_0$. For any $V'$ of the form $\varphi_{-1}^{-1}(V)$ for some $\alpha$ and a neighbourhood $V$ of $P_{0,\alpha} = \varphi_{\alpha}(P_0)$ in $S_N$, the question whether $\Pi(V'|X^n)$ converges to one in $P_0$-probability reduces to the question whether $\Pi(V|Z_n^a)$ converges to one in $P_{0,\alpha}$-probability. Remote contiguity then only has to hold as in example 7.13.

Another way of saying this is to note directly that, because $X^n \rightarrow \Pi(V'|X^n)$ is $\sigma_a$-measurable, remote contiguity (as in definition 7.2) is to be imposed only for $\phi_n : \mathcal{X} \rightarrow [0,1]$ that are measurable with respect to $\sigma_{\alpha,n}$ (rather than $\mathcal{B}^n$) for every $n \geq 1$. That conclusion again reduces the remote contiguity requirement necessary for the consistency of the posterior for the parameter $(P(A_1),\ldots,P(A_N))$ to that of a finite sample space, as in example 7.13. Full supports of the priors $\Pi_\alpha$ then guarantee remote contiguity for exponential rates as required in condition (ii) of theorem 7.2. The uniform tests of exponential power for weak neighbourhoods, as in proposition A.1, complete the proof that tailfree priors lead to consistent posterior distributions. The implication of consistency of the posteriors $\Pi(|Z_n^a)$ for all $\alpha$ for the posterior $\Pi(|X^n)$, is that it is consistent in the so-called inverse limit topology that the spaces $(S_\alpha : \alpha \in \mathcal{A})$ induce on $M^1(\mathcal{X})$.

The following theorem is more specific than the preceding discussion suggests: the reason is that sufficient conditions have been included for the existence of the inverse limit prior $\Pi$ we construct when we specify the priors $\Pi_\alpha$. See appendix D for details.

**Theorem 7.3.** Let $\mathcal{X}$ be a Polish space and let $\mathcal{P}$ be the full model $M^1(\mathcal{X})$, endowed with Prokhorov’s weak topology. For given $P_0 \in \mathcal{P}$, consider i.i.d. samples $(X_1,\ldots,X_n) \sim P_n$. Let $(\alpha_n)$ be a sequence of finite partitions of $\mathcal{X}$ such that,

(i) $\alpha_n$ consists only of open and closed subsets, for all $n \geq 1$,
(ii) $\alpha_{n+1}$ refines $\alpha_n$, for all $n \geq 1$,
(iii) for every $x,y \in \mathcal{X}$, if $x \neq y$, there is an $n \geq 1$ such that $x$ and $y$ lie in distinct elements of $\alpha_n$,
(iv) for every open $U \subset \mathcal{X}$, there exist an $n \geq 1$ and an element $A \in \alpha_n$ such that $A \subset U$.

Let $\Pi_n$ be a tailfree sequence of priors on the simplices $S_n = S_{N(\alpha_n)}$ of full support, satisfying inverse limit condition (??). Then the posterior is consistent, i.e.,

$$\int f(P) d\Pi_n(P|X^n) \stackrel{P_0}{\rightarrow} f(P_0),$$

for any bounded, continuous $f : \mathcal{P} \rightarrow \mathbb{R}$.

In the case of the Dirichlet process prior condition (??) is satisfied and full support of the base measure $\mu$ implies full support for all $\Pi_n$, if we restrict attention to partitions $\alpha = (A_1,\ldots,A_N)$ such that $\mu(A_i) > 0$ for all $1 \leq i \leq N$. (Particularly, we require $P_0 \ll \mu$ for consistent estimation.)
7.5 Rates of posterior concentration

A significant extension to the theory on posterior convergence is formed by results concerning posterior convergence in metric spaces at a rate. Minimax rates of convergence for (estimators based on) posterior distributions were considered more or less simultaneously in Ghosal-Ghosh-van der Vaart [101] and Shen-Wasserman [204]. Both propose an extension of Schwartz’s theorem to posterior rates of convergence [101, 204] and apply Barron’s sieve idea with a well-known entropy argument [32, 33] to a shrinking sequence of Hellinger neighbourhoods and employs a more specific, rate-related version of the Kullback-Leibler condition (6.15) for the prior. Both appear to be inspired by contemporary results regarding Hellinger rates of convergence for sieve MLE’s, as well as on Barron-Schervish-Wasserman [14], which concerns posterior consistency based on controlled bracketing entropy for a sieve, up to subsets of negligible prior mass, following ideas that were first laid down in [9]. It is remarked already in [14] that their main theorem is easily re-formulated as a rate-of-convergence theorem, with reference to [204]. More recently, Walker, Lijoi and Prünster [224] have added to these considerations with a theorem for Hellinger rates of posterior concentration in models that are separable for the Hellinger metric, with a central condition that calls for summability of square-roots of prior masses of covers of the model by Hellinger balls, based on analogous consistency results in Walker [221]. More recent is [145], which shows that alternatives for the priors of [101, 204] exist.

Theorem 7.4. Assume that for all \( n \geq 1 \), the data \( X^n \sim P_{\theta_0,n} \) for some \( \theta_0 \in \Theta \). Fix priors \( \Pi_n : \mathcal{G} \to [0,1] \) and assume that for given \( B_n \), \( V_n \in \mathcal{G} \) with \( \Pi_n(B_n) > 0 \) and \( a_n, b_n \downarrow 0 \) such that \( a_n = o(b_n) \).

(i) there are Bayesian tests \( \phi_n : X_n \to [0,1] \) such that,

\[
\int_{B_n} P_{\theta_0,n}(1 - \phi_n) d\Pi_n(\theta) + \int_{V_n} P_{\theta_0,n}(\phi_n) d\Pi_n(\theta) = o(a_n),
\]

(ii) The prior mass of \( B_n \) is lower-bounded, \( \Pi_n(B_n) \geq b_n \).

(iii) The sequence \( P_{\theta_0,n} \) satisfies \( P_{\theta_0,n} \prec b_n a_n^{-1} \Pi_n|B_n \).

Then \( \Pi(V_n|X^n)^{P_{\theta_0,n}} \rightarrow 0 \).

Proof. Proposition 7.1 says that \( P_n^{\Pi_n|B_n} \Pi(V_n|X^n) \) is of order \( o(b_n^{-1}a_n) \). Condition (iii) then implies that \( P_{\theta_0,n} \Pi(V_n|X^n) = o(1) \), which is equivalent to \( \Pi(V_n|X^n)^{P_{\theta_0,n}} \rightarrow 0 \) since \( 0 \leq \Pi(V_n|X^n) \leq 1 \), \( P_{\theta_0,n} \)-almost-surely for all \( n \geq 1 \).

To connect with the literature we interpret lower bound (7.21) again, reformulating lemma 6.7 as a statement of remote contiguity.

Lemma 7.4. For all \( n \geq 1 \), assume that \( (X_1, X_2, \ldots, X_n) \in \mathcal{G}^n \sim P_0^n \) for some \( P_0 \in \mathcal{P} \) and let \( \epsilon_n \downarrow 0 \) be given. Let \( B_n \) be as in example 7.17. Then, for any priors \( \Pi_n \) such that \( \Pi_n(B_n) > 0 \),
for any constant $c > 1$.

**Example 7.16.** To apply theorem 7.4, consider again the situation of a uniform distribution with an unknown location, as in examples 6.4 and 7.7. Take $V_n$ equal to $\{\theta : \theta - \theta_0 > \varepsilon_n\}$ $\{\theta : \theta - \theta > \varepsilon_n\}$ respectively, with $\varepsilon_n = M_n/n$ for some $M_n \to \infty$. It is noted that, for every $0 < c < 1$, the likelihood ratio test,

$$
\phi_n(X^n) = 1\{dP_{\theta_0+\varepsilon_n/n} / dP_{\theta_0,n}(X^n) > c\} = 1\{X_{(1)} > \theta_0 + \varepsilon_n\},
$$

satisfies $P_n(b_n) = 0$ for all $\theta \in V_n$, and if we choose $\delta_n = 1/2$ and $\varepsilon_n = M_n/n$ for some $M_n \to \infty$, $P_n(b_n) \leq e^{-M_n+1}$ for all $\theta \in B_n$, so that,

$$
\int_{B_n} P_n b_n(\theta) \frac{d\pi(\theta)}{\pi(\theta)} + \int_{V_n} P_n(b_n) d\pi(\theta) \leq \pi(B_n) e^{-M_n+1},
$$

Using lemma 7.1, we see that $P_n(b_n) \pi(V_n|X^n) \leq e^{-M_n+1}$. Based on the conclusion of example ?? above, remote contiguity implies that $P_n(b_n) \pi(V_n|X^n) \to 0$. Treating the case $\theta < \theta_0 - \varepsilon_n$ similarly, we conclude that the posterior is consistent at (any $\varepsilon_n$ slower than) rate $1/n$.

**Example 7.17.** Let us briefly review the conditions of [14, 101, 204] in light of theorem 7.4: let $\varepsilon_n \downarrow 0$ denote the Hellinger rate of convergence we have in mind, let $M > 1$ be some constant and define,

$$
V_n = \{P \in \mathcal{P} : H(P, P_0) \geq M \varepsilon_n\},
$$

$$
B_n = \{P \in \mathcal{P} : -P_0 \log dP / dP_0 < \varepsilon_n^2, P_0 \log^2 dP / dP_0 < \varepsilon_n^3\}.
$$

If $\varepsilon = \varepsilon_n$ with $\varepsilon_n \downarrow 0$ and $n\varepsilon_n^2 \to \infty$, and the model’s Hellinger entropy is upper-bounded by $\log N(\varepsilon_n, \mathcal{P}, H) \leq Kn\varepsilon_n^2$ for some $K > 0$, the construction of example 7.11 extends to tests that separate $V_n = \{P \in \mathcal{P} : H(P, P_0) \geq 4\varepsilon_n\}$ from $B_n = \{P \in \mathcal{P} : H(P_0, P) < \varepsilon_n\}$ asymptotically, with power $\exp(-nL\varepsilon_n^2)$ for some $L > 0$. (See also the so-called Le Cam dimension of a model [160] and Birgê’s rate-oriented work [32, 33].) It is worth pointing out at this stage that posterior inconsistency due to the phenomenon of ‘data tracking’ [14, 223], whereby weak posterior consistency holds but Hellinger consistency fails, can only be due to failure of the testing condition in the Hellinger case.

Note that $B_n$ is contained in the Hellinger ball of radius $\varepsilon_n$ around $P_0$, so (7.20) holds. New in [101, 204] is the condition for the priors $\pi_n$,

$$
\pi_n(B_n) \geq e^{-C \varepsilon_n^2},
$$

for some $C > 0$. With the help of lemmas 7.4 and 7.2-(ii), we conclude that,

$$
P_0(n) \leq e^{C \varepsilon_n^2} P_{\pi, B_n}^{\pi, B_n},
$$
7.5 Rates of posterior concentration

for any \( c > 1 \). If we choose \( M \) such that \( DM^2 - C > 1 \), theorem 7.4 proves that
\[
\Pi(V_n | X^n) \xrightarrow{P_0} 0,
\]
i.e. the posterior is Hellinger consistent at rate \( \varepsilon_n \).

Note that the argument also extends to models that are Hellinger separable: in that case (7.18) remains valid, but with \( N(\varepsilon) = \infty \). The mass fractions \( \Pi(V_i) / \Pi(B') \) become important (we point to strong connections with Walker’s theorem [221, 224]). Here we see the balance between prior mass and testing power for Bayesian tests, as intended by the remark that closes the subsection on the existence of Bayesian test sequences in section 7.1.

Certain (simple, parametric) models do not allow the definition of priors that satisfy (7.21), and alternative less restrictive choices for the sets \( B_n \) are possible under mild conditions on the model [145].

7.5.1 Remote contiguity and the LAN condition

To conclude we consider remote contiguity under the condition that the model is LAN (see definition 4.1 and LeCam (1960) [157]).

Lemma 7.5. Assume that the model satisfies LAN condition (4.2) with non-singular \( I_{\theta_0} \) and that the prior \( \Pi \) for \( \theta \) has a Lebesgue-density \( \pi : \mathbb{R}^d \to \mathbb{R} \) that is continuous and strictly positive in all of \( \Theta \). For given \( H > 0 \), define the subsets \( B_n = \{ \theta \in \Theta : \theta = \theta_0 + n^{-1/2} h, ||h|| \leq H \} \). Then,
\[
P_{0,n} < c_n^{-1} P_{n,B_n},
\]
for any \( c_n \downarrow 0 \).

Proof. According to lemma 3 in section 8.4 of Le Cam and Yang (1990) [166], \( P_{0,n} \) is contiguous with respect to \( P_{n,B_n} \). That implies the assertion.

Note that for some \( K > 0 \), \( \Pi(B_n) \geq b_n := K(H / \sqrt{n})^d \). Assume again the existence of Bayesian tests for \( V = \{ \theta \in \Theta : ||\theta - \theta_0|| > \rho \} \) (for some \( \rho > 0 \)) versus \( B_n \) (or some \( B \) such that \( B_n \subset B \)), of power \( a_n = \exp(-\frac{1}{2} n \tau^2) \) (for some \( \tau > 0 \)). Then \( a_n b_n^{-1} = o(1) \), and, assuming (7.23), theorem 7.4 implies that \( \Pi(||\theta - \theta_0|| > \rho | X^n) \xrightarrow{P_{0,n}} 0 \), so consistency is straightforwardly demonstrated.

The case becomes somewhat more complicated if we are interested in optimality of parametric rates: following the above, a logarithmic correction arises from the lower bound \( \Pi(B_n) \geq K(H / \sqrt{n})^d \) when combined in the application of theorem 7.4. To alleviate this, we adapt the construction somewhat: define \( V_n = \{ \theta \in \Theta : ||\theta - \theta_0|| \leq M_n n^{-1/2} \} \) for some \( M_n \to \infty \) and \( B_n \) like above. Under the condition that there exists a uniform test sequence for any fixed \( V = \{ \theta \in \Theta : ||\theta - \theta_0|| > \rho \} \) versus \( B_n \) (see, for example, [?]'), uniform test sequences for \( V_n \) versus \( B_n \) of power \( e^{-K' M_n^2} \) exist, for some \( k' > 0 \). Alternatively, assume that the Hellinger distance and the norm on \( \Theta \) are related through inequalities of the form,
for some constants $K_1, K_2 > 0$. Then cover $V_n$ with rings,

$$V_{n,k} = \left\{ \theta \in V_n : \frac{(M_n + k - 1)}{\sqrt{n}} \leq ||\theta - \theta_0|| \leq \frac{(M_n + k)}{\sqrt{n}} \right\},$$

for $k \geq 1$ and cover each ring with balls $V_{n,k,l}$ of radius $n^{-1/2}$, where $1 \leq l \leq L_{n,k}$ and $L_{n,k}$ the minimal number of radius-$n^{-1/2}$ balls needed to cover $V_{n,k}$, related to the Le Cam dimension [160]. With the $B_n$ defined like above, and the inequality,

$$\int P_{\theta,n}(V_{n,k,l}) d\Pi_n(\theta|B_n) \leq \sup_{\theta \in B_n} P_{\theta,n}(1 - \phi_{n,k,l}) \Pi_n(V_{n,k,l}) \Pi_n(B_n),$$

where the $\phi_{n,k,l}$ are the uniform minimax tests for $B_n$ versus $V_{n,k,l}$ of lemma 6.3, of power $\exp(-K'(M_n + k - 1)^2)$ for some $K' > 0$. We define $\phi_{n,k} = \max\{\phi_{n,k,l} : 1 \leq l \leq L_{n,k}\}$ for $V_{n,k}$ versus $B_n$ and note,

$$\int P_{\theta,n}(V_{n,k,l}) d\Pi_n(\theta|B_n) \leq \left( L_{n,k} + \frac{\Pi_n(V_{n,k})}{\Pi_n(B_n)} \right) e^{-K(M_n + k - 1)^2},$$

where the numbers $L_{n,k}$ are upper bounded by a multiple of $(M_n + k)^d$ and the fraction of prior masses $\Pi_n(V_{n,k})/\Pi_n(B_n)$ can be controlled without logarithmic corrections when summing over $k$ next.

### 7.6 Consistent hypothesis testing with Bayes factors

The Neyman-Pearson paradigm notwithstanding, hypothesis testing and classification concern the same fundamental statistical question, to find a procedure to choose one subset from a given partition of the parameter space as the most likely to contain the parameter value of the distribution that has generated the data observed. Asymptotically one wonders whether choices following such a procedure focus on the correct subset with probability growing to one.

From a somewhat shifted perspective, we argue as follows: no statistician can be certain of the validity of specifics in his model choice and therefore always runs the risk of biasing his analysis from the outset. Non-parametric approaches alleviate his concern but imply greater uncertainty within the model, leaving the statistician with the desire to select the correct (sub)model on the basis of the data before embarking upon the statistical analysis proper (for a recent overview, see [213]). The issue also makes an appearance in asymptotic context, where over-parametrized models leave room for inconsistency of estimators, requiring regularization [34, 35, 48].
Model selection describes all statistical methods that attempt to determine from the data which model to use. (Take for example sparse variable selection, where one projects out the majority of covariates prior to actual estimation, and the model-selection question is which projection is optimal.) Methods for model selection range from simple rules-of-thumb, to cross-validation and penalization of the likelihood function. Here we propose to conduct the frequentist analysis with the help of a posterior: when faced with a (dichotomous) model choice, we let the so-called Bayes factor formulate our preference. For an analysis of hypothesis testing that compares Bayesian and frequentist views, see [12]. An objective Bayesian perspective on model selection is provided in [226].

**Definition 7.4.** For all \( n \geq 1 \), let the model be parametrized by maps \( \theta \mapsto P_{\theta,n} \) on a parameter space \((\Theta, \mathcal{G})\) with priors \( \Pi_n : \mathcal{G} \to [0,1] \). Consider disjoint, measurable \( B, V \subset \Theta \). For given \( n \geq 1 \), we say that the Bayes factor for testing \( B \) versus \( V \),

\[
F_n = \frac{\Pi(B|X^n) \Pi_n(V)}{\Pi(V|X^n) \Pi_n(B)},
\]

is consistent for testing \( B \) versus \( V \), if for all \( \theta \in V \), \( F_n \xrightarrow{P_{\theta,n}} 0 \) and for all \( \theta \in B \), \( F_n^{-1} \xrightarrow{P_{\theta,n}} 0 \).

### 7.6.1 Frequentist model selection with posteriors

Let us first consider this from a purely Bayesian perspective: for fixed prior \( \Pi \) and i.i.d. data, theorem 8.7 says that the posterior gives rise to consistent Bayes factors for \( B \) versus \( V \) in a Bayesian (that is, \( \Pi \)-almost-sure) way, iff a Bayesian test sequence for \( B \) versus \( V \) exists. If the parameter space \( \Theta \) is Polish and the maps \( \theta \mapsto P_{\theta}(A) \) are Borel measurable for all \( A \in \mathcal{G} \), proposition 7.2 says that any Borel set \( V \) is Bayesian testable versus \( \Theta \setminus V \), so in Polish models for i.i.d. data, model selection with Bayes factors is \( \Pi \)-almost-surely consistent for all Borel measurable \( V \subset \Theta \).

The frequentist requires strictly more, however, so we employ remote contiguity again to bridge the gap with the Bayesian formulation.

**Theorem 7.5.** For all \( n \geq 1 \), let the model be parametrized by maps \( \theta \mapsto P_{\theta,n} \) on a parameter space \((\Theta, \mathcal{G})\) with priors \( \Pi_n : \mathcal{G} \to [0,1] \). Consider disjoint, measurable \( B, V \subset \Theta \) with \( \Pi_n(B), \Pi_n(V) > 0 \) such that,

(i) There exist Bayesian tests for \( B \) versus \( V \) of power \( a_n \downarrow 0 \),

\[
\int_B P^\theta \phi_n d\Pi_n(P) + \int_V Q^\theta (1 - \phi_n) d\Pi_n(Q) = o(a_n),
\]

(ii) For every \( \theta \in B \), \( P_{\theta,n} \preceq a_n^{-1} P_n^{\Pi_n|B} \), and for every \( \theta \in V \), \( P_{\theta,n} \preceq a_n^{-1} P_n^{\Pi_n|V} \).
Then the Bayes factor for $B$ versus $V$ is consistent.

Note that the second condition of theorem 8.10 can be replaced by a local condition: if, for every $\theta \in B$, there exists a sequence $B_n(\theta) \subset B$ such that $I_n(B_n(\theta)) \geq b_n$ and $P_{\theta,n} \leq a_n^{-1} b_n P_{n\Pi|B_n}^{\Pi|B}$, then $P_{\theta,n} \leq a_n^{-1} P_{n\Pi|B}$ (as a consequence of lemma 7.3 with $C_n = B$).

### 7.6.2 Goodness-of-fit Bayes factors for random walks

Consider the asymptotic consistency of goodness-of-fit tests for the transition kernel of a Markov chain with posterior odds or Bayes factors. Bayesian analyses of Markov chains on a finite state space are found in [209] and references therein. Consistency results c.f. [221] for random walk data are found in [104]. Large-deviation results for posterior distributions are derived in [185, 80]. The examples below are based on ergodicity for remote contiguity and Hoeffding’s inequality for uniformly ergodic Markov chains [173, 108] to construct suitable tests. We first prove the analogue of Schwartz’s construction in the case of an ergodic random walk.

Let $(S, \mathcal{S})$ denote a measurable state space for a discrete-time, stationary Markov process $P$ describing a random walk $X^n = \{X_i \in S : 0 \leq i \leq n\}$ of length $n \geq 1$ (conditional on a starting position $X_0$). The chain has a Markov transition kernel $P(\cdot | \cdot) : \mathcal{S} \times S \to [0, 1]$ that describes $X_i | X_{i-1}$ for all $i \geq 1$.

Led by Pearson’s approach to goodness-of-fit testing, we choose a finite partition $\alpha = \{A_1, \ldots, A_N\}$ of $S$ and ‘bin the data’ in the sense that we switch to a new process $Z^n$ taking values in the finite state space $S_\alpha = \{e_j : 1 \leq j \leq N\}$ (where $e_j$ denotes the $j$-th standard basis vector in $\mathbb{R}^n$), defined by $Z^n = \{Z_i \in S_\alpha : 0 \leq i \leq n\}$, with $Z_i = (\{X_i \in A_1\}, \ldots, \{X_i \in A_N\})$. The process $Z^n$ forms a stationary Markov chain on $S_\alpha$ with distribution $P_{\alpha,n}$. The model is parametrized in terms of the convex set $\Theta$ of $N \times N$ Markov transition matrices $p_\alpha$ on the finite state space $S_\alpha$.

\begin{equation}
    p_\alpha(k|l) = P_{\alpha,n}(Z_i = e_k | Z_{i-1} = e_l) = P(X_i \in A_k | X_{i-1} \in A_l),
\end{equation}

for all $0 \leq i \leq n$ and $1 \leq k, l \leq N$. We assume that $P_{\alpha,n}$ is ergodic with equilibrium distribution that we denote by $\pi_\alpha$, and $\pi_\alpha(k) := \pi_\alpha(Z = k)$. We are interested in Bayes factors for goodness-of-fit type questions, given a parameter space consisting of transition matrices.

**Example 7.18.** Assume that the true transition kernel $P_0$ gives rise to a matrix $p_0 \in \Theta$ that generates an ergodic Markov chain $Z^n$. Denote the true distribution of $Z^n$ by $P_{0,n}$ and the equilibrium distribution by $\pi_0$ (with $\pi_0(k) := \pi_0(Z = k)$). For given $\epsilon > 0$, define,

\begin{equation}
    B' = \left\{ p_\alpha \in \Theta : \sum_{k,l=1}^N -p_0(l|k)\pi_0(k) \log \frac{p_\alpha(l|k)}{p_0(l|k)} < \epsilon^2 \right\}.
\end{equation}

Assume that $\Pi(B') > 0$. According to the ergodic theorem, for every $p_\alpha \in B'$,
For some fixed $P$ distribution for two consecutive steps in the random walk. Like Pearson, we fix some such example 7.19. We formulate goodness-of-fit hypotheses in terms of the joint distribution. Instead, we shall resort to local asymptotic normality for a sharper result.

However, exponential remote contiguity will turn out not to be enough for goodness-of-fit tests below, unless we impose stringent model conditions. Instead, we shall consider the set \( \Pi(B) = \{ \Pi : p_0 \ll \Pi \} \) such that hypotheses \( \Pi(B) > 0 \) and Fatou’s lemma imply remote contiguity because, \[
P_{0,n} \left( \int \frac{dP_{\alpha,n}}{dP_{0,n}}(Z^n) d\Pi(p|B') < e^{\frac{n}{2} \epsilon^2} \right) \to 0.
\]

So lemma 7.2 says that \( P_{0,n} < \exp(\frac{n}{2} \epsilon^2) P_n^{\Pi|B'} \).

Example 7.19. We formulate goodness-of-fit hypotheses in terms of the joint distribution for two consecutive steps in the random walk. Like Pearson, we fix some such distribution \( P_0 \) and consider hypotheses based on differences of ‘bin probabilities’ \( p_\alpha(k,l) = p_\alpha(k|l)\pi_\alpha(l) \).

\[
H_0 : \max_{1 \le k,l \le N} |p_\alpha(k,l) - p_0(k,l)| < \epsilon,
\]

\[
H_1 : \max_{1 \le k,l \le N} |p_\alpha(k,l) - p_0(k,l)| \ge \epsilon, \tag{7.25}
\]

for some fixed \( \epsilon > 0 \). The sets \( B \) and \( V \) are defined as the sets of transition matrices \( p_\alpha \in \Theta \) that satisfy hypotheses \( H_0 \) and \( H_1 \) respectively. We assume that the prior is chosen such that \( \Pi(B) > 0 \) and \( \Pi(V) > 0 \).

Endowed with some matrix norm, \( \Theta \) is compact and a Borel prior on \( \Theta \) can be defined in various ways. For example, we may assign the vector \( (p_\alpha(1), \ldots, p_\alpha(N)) \) a product of Dirichlet distributions. Conjugacy applies and the posterior for \( p_\alpha \) is again a product of Dirichlet distributions [209]. For an alternative family of priors, consider the set \( \delta \) of \( N \times N \)-matrices \( E \) that have standard basis vectors \( e_k \) in \( \mathbb{R}^N \) as columns. Each \( E \in \delta \) is a deterministic Markov transition matrix on \( S_\alpha \) and \( \delta \) is the extremal set of the polyhedral set \( \Theta \). According to Choquet’s theorem, every transition matrix \( p_\alpha \) can then be written in the form,

\[
p_\alpha = \sum_{E \in \delta} \lambda_E E, \tag{7.26}
\]

for a (non-unique) combination of \( \lambda_\delta := \{ \lambda_E : E \in \delta \} \) such that \( \lambda_E \ge 0, \sum_\delta \lambda_E = 1 \). If \( \lambda_E > 0 \) for all \( E \in \delta \), the resulting Markov chain is ergodic and we denote the...
corresponding distributions for $Z^n$ by $P_{a,n}$. Any Borel prior $\Pi'$ (e.g. a Dirichlet distribution) on the simplex $S_{nN}$ in $\mathbb{R}^{nN}$ is a prior for $\lambda_\varepsilon$ and induces a Borel prior $\Pi$ on $\Theta$. Note that all non-ergodic transition matrices lie in the boundary $\partial \Theta$, so if we choose $\Pi'$ such that $\Pi(\Theta) = 1$, ergodicity may be assumed in all prior-almost-sure arguments. This is true for any $\Pi'$ that is absolutely continuous with respect to the $(N^N - 1)$-dimensional Lebesgue measure on $S_{nN}$ (for example when we choose $\Pi'$ equal to a Dirichlet distribution). Note that if the associated density is continuous and strictly positive, $\Pi(B) > 0$ and $\Pi(V) > 0$.

We intend to use theorem 8.10 with $B$ and $V$ defined by $H_0$ and $H_1$, so we first demonstrate that a Bayesian test sequence for $B$ versus $V$ exists, based on a version of Hoeffding’s inequality valid for random walks [108]. First, define, for given $0 < \lambda_n \leq N^{-N}$ such that $\lambda_n \downarrow 0$,

$$S'_n := \{ \lambda_\varepsilon \in S^{nN} : \lambda_E \geq \lambda_n/N^{N-1}, \text{for all } E \in \mathcal{E} \},$$

and denote the image of $S'_n$ under (7.26) by $S_n$. Note that if $\Pi(\partial \Theta) = 0$, then $\pi_{S,n} := \Pi(\Theta \setminus S_n) \to 0$.

Now fix $n \geq 1$ for the moment. Recalling the nature of the matrices $E$, we see that for every $1 \leq k, l \leq N$, $p_a(k|l)$ as in equation (7.26) is greater than or equal to $\lambda_n$. Consequently, the corresponding Markov chain satisfies condition (A.1) of Glynn and Ormoneit [108] (closely related to the notion of uniform ergodicity [173]): starting in any point $X_0$ under a transition from $S_n$, the probability that $X_1$ lies in $A \subset S_n$ is greater than or equal to $\lambda_n \phi(A)$, where $\phi$ is the uniform probability measure on $S_n$. This mixing condition enables a version of Hoeffding’s inequality (see theorem 2 in [108]): for any $\lambda_\varepsilon \in S'_n$ and $1 \leq k, l \leq N$, the transition matrix of equation (7.26) is such that, with $\hat{\rho}_n(k,l) = n^{-1} \sum i \{Z_i = k, Z_{i-1} = l\}$,

$$P_{a,n}(\hat{\rho}_n(k,l) - p_a(k,l) \geq \delta) \leq \exp\left(-\frac{\lambda_n^2(n\delta - 2\lambda_n^{-1})^2}{2n}\right). \quad (7.27)$$

Now define for a given sequence $\varepsilon_n > 0$ with $\varepsilon_n \downarrow 0$ and all $n \geq 1, 1 \leq k, l \leq N$,

$$B_n = \{ p_a \in \Theta : \max_{k,l} |p_a(k,l) - p_0(k,l)| < \varepsilon - \delta_n \},$$

$$V_{k,l} = \{ p_a \in \Theta : |p_a(k,l) - p_0(k,l)| \geq \varepsilon \},$$

$$V_{+,k,l,n} = \{ p_a \in \Theta : p_a(k,l) - p_0(k,l) \geq \varepsilon + \delta_n \},$$

$$V_{-,k,l,n} = \{ p_a \in \Theta : p_a(k,l) - p_0(k,l) \leq -\varepsilon - \delta_n \}. $$

Note that if $\Pi'$ is absolutely continuous with respect to the Lebesgue measure on $S^{nN}$, then $\pi_{B,n} := \Pi(B \setminus B_n) \to 0$ and $\pi_{n,k,l} := \Pi(V_{k,l,n} \setminus (V_{+,k,l,n} \cup V_{-,k,l,n})) \to 0$.

If we define the test $\phi_{+,k,l,n}(Z^n) = 1\{\hat{\rho}_n(k,l) - p_0(k,l) \geq \varepsilon \}$, then for any $p_a \in B_n \cap S_n$,

$$P_{a,n}(\hat{\rho}_n(k,l) - p_a(k,l) \geq \delta_n) \leq \exp\left(-\frac{\lambda_n^2(n\delta_n - 2\lambda_n^{-1})^2}{2n}\right).$$

$$P_{a,n}(\hat{\rho}_n(k,l) - p_a(k,l) \geq \delta_n) \leq \exp\left(-\frac{\lambda_n^2(n\delta_n - 2\lambda_n^{-1})^2}{2n}\right).$$
If on the other hand, \( p_\alpha \) lies in the intersection of \( V_{+,n,k,l} \) with \( S_n \), we find,

\[
P_{\alpha,n}(1 - \phi_{+,n,k,l}(Z^n)) = P_{\alpha,n}(\hat{p}_{\alpha}(k,l) - p_\alpha(k,l) < -\delta_n) \\
\leq \exp\left(-\frac{\lambda_n^2(n\delta_n - 2\lambda_n^{-1})^2}{2n}\right).
\]

Choosing the sequences \( \delta_n \) and \( \lambda_n \) such that \( n\delta_n^2\lambda_n^2 \to \infty \), we also have \( \lambda_n^{-1} = o(n\delta_n) \), so the exponent on the right is smaller than or equal to \( -\frac{1}{4} n\lambda_n^2\delta_n^2 \).

So if we define \( \phi_n(Z^n) = \max_{k,l} \{ \phi_{-,k,l,n}(Z^n), \phi_{+,k,l,n}(Z^n) \} \),

\[
\int_B P_{\alpha,n}\phi_n d\Pi(p_\alpha) + \int_{V_n} \mathcal{Q}_{\alpha,n}(1 - \phi_n) d\Pi(q_\alpha) \\
\leq \int_{B \cap S_n} P_{\alpha,n}\phi_n d\Pi(p_\alpha) + \int_{V_n \cap S_n} \mathcal{Q}_{\alpha,n}(1 - \phi_n) d\Pi(q_\alpha) + \Pi(\Theta \setminus S_n) \\
\leq \sum_{k,l=1}^{N} P_{\alpha,n}(\phi_{-,k,l,n} + \phi_{+,k,l,n}) d\Pi(p_\alpha) \\
+ \sum_{k,l=1}^{N} \left( \int_{V_{-,k,l}} \mathcal{Q}_{\alpha,n}(1 - \phi_{-,k,l,n}) d\Pi(q_\alpha) \right) \\
+ \sum_{k,l=1}^{N} \left( \mathcal{Q}_{\alpha,n}(1 - \phi_{+,k,l,n}) d\Pi(q_\alpha) \right) \\
+ \sum_{k,l=1}^{N} \Pi(V_{+,k,l} \setminus (V_{+,n,k,l} \cup V_{+,n,k,l})) + \Pi(\Theta \setminus S_n) + \Pi(B \setminus B_n) \\
\leq 2N^2 e^{-\frac{1}{4} n\lambda_n^2\delta_n^2} + \pi_{B,n} + \pi_{S,n} + \sum_{k,l=1}^{N} \pi_{n,k,l}.
\]

So if we choose a prior \( \Pi' \) on \( S_N^W \) that is absolutely continuous with respect to Lebesgue measure, then \( (\phi_n) \) defines a Bayesian test sequence for \( B \) versus \( V \).

Because we have not imposed control over the rates at which the terms on the \( r.h.s. \) go to zero, remote contiguity at exponential rates is not good enough. Even if we would restrict supports of a sequence of priors such that \( \pi_{B,n} = \pi_{S,n} = \pi_{n,k,l} = 0 \), the first term on the \( r.h.s. \) is sub-exponential. To obtain a rate sharp enough, we note that the chain \( Z^n \) is positive recurrent, which guarantees that the dependence \( p_\alpha \to dP_{\alpha,n}/dp_{0,n} \) is locally asymptotically normal \([121, 109]\). According to lemma 7.5, this implies that local prior predictive distributions based on \( n^{-1/2} \)-neighbourhoods of \( p_0 \) in \( \Theta \) are \( c_n \)-remotely contiguous to \( p_{0,n} \) for any rate \( c_n \), if the prior has full support. If we require that the prior density \( \pi' \) with respect to Lebesgue measure on \( S_N^W \) is continuous and strictly positive, then we see that there exists a constant \( \pi > 0 \) such that \( \pi'(\lambda) \geq \pi \) for all \( \lambda \in S_N^W \), so that for every \( n^{-1/2} \)-neighbourhood \( B_n \) of \( p_0 \), there exists a \( K > 0 \) such that \( \Pi(B_n) \geq b_n := Kn^{-1}\). Although local asymptotic normality guarantees remote contiguity at arbitrary rate, we still have to
make sure that \( c_n \to 0 \) in lemma 7.5, i.e. that \( a_n = o(b_n) \). Then the remark directly after theorem 8.10 shows that condition (ii) of said theorem is satisfied.

The above leads to the following conclusion concerning goodness-of-fit testing c.f. (7.25).

**Proposition 7.7.** Let \( X^n \) be a stationary, discrete time Markov chain on a measurable state space \((S, \mathcal{F})\). Choose a finite, measurable partition \( \alpha \) of \( S \) such that the Markov chain \( Z^n \) is ergodic. Choose a prior \( \Pi' \) on \( S^N \) absolutely continuous with respect to Lebesgue measure with a continuous density that is everywhere strictly positive. Assume that,
\[
(i) \quad n \lambda_2^2 S^2 / \log(n) \to \infty, \\
(ii) \quad \Pi(B \setminus B_n), \Pi(\Theta \setminus S_n) = o(n^{-(N^N/2)}), \\
(iii) \quad \max_{k,l} \Pi(V_{k,l} \setminus (V_{k,l,n} \cup V_{-k,l,n})) = o(n^{-(N^N/2)}).
\]

Then for any choice of \( \epsilon > 0 \), the Bayes factors \( F_n \) are consistent for \( H_0 \) versus \( H_1 \).

To guarantee ergodicity of \( Z^n \) one may use an empirical device, i.e. we may use an independent, finite-length realization of the random walk \( X^n \) to find a partition \( \alpha \) such that for all \( 1 \leq k, l \leq N \), we observe some \( m \)-step transition from \( l \) to \( k \). An interesting generalisation concerns a hypothesized Markov transition kernel \( P_0 \) for the process \( X^n \) and partitions \( \alpha_n \) (with projections \( p_0, \alpha_n \) as in (7.24)), chosen such that \( \alpha_{n+1} \) refines \( \alpha_n \) for all \( n \geq 1 \). Bayes factors then test a sequence of pairs of hypotheses (7.25) centred on the \( p_0, \alpha_n \). The arguments leading to proposition 7.7 do not require modification and the rate of growth \( N_n \) comes into the conditions of proposition 7.7.

Example 7.19 demonstrates the enhancement of the role of the prior as intended by the remark that closes the subsection on the existence of Bayesian test sequences in section 7.1: where testing power is relatively weak, prior mass should be scarce to compensate and where testing power is strong, prior mass should be plentiful. A random walk for which mixing does not occur quickly enough does not give rise to (7.27) and alternatives for which separation decreases too fast lose testing power, so the difference sets of proposition 7.7 are the hard-to-test parts of the parameter space and conditions (ii)–(iii) formulate how scarce prior mass in these parts has to be.

### 7.7 Confidence sets from credible sets

The Bernstein-von Mises theorem [166] asserts that the posterior for a smooth, finite-dimensional parameter converges in total variation to a normal distribution centred on an efficient estimate with the inverse Fisher information as its covariance, if the prior has full support. The methodological implication is that Bayesian credible sets derived from such a posterior can be reinterpreted as asymptotically efficient confidence sets. This parametric fact begs for the exploration of possible
non-parametric extensions but Freedman discourages us [96] with counterexamples (see also [57]) and concludes that: “The sad lesson for inference is this. If frequentist coverage probabilities are wanted in an infinite-dimensional problem, then frequentist coverage probabilities must be computed.”

In recent years, much effort has gone into calculations that address the question whether non-parametric credible sets can play the role of confidence sets nonetheless. The focus lies on well-controlled examples in which both model and prior are Gaussian so that the posterior is conjugate and analyse posterior expectation and variance to determine whether credible metric balls have asymptotic frequentist coverage (for examples, see Szabó, van der Vaart and van Zanten [211] and references therein). Below, we change the question slightly and do not seek to justify the use of credible sets as confidence sets; from the present perspective it appears more natural to ask in which particular fashion a credible set is to be transformed in order to guarantee the transform is a confidence set, at least in the large-sample limit.

In previous subsections, we have applied remote contiguity after the concentration inequality to control the $P_{\theta_n}^{\Pi}$-expectation of the posterior probability for the alternative $V$ through its $P_n^{\Pi_{\theta}}$-expectation. In the discussion of the coverage of credible sets that follows, remote contiguity is applied to control the $P_{\theta_n}^{\Pi}$-probability that $\theta_0$ falls outside the prospective confidence set through its $P_n^{\Pi_{\theta_n}}$-probability. The theorem below then follows from an application of Bayes’s rule (A.4). Credible levels provide the sequence $a_n$.

**Definition 7.5.** Let $(\Theta, \mathcal{F})$ with prior $\Pi$, denote the sequence of posteriors by $\Pi(\cdot|\cdot) : \mathcal{F} \times \mathcal{F}_n \to [0, 1]$. Let $\mathcal{D}$ denote a collection of measurable subsets of $\Theta$. A sequence of credible sets $(D_n)$ of credible levels $1 - a_n$ (where $0 \leq a_n \leq 1$, $a_n \downarrow 0$) is a sequence of set-valued maps $D_n : \mathcal{F}_n \to \mathcal{D}$ such that $\Pi(\Theta \setminus D_n(x)|x) = o(a_n)$ for $P_n^{\Pi_{\theta}}$-almost-all $x \in \mathcal{F}_n$.

**Definition 7.6.** For $0 \leq a \leq 1$, a set-valued map $x \mapsto C(x)$ defined on $\mathcal{F}$ such that, for all $\theta \in \Theta$, $P_{\theta}(\theta \notin C(x)) \leq a$, is called a confidence set of level $1 - a$. If the levels $1 - a_n$ of a sequence of confidence sets $C_n(x)$ go to 1 as $n \to \infty$, the $C_n(x)$ are said to be asymptotically consistent.

**Definition 7.7.** Let $D$ be a (credible) set in $\Theta$ and let $B = \{B(\theta) : \theta \in \Theta\}$ denote a collection of model subsets such that $\theta \in B(\theta)$ for all $\theta \in \Theta$. A model subset $C'$ is said to be (a confidence set) associated with $D$ under $B$, if for all $\theta \in \Theta \setminus C'$, $B(\theta) \cap D = \emptyset$. The intersection $C$ of all $C'$ like above equals $\{\theta \in \Theta : B(\theta) \cap D \neq \emptyset\}$ and is called the minimal (confidence) set associated with $D$ under $B$ (see Fig 7.1).

**Example 7.20** makes this construction explicit in uniform spaces and specializes to metric context.

**Theorem 7.6.** Let $\theta_0 \in \Theta$ and $0 \leq a_n \leq 1$, $b_n > 0$ such that $a_n = o(b_n)$ be given. Choose priors $\Pi_n$ and let $D_n$ denote level-$(1 - a_n)$ credible sets. Furthermore, for all $\theta \in \Theta$, let $B_n = \{B_n(\theta) \in \mathcal{F} : \theta \in \Theta\}$ denote a sequence such that,

(i) $\Pi_n(B_n(\theta_0)) \geq b_n$.
(ii) $P_{\theta_0, n} \prec b_n a_n^{-1} \Pi_n(B_n(\theta_0))$.

Then any confidence sets $C_n$ associated with the credible sets $D_n$ under $B_n$ are asymptotically consistent, i.e. for all $\theta_0 \in \Theta$,

$$P_{\theta_0, n}(\theta_0 \in C_n(X_n)) \to 1. \quad (7.28)$$

**Proof.** Fix $n \geq 1$ and let $D_n$ denote a credible set of level $1 - o(a_n)$, defined for all $x \in F_n \subset \mathcal{X}_n$ such that $\Pi_n(F_n) = 1$. For any $x \in F_n$, let $C_n(x)$ denote a confidence set associated with $D_n(x)$ under $B$. Due to definition 7.7, $\theta_0 \in \Theta \setminus C_n(x)$ implies that $B_n(\theta_0) \cap D_n(x) = \emptyset$. Hence the posterior mass of $B(\theta_0)$ satisfies $\Pi(B_n(\theta_0)|x) = o(a_n)$. Consequently, the function $x \mapsto 1\{\theta_0 \in \Theta \setminus C_n(x)\} \Pi(B(\theta_0)|x)$ is $o(a_n)$ for all $x \in F_n$. Integrating with respect to the $n$-th prior predictive distribution and dividing by the prior mass of $B_n(\theta_0)$, one obtains,

$$\frac{1}{\Pi_n(B_n(\theta_0))} \int 1\{\theta_0 \in \Theta \setminus C_n\} \Pi(B_n(\theta_0)|X_n) d\Pi_n \leq \frac{a_n}{b_n}.$$

Applying Bayes’s rule in the form (A.4), we see that,

$$P_{\theta_0, n}(\theta_0 \in \Theta \setminus C_n(X_n)) = \int P_{\theta_0, n}(\theta_0 \in \Theta \setminus C_n(X_n)) d\Pi_n(\theta|B_n) \leq \frac{a_n}{b_n}.$$

By the definition of remote contiguity, this implies asymptotic coverage c.f. (7.28).
7.7 Confidence sets from credible sets

Proof. (corollary 7.2)
Define \(a_n = \exp(-C'n\varepsilon_n^2)\), \(b_n = \exp(-Cn\varepsilon_n^2)\), so that the \(D_n\) are credible sets of level \(1 - o(a_n)\), the sets \(B_n\) of example 7.17 satisfy condition (i) of theorem 7.6 and \(b_n a_n^{-1} = \exp(cn\varepsilon_n^2)\) for some \(c > 0\). By (7.22), we see that condition (ii) of theorem 7.6 is satisfied. The assertion now follows.

This refutes Freedman’s lesson, showing that the asymptotic identification of credible sets and confidence sets in smooth parametric models (the main inferential implication of the Bernstein-von Mises theorem) generalises to the above form of asymptotic congruence in non-parametric models. The fact that this statement holds in full generality implies very practical ways to obtain confidence sets from posteriors, calculated, simulated or approximated. A second remark concerns the confidence levels of associated confidence sets. In order for the assertion of theorem 7.6 to be specific regarding the confidence level (rather than just resulting in asymptotic coverage), we re-write the last condition of theorem 7.6 as follows,

\[(ii') c_n^{-1} P_{\theta_0} \left( \frac{\theta}{b_n a_n^{-1}} \right) \}

so that the last step in the proof of theorem 7.6 is more specific; particularly, assertion (7.28) becomes,

\[P_{\theta_0} (\theta \in D_n(X^n)) = o(c_n),\]

i.e. the confidence level of the sets \(D_n(X^n)\) is \(1 - Kc_n\) asymptotically (for some constant \(K > 0\) and large enough \(n\)).

The following corollary that specializes to the i.i.d. situation is immediate (see example 7.21). Let \(\mathcal{P}\) denote a model of single-observation distributions, endowed with the Hellinger or total-variational topology.

**Corollary 7.2.** For \(n \geq 1\) assume that \((X_1, X_2, \ldots, X_n) \in \mathcal{X}^n \sim P_0^n\) for some \(P_0 \in \mathcal{P}\). Let \(\Pi_n\) denote Borel priors on \(\mathcal{P}\), with constant \(C > 0\) and rate sequence \(\varepsilon_n \downarrow 0\) such that (7.21) is satisfied. Denote by \(D_n\) credible sets of level \(1 - \exp(-C'n\varepsilon_n^2)\), for some \(C' > C\). Then the confidence sets \(C_n\) associated with \(D_n\) under radius-\(\varepsilon_n\)-Hellinger-enlargement are asymptotically consistent.

Note that in the above corollary,

\[\text{diam}_H(C_n(X^n)) = \text{diam}_H(D_n(X^n)) + 2\varepsilon_n,\]

\(P_0^n\)-almost surely. If, in addition to the conditions in the above corollary, tests satisfying (7.20) with \(a_n = \exp(-C'n\varepsilon_n^2)\) exist, the posterior is consistent at rate \(\varepsilon_n\) and sets \(D_n(X^n)\) have diameters decreasing as \(\varepsilon_n\), c.f. theorem 7.4. In the case \(\varepsilon_n\) is the minimax rate of convergence for the problem, the confidence sets \(C_n(X^n)\) attain rate-optimality [170]. Rate-adaptivity [118, 50, 211] is not possible like this because a definite, non-data-dependent choice for the \(B_n\) is required.
7.7.1 Credible/confidence sets in metric spaces

First we come back to the remark following proposition ??, concerning shrinking confidence balls centred on small-ball estimators.

Proposition 7.8. Let ...

Proof. Denote ...

When enlarging credible sets to confidence sets using a collection of subsets $B$ as in definition 7.7, measurability of confidence sets is guaranteed if $B(\theta)$ is open in $\Theta$ for all $\theta \in \Theta$.

Example 7.20. Let $\mathcal{G}$ be the Borel $\sigma$-algebra for a uniform topology on $\Theta$, like the weak and metric topologies of appendix ??, $\mathcal{G}$ denote a symmetric entourage and, for every $\theta \in \Theta$, define $B(\theta) = \{(\theta', \theta') \in W\}$, a neighbourhood of $\theta$. Let $D$ denote any credible set. A confidence set associated with $D$ under $B$ is any set $C'$ such that the complement of $D$ contains the $W$-enlargement of the complement of $C'$. Equivalently (by the symmetry of $W$), the $W$-enlargement of $D$ does not meet the complement of $C'$. Then the minimal confidence set $C$ associated with $D$ is the $W$-enlargement of $D$. If the $B(\theta)$ are all open neighbourhoods (e.g. whenever $W$ is a symmetric entourage from a fundamental system for the uniformity on $\Theta$), the minimal confidence set associated with $D$ is open. The most common examples include the Hellinger or total-variational metric uniformities, but weak topologies (like Prohorov’s or $\mathcal{B}_n$-topologies) and polar topologies are uniform too.

Example 7.21. To illustrate example 7.20 with a customary situation, consider a parameter space $\Theta$ with parametrization $\theta \mapsto \theta^*$, to define a model for i.i.d. data $X = (X_1, \ldots, X_n) \sim P_{\theta^*}$, for some $\theta^* \in \Theta$. Let $\mathcal{G}$ be the class of all pre-images of Hellinger balls, i.e. sets $D(\theta, \varepsilon) \subset \Theta$ of the form,

$$D(\theta, \varepsilon) = \{ \theta' \in \Theta : H(P_\theta, P_{\theta'}) < \varepsilon \},$$

for any $\theta \in \Theta$ and $\varepsilon > 0$. After choice of a Kullback-Leibler prior $\Pi$ for $\theta$ and calculation of the posteriors, choose $D_n$ equal to the pre-image $D(\hat{\theta}_n, \hat{\varepsilon}_n)$ of a (e.g. the one with the smallest radius, if that exists) Hellinger ball with credible level $1 - o(a_n)$, $a_n = \exp(-n\alpha^2)$ for some $\alpha > 0$. Assume, now, that for some $0 < \varepsilon < \alpha$, the $W$ of example 7.20 is the Hellinger entourage $W = \{(\theta, \theta') : H(P_\theta, P_{\theta'}) < \varepsilon \}$. Since Kullback-Leibler neighbourhoods are contained in Hellinger balls, the sets $D(\hat{\theta}_n, \hat{\varepsilon}_n + \varepsilon)$ (associated with $D_n$ under the entourage $W$), is a sequence of asymptotic confidence sets, provided the prior satisfies (6.15). If we make $\varepsilon$ vary with $n$, neighbourhoods of the form $B_n$ in example 7.17 are contained in Hellinger balls of radius $\varepsilon_{\alpha}$, and in that case,

$$C_n(X^n) = D(\hat{\theta}_n, \hat{\varepsilon}_n + \varepsilon_{\alpha}),$$

is a sequence of asymptotic confidence sets, provided that the prior satisfies (7.21).
7.8 Conclusions

We list and discuss the main conclusions of this chapter below.

**Frequentist validity of Bayesian limits**

There exists a systematic way of taking Bayesian limits into frequentist ones, if priors satisfy an extra condition relating true data distributions to localized prior predictive distributions. This extra condition generalises Schwartz’s Kullback-Leibler condition and amounts to a weakened form of contiguity, termed *remote contiguity*.

For example regarding consistency with *i.i.d.* data, Doob shows that a Bayesian form of posterior consistency holds without any real conditions on the model. To the frequentist, ‘holes’ of potential inconsistency remain, in null-sets of the prior. Remote contiguity ‘fills the holes’ and elevates the Bayesian form of consistency to the frequentist one. Similarly, prior-almost-surely consistent tests are promoted to frequentist consistent tests and Bayesian credible sets are converted to frequentist confidence sets.

**The nature of Bayesian test sequences**

The existence of a Bayesian test sequence is equivalent to consistent posterior convergence in the Bayesian, prior-almost-sure sense. In theorems above, a Bayesian test sequence thus represents the Bayesian limit for which we seek frequentist validity through remote contiguity. Bayesian test sequences are more abundant than the more familiar uniform test sequences. Aside from prior mass requirements arising from remote contiguity, *the prior should assign little weight where testing power is weak and much where testing power is strong*, ideally.

Example 7.19 illustrates the influence of the prior when constructing a test sequence. Aside from the familiar lower bounds for prior mass that arise from remote contiguity, existence of Bayesian tests also poses upper bounds for prior mass.

**Systematic analysis of complex models and datasets**

Although many examples have been studied on a case-by-case basis in the literature, the systematic analysis of limiting properties of posteriors in cases where the data is dependent, or where the model, the parameter space and/or the prior are sample-size dependent, requires generalisation of Schwartz’s theorem and its variations, which the formalism presented here provides.

To elaborate, given the growing interest in the analysis of dependent datasets gathered from networks (*e.g.* by webcrawlers that random walk linked webpages), or from time-series/stochastic processes (*e.g.* financial data of the high-frequency type), or in the form of high-dimensional or even functional data (biological, financial, medical and meteorological fields provide many examples), the development of new Bayesian methods involving such aspects benefits from a simple, insightful, systematic perspective to guide the search for suitable priors in concrete examples.

To illustrate the last point, let us consider consistent community detection in stochastic block models [74, 30]. Bayesian methods have been developed for consistent selection of the number of communities [117], for community detection with
a controlled error-rate with a growing number of communities [53] and for consistent community detection using empirical priors [210]. A moment’s thought on the discrete nature of the community assignment vector suggests a sequence of uniform priors, for which remote contiguity (of \( B_n = \{ P_{0,n} \} \)) is guaranteed (at any rate) and prior mass lower bounded by \( b_n = K_n!K_n^{-n} \) (where \( K_n \) is the number of communities at ‘sample size’ \( n \)). It would be interesting to see under which conditions a Bayesian test sequence of power \( a_n = o(b_n) \) can be devised that tests the true assignment vector versus all alternatives (in the sparse regime [61, 2, 178]). Rather than apply a Chernoff bound like in [53], one would probably have to start from the probabilistic [178] or information-theoretic [2] analyses of respective algorithmic solutions in the (very closely related) planted bi-section model. If a suitably powerful test can be shown to exist, theorem 7.4 proves frequentist consistency of the posterior.

**Methodology for uncertainty quantification**

Use of a prior that induces remote contiguity allows one to convert credible sets of a calculated, simulated or approximated posterior into asymptotically consistent confidence sets, in full generality. This extends the main inferential implication of the Bernstein-von Mises theorem to non-parametric models without smoothness conditions.

The latter conclusion forms the most important and practically useful aspect of this book.

### 7.9 Exercises [EMPTY]
Chapter 8
Consistent tests and model selection

8.1 Asymptotic testability

The question, “Which pairs of model subsets can be told apart asymptotically and which cannot?”, is not just of direct practical importance (e.g. for model selection with large amounts of data) and of essential value in the development of theory. It is also a fundamental matter at the heart of statistics: which model questions have a truly statistical nature (that is, questions answerable from the data), and which do not? Of course, there are two versions of this question, one that requires only a proof of the existence of tests, and another that asks for the actual construction of such tests. In this chapter, the existence question is answered first and the constructive question is considered as in [146], promoting the existence result to a guarantee that posteriors achieve the correct conclusion, also for the frequentist.

To make the issue precise, consider a situation where we observe i.i.d. data $X^n \sim P^n, (n \geq 1)$, with a model $\mathcal{P}$ such that $P \in \mathcal{P}$. Suppose that, for disjoint $B, V \subset \mathcal{P}$, we are interested whether, $H_0 : P \in B$, or $H_1 : P \in V$.

In an asymptotic, symmetric testing procedure, one requires a sequence of test functions $(\phi_n)$ with type-I and type-II error probabilities (resp. $P^n \phi_n$ for $P \in B$ and $P^n(1 - \phi_n)$ for $P \in V$) that go to zero. Equivalently (see [181, 86] and proposition 8.4) one requires existence of some testing procedure with the following property,

A testing procedure that chooses for $B$ or $V$ based on $X^n$ for every $n \geq 1$, has property (D) if it is wrong only a finite number of times with $P^n$-probability one.

Property (D) is referred to as “discernibility” in [64, 181, 86] and it is also the basis for the tests in many other publications, for example [56, 65].

To do justice to the level of generality that the title promises, a real answer requires various things: ideally, there should be no restrictions on the model at all; furthermore, the answer should characterise the pairs $B, V$ for which test sequences
exist, as well as the pairs for which this is not the case, preferably in the form of an equivalence: given any model $\mathcal{P}$, whether $P$ belongs to $B$ or to $V$ can be tested asymptotically, if and only if, etcetera. Answers depend crucially on the formal/philosophical framework: a Bayesian who gives his answer based on posterior odds or a Bayes factor, and who disregards potential prior null-sets of exceptions, answers this question differently from a frequentist, who formulates his answer in terms of test functions and insists on asymptotic consistency for all points in the model (or even uniformly). These distinctions lead to differing answers to the question “What is asymptotically testable and what is not?”, and hence, to differing notions of answerable and non-answerable statistical questions.

In section 8.2, we consider three forms of asymptotic testability: uniform testability, pointwise testability and Bayesian testability. In subsequent sections we prove for each form an equivalence characterising pairs $B, V$ for which consistent tests exist: in section 8.3 equivalent formulations of uniform testability are given; in section 8.4 we characterise hypotheses that are pointwise testable; and in section 8.5, it is shown that Bayesian tests exist for a very wide variety of hypotheses.

As stressed already, we do not restrict attention to subclasses of models, the model choice is left completely free. (We make one exception: in theorem 8.6 we require a dominated model, see the discussion in section 8.7.) Characterisations of testability are formulated in terms of conditions on the sets $B, V$ only. Otherwise, we would not characterise testability itself but how it manifests in subclasses. Of course, it is possible that a hypotheses $B, V$ are not testable versus its complement in a large model $\mathcal{P}$, while becoming testable when $\mathcal{P}$ is restricted. That type of condition does not play a role in the theorems below; useful general restrictions (like completeness, metric totally-boundedness, or weak-relative-compactness) are accommodated in corollaries. Such restrictions form connections with previous work and motivate examples.

### 8.1.1 Some examples and unexpected answers

Intuition regarding the existence problem of asymptotic tests is greatly helped by some examples that typify the nature of possible answers: distinctions between smoothness classes for a regression function $f : X \to \mathbb{R}$:

$$H_0 : f \in C^1(X \to \mathbb{R}) , \quad H_1 : f \in C^2(X \to \mathbb{R}) ,$$

cannot be tested consistently according to the frequentist. However, to the Bayesian using the posterior, smoothness classes are asymptotically testable without any reservations, for prior-almost-all points in the model. To mention another instance, the frequentist cannot test to distinguish asymptotically between classes of densities $p$ on $\mathbb{R}$ with or without a second moment:

$$H_0 : \int x^2 \ p(x) \ dx < \infty , \quad H_1 : \int x^2 \ p(x) \ dx = \infty .$$
That simple fact implies no statistician dealing with unbounded data can ever use the central limit theorem with asymptotic certainty that it applies for the true distribution \( P \) of the data. Bayesians can make the distinction asymptotically (but, again, prior-almost-surely). Similarly, Bayesians can test consistently whether a distribution on \( \mathbb{R} \) is compactly supported, whether its Lebesgue density is square-integrable, etcetera, distinctions that are not testable for frequentists [65].

The intricacy of the question is underlined further by the unexpected answer to Cover’s rational mean problem [56]: for an i.i.d. sequence of coin-flips \( X_1, X_2, \ldots \) (with all \( X_i \) distributed marginally Bernoulli-\( p \) with \( p \in [0, 1] \)), consider the hypotheses:

\[
H_0 : p \in [0, 1] \cap \mathbb{Q}, \quad H_1 : p \in [0, 1] \setminus \mathbb{Q}.
\]  

(8.1)

Rather surprisingly, Cover shows that there exists a test sequence \( \phi_n(X_1, \ldots, X_n) \) that goes to one if \( p \in [0, 1] \cap \mathbb{Q} \), and to zero for Lebesgue-almost-all \( p \in [0, 1] \setminus \mathbb{Q} \). It is not possible to find a test sequence for Cover’s problem without such an exceptional null-set (see corollary 8.4.2). However, it is possible to restrict the model to enable testability: Dembo and Peres [64] show that there exist asymptotically consistent tests for,

\[
H_0 : p \in [0, 1] \cap \mathbb{Q}, \quad H_1 : p \in [0, 1] \cap \sqrt{2} + \mathbb{Q},
\]  

(8.2)

without measure-theoretic exceptions. But one does not have to restrict to countable hypotheses to find testability for apparently deeply intertwined hypotheses: example 8.10 shows it is possible to test whether \( p \) lies in the Cantor set \( C \) or not,

\[
H_0 : p \in C, \quad H_1 : p \in [0, 1] \setminus C,
\]  

(8.3)

It is noted that \( C \) is zero-dimensional and nowhere-dense, while both \( C \) and its complement are uncountable. And although \( C \) has Lebesgue measure zero, there are Cantor sets of non-zero Lebesgue measure for which testability also holds. So if testability is ruined by certain forms of denseness but not for others, and maintained for self-similar sets like \( C \), what does the distinction depend on?

8.1.2 Testability over the decades

Of course the question has a long history: the first attempts to answer general questions on testability appear already in the 1950’s: Hoeffding and Wolfowitz [119] give sufficient conditions that are also necessary in some cases (see also, [156]). Kraft [148] studied consistent tests for families of general, dependent data distributions and gives a separation condition in terms of the separation of convex hulls of finite-dimensional projections, much like the Hahn-Banach theorem and its specializations suggest [46]. Berger [17] gives necessary and sufficient conditions for the existence of uniformly consistent tests, and subsequent work [18, 156] extends the approach to pointwise consistent estimation problems. It is noted that the present work is inspired first and foremost on the Le Cam-Schwartz theorem [156], which provides
necessary and sufficient conditions for the existence of uniform and pointwise tests, in terms of a particular uniformity we denote $U_\infty$ with associated topology $T_\infty$. (Appendix ?? offers a comparison of $T_\infty$ with other, better-known model topologies.) In the form applicable to pointwise testing, the Le Cam-Schwartz theorem states that,

**Theorem 8.1.** (Le Cam-Schwartz, 1960) Let $\mathcal{P}$ be a model for i.i.d. data $X^n$ with disjoint subsets $B, V$. There exist (uniformly) consistent tests for $B$ versus $V$, if and only if, there exists a sequence of $U_\infty$-uniformly continuous functions $\psi_n : \mathcal{P} \to [0, 1]$ such that,

$$\psi_n(P) \to 1_{V}(P),$$

(uniformly) over all $P \in \mathcal{P}$.

So the Le Cam-Schwartz theorem provides the definitive answer to our question. However, its formulation is in terms of a uniformity $U_\infty$ that is “rather inaccessible” [162] (see [64, 181, 86] for more detailed comments), and it is perhaps this inaccessibility that explains why the entire body of subsequent work on the subject mentions the Le Cam-Schwartz theorem but does not relate to it at any formal level. Most sensitive to the argument put forth by Le Cam and Schwartz appears to be the insightful work of Ermakov [86], which departs from necessary conditions for the existence of pointwise consistent tests in terms of uniformly consistent tests. However, a weakly compact, dominated model is required for Ermakov’s results and Prokhorov’s weak topology rather than Le Cam-Schwartz’s uniformity $U_\infty$ is used to formulate testability conditions.

A separate but related historical line of research originates from Cover’s rational mean problem [56], and answers Cover’s specific (but prototypical, see theorem 8.5) question from the probabilistic point of view (see also, [187]). As a second inspirational reference for this work, we mention Dembo and Peres [64], who show that the limited version of Cover’s problem in (8.2) has a solution and subsequently prove the following theorem.

**Theorem 8.2.** (Dembo and Peres, 1995) Let $\mathcal{P}$ be a model dominated by Lebesgue measure $\mu$ for i.i.d. data $X^n$. Model subsets $B, V$ that are contained in disjoint countable unions of closed sets for Prokhorov’s weak topology have tests with property (D). If there exists an $\alpha > 1$ such that $\int (dP/d\mu)^\alpha d\mu < \infty$ for all $P \in \mathcal{P}$, then the converse is also true.

Note the recurrence of weakly compact, dominated models with Prokhorov’s weak topology. Kulkarni and Zeitouni [150] accept Cover’s exceptional null-set and consider the question when such tests (which we call Bayesian, see definition 8.3) exist in more general setting. Nobel [181] notices that the approach of Dembo and Peres can be extended from i.i.d. setting to a framework where the data is dependent, e.g. to test between disjoint families of (uniformly) ergodic processes. Ermakov’s work also appears to be inspired by the results of Dembo and Peres.

If one departs from the strictly constructive, statistical perspective (e.g. in non-parametric density estimation), first of all, there are many solutions that are specific to model and hypotheses presented, often involving specific test-statistics and
critical regions, roughly following the classical approach of [154]. If we restrict attention to (a non-exhaustive list of) references that aim to answer the more general question, it is worth mentioning Donoho (1988) [72], who discusses non-parametric confidence sets and testing of hypotheses for aspects of the density of the data in dominated, non-parametric models. Similar in intention, and a third major inspiration for this chapter, is Devroye and Lugosi (2003) [65] who construct solutions in many diverse and practical examples of non-parametric testing problems for densities, based largely on contemporary methods of kernel estimation.

8.1.3 The forms that answers take

Given the rather intricate examples of subsection 8.1.1, one wonders which expectations one should have regarding the forms in which answers to the testability question are formulated. Based on the examples of pointwise testing in subsection 8.1.1, it is clear that model topology plays a central role in characterising which disjoint pairs \( B, V \) are testable and which are not. Exactly which topology we deploy here, is prescribed by the necessary and sufficient conditions that the Le Cam-Schwartz theorem formulates: we are obliged to view the model as a uniform space with the uniformity \( \mathcal{U}_\infty \). (In the examples of subsection 8.1.1, the topology \( \mathcal{T}_\infty \) coincides with the usual topology of \([0,1]\).) This rather technical starting point is not a choice but an imperative (if we insist on total freedom of model choice); only by setting model conditions (e.g. like uniform integrability, as in [64, 181, 86] and corollary 8.5) can this be avoided.

But having decided which topology is relevant, we also need to determine what type of topological condition we expect to determine testability of disjoint pairs \( B, V \). For uniform testability of disjoint \( B, V \), it is necessary and sufficient (see theorem 8.4) that \( B \) and \( V \) are \( \mathcal{U}_\infty \)-uniformly separated: there exists an entourage \( U \) that does not meet \( B \times V \cup V \times B \). Regarding pointwise testability one expects countable unions of weakly closed sets to be important, based on [64, 181, 86]; as we shall see, disjoint \( B, V \) that are pointwise testable can be characterised as sets that are both countable unions of closed sets and countable intersections of open sets in \( B \cup V \). This condition holds for the Cantor set \( C \) and its complement in \([0,1]\) and for the countable sets \([0,1] \cap \mathbb{Q}\) and \([0,1] \cap \sqrt{2} + \mathbb{Q}\), but not for \([0,1] \cap \mathbb{Q}\) and its complement in \([0,1]\).

Bayesian testability is different and forms the constructive contribution. Bayesian testability does not fit the formulation of the Le Cam-Schwartz theorem and, as such, escapes its topological imperatives. The existence of a Bayesian test sequence is equivalent to the consistency of posterior odds or Bayes factors (see theorem 8.7), at least, if one is willing to permit prior null-sets of exceptions. This presents the opportunity to promote mere existence proofs for (Bayesian) testability, to constructive proofs in the sense that they imply an actual way to perform the test based on the data, using the posterior. This resolves the matter of testability for the Bayesian, but for the frequentist there remains the rather unwelcome possibility of exceptional
null-sets [93]. To bridge the discrepancy, two things are required [146]: a Bayesian test sequence with known testing power and a prior that induces a property termed remote contiguity for local prior predictive distributions. Then the Bayesian conclusion that the posterior provides a consistent test sequence remains valid for the frequentist, that is, without exceptional null-sets. Using a generality concerning the testing power of uniform test sequences (see proposition 8.1), and remote contiguity as it applies for Kullback-Leibler priors, we indicate a practical way to perform consistent, frequentist model selection with posteriors, and demonstrate how to use it in two model selection problems, selection of the number of clusters in a clustering problem, and selection of a directed acyclical graph in a graphical model.

8.2 Existence of test sequences

Let the model $\mathcal{P}$ be a collection of distributions $P$ on a measurable space $(\mathcal{X}, \mathcal{B})$, to model i.i.d. samples $X^n = (X_1, X_2, \ldots, X_n) \in \mathcal{X}^n, X_n \sim P^n$. The relevant model topology is the (subspace) topology $\mathcal{T}_m$ defined in definition ???. We consider two disjoint model subsets $B, V$ and wonder whether there is a way to tell whether the true distribution of the data lies in $B$ or in $V$ with asymptotic certainty. More particularly, we wonder whether there exists a sequence of test functions $\phi_n : \mathcal{X}^n \to [0, 1]$ that converge to one or to zero, depending on $P \in B$ or $P \in V$ (in probability/expectation or almost-surely, see proposition 8.4). Given a topological space $X$, we say that the testing problem has a (uniform) representation on $X$, if there exists a $\mathcal{T}_\infty$-(uniformly-)continuous surjective map $f : B \cup V \to X$ such that $f(B) \cap f(V) = \emptyset$. Given a Hausdorff topological space $\Theta$, we say that the model is parametrized by $\Theta$, if there exists a $\mathcal{T}_\infty$-continuous bijection $P : \Theta \to \mathcal{P}$ (i.e. for every $m \geq 1$ and measurable $f : \mathcal{P}^m \to [0, 1]$, the map $\theta \mapsto P^m_\theta f$ is continuous). This condition is satisfied quite easily, for example, it is weaker than continuity with respect to the total-variational topology (see proposition C.4). It does not imply that $\Theta$ and $\mathcal{P}$ are homeomorphic, unless $\Theta$ is compact. If $\Theta$ is compact and $P$ is $\mathcal{T}_1$-continuous (see proposition C.3), then $\mathcal{P}$ is parametrized by $\Theta$ and $P$ is a homeomorphism. When considering a represented testing problem on a parametrized model with $X = \Theta, \Theta$ and the model are homeomorphic.

Test sequences come in various kinds, e.g. uniform or pointwise, or Bayesian in nature.

**Definition 8.1.** We say that $(\phi_n)$ is a *uniform test sequence* for $B$ versus $V$, if,

$$
\sup_{P \in B} P^n \phi_n \to 0, \quad \sup_{Q \in V} Q^n (1 - \phi_n) \to 0.
$$

(8.5)

Existence of a uniform test sequence for $B$ versus $V$ implies the existence of a uniform test sequence of exponential power (see proposition 8.1), i.e. (8.5) implies there exist a test sequence $\psi_\alpha$ whose sum of type-I and type-II errors goes to zero exponentially fast,
8.2 Existence of test sequences

\[
\sup_{P \in B} P^n \phi_n + \sup_{Q \in V} Q^n (1 - \phi_n) \leq e^{-nD},
\]

for some constant \(D > 0\). This fact is exploited in section 8.7, where it is used together with a suitable prior to do demonstrate model-selection with Bayes factors, in a constructive way, while satisfying frequentist consistency criteria.

**Definition 8.2.** The \( (\phi_n) \) are a pointwise test sequence for \( B \) versus \( V \), if,

\[
\phi_n(X^n) \xrightarrow{P} 0, \quad \phi_n(X^n) \xrightarrow{Q} 1,
\]

(8.6)

for all \( P \in B \) and \( Q \in V \).

Existence of a pointwise test sequence for \( B \) versus \( V \) is equivalent to the existence of test sequence with property (D), (see proposition 8.4).

Aside from these two frequentist notions of testability, we also consider a version of the pointwise test that is strictly Bayesian, because it leaves room for a prior-null-set of exceptions [56, 150, 146].

**Definition 8.3.** Let \( (\mathcal{P}, \mathcal{G}) \) be a measurable space with prior \( \Pi \) and assume \( B, V \in \mathcal{G} \). We say that \( (\phi_n) \) is a Bayesian test sequence for \( B \) versus \( V \) (under \( \Pi \)), if,

\[
\phi_n \xrightarrow{P} 0, \quad \phi_n \xrightarrow{Q} 1,
\]

(8.7)

for \( \Pi \)-almost-all \( P \in B \) and \( Q \in V \).

The goal of this chapter is to characterize the existence of the test sequences with as much precision and in as much detail as possible, for the three definitions 8.1–8.3. We require an “accessible” form, that is, firstly we insist on easy illustration with a wide variety of examples and counterexamples, and secondly, that we elevate results of existence to constructive results (by applying the methods of [146]).

### 8.2.1 The Le Cam-Schwartz theorem

The basis for sections 8.3 and 8.4 is the Le Cam-Schwartz theorem. The following theorem is the Le Cam-Schwartz theorem, restated in test-specific form. Below \( \mathcal{F} \) denotes a increasingly directed collection of model subsets (for any finite subset \( \{F_1, \ldots, F_m\} \subset \mathcal{F} \), there exists an \( F \in \mathcal{F} \) such that \( F_1 \cup \ldots \cup F_m \subset F \). Examples: \( \mathcal{F} = \{\mathcal{P}\}; \mathcal{F} \) consists of all finite subsets of \( \mathcal{P} \); \( \mathcal{F} \) consists of all compact subsets of \( \mathcal{P} \), etcetera).

**Theorem 8.3.** (Le Cam-Schwartz, 1960) Let \( \mathcal{P} \) with hypotheses \( B, V \) \( B \cap V = \emptyset \) be given. There exists an \( \mathcal{F}\)-uniform test sequence for \( B \) versus \( V \), if and only if, there exists a sequence \( (\psi_n) \) of \( \mathcal{U}_n\)-uniformly continuous functions \( \psi_n : B \cup V \to [0, 1] \) such that,

\[
\sup_{F \in \mathcal{F}} |\psi_n(P) - 1_V(P)| \to 0,
\]

(8.8)
for every $F \in \mathcal{F}$.

Proof. (See [156] and also section 17.5 of [162].) If there exists an $\mathcal{F}$-uniform test sequence $(\phi_n)$ for $B$ versus $V$, then the functions $\psi_n : B \cup V \rightarrow [0,1],$

$$\psi_n(P) = P^n \phi_n(X^n),$$

are $\mathcal{U}_m$-uniformly continuous and converge $\mathcal{F}$-uniformly to $1_V$ on $B \cup V$. Conversely, suppose first that $\psi_n = \psi$ for some $\mathcal{U}_m$-uniformly continuous $\psi : B \cup V \rightarrow [0,1]$. Let $\varepsilon > 0$ be given. There exist $m, J \geq 1$, $\delta > 0$ and $f_j : \mathcal{X}^m \rightarrow [0,1]$ $(1 \leq j \leq J)$, such that for all $P, Q \in B \cup V$,

$$\rho(P, Q) := \max_{1 \leq j \leq J} |P^n f_j - Q^n f_j| < \delta,$$

implies that $|\psi(P) - \psi(Q)| < \varepsilon$. Define $M$ to be the smallest integer greater than $9J\varepsilon^{-1}\delta^{-2}$ and $\mathbb{P}_M f_j = M^{-1}\sum_{i=1}^M f_j(X_i)$, for random $X_1, \ldots, X_M \in \mathcal{X}^m$. Because any probability model is pre-compact for the uniform structure $\mathcal{U}_m$, there exist $L \geq 1$ and $\{Q_1, \ldots, Q_L\}$ such that, for all $P \in B \cup V$,

$$\min_{1 \leq j \leq L} \rho(P, Q_j) < \frac{1}{4}\delta.$$

Let $\hat{Q}_M$ denote the minimizer of $P \mapsto \rho(\mathbb{P}_M, P)$ over $\{Q_1, \ldots, Q_L\}$. For any $P$ and $l$ such that $\rho(P, Q_l) < \frac{1}{4}\delta$, we have,

$$\rho(\hat{Q}_M, P) \leq \rho(\mathbb{P}_M, P) + \rho(\mathbb{P}_M, \hat{Q}_M) \leq \rho(\mathbb{P}_M, P) + \rho(\mathbb{P}_M, Q_l) \leq 2\rho(\mathbb{P}_M, P) + \frac{1}{4}\delta.$$

For any $P \in B \cup V$, Chebyshev’s inequality gives,

$$P^M M(\rho(\mathbb{P}_M, P) \geq \frac{1}{4}\delta) \leq \sum_{j=1}^J P^M M(\rho(\mathbb{P}_M, f_j) \geq \frac{1}{4}\delta) \leq \sum_{j=1}^J \frac{9}{\delta^2} \text{Var}(\mathbb{P}_M f_j) \leq \frac{9J}{\delta^2 M} < \varepsilon.$$

Conclude that for all $P \in B \cup V$, $P^M M(\rho(\hat{Q}_M, P) \geq \varepsilon) \geq 1 - \varepsilon$. This proves the following intermediate result: for every $\varepsilon > 0$ and uniformly continuous $\psi$, there exists a sequence of estimators $(\hat{Q}_n)$ which satisfy,

$$\sup_{P \in B \cup V} P^n |\psi(\hat{Q}_n) - \psi(P)| < \varepsilon, \quad (8.9)$$

for large enough $n$. Generalizing to the sequential case of uniformly continuous $(\psi_k)$ satisfying (8.8), let a sequence $(\varepsilon_k)$ be given that decreases to zero. For every $k \geq 1$, let $(\hat{Q}_{k(n)})$ denote the estimator sequence that satisfies (8.9) for $\psi_k$ and $\varepsilon_k$. By traversing the sequences labelled with $k$ slowly enough with increasing $n$, we can guarantee that,

$$\sup_{P \in B \cup V} P^n |\psi(\hat{Q}_{k(n)}(n)) - \psi_k(n)(P)| < \varepsilon_k(n).$$

Combined with assumption (8.8), we see that,
A very natural question concerns conditions under uniform test sequences exists [17, 187]. Let us first establish the following useful equivalence [201, 160, 10, 63].

Proposition 8.1. Let \( \mathcal{P} \) be a model with hypotheses \( B \) and \( V \), \( B \cap V = \emptyset \). The following are equivalent:

1. there exists a uniform test sequence \( (\phi_n) \) such that,
   \[
   \sup_{P \in B} P^n \phi_n \to 0, \quad \sup_{Q \in V} Q^n (1 - \phi_n) \to 0,
   \]

2. there exists a test sequence \( (\phi_n) \) and a constant \( D > 0 \) such that,
   \[
   \sup_{P \in B} P^n \phi_n \leq e^{-nD}, \quad \sup_{Q \in V} Q^n (1 - \phi_n) \leq e^{-nD}.
   \]

This fact can be exploited, for example, in Bayesian model selection, and consequently, in frequentist model selection with posteriors as well if a Kullback-Leibler prior is used; \textit{c.f.} proposition 8.1, existence of a uniform test implies existence of an exponentially powerful uniform test, which is enough to compensate for prior-mass lower-bounds for Kullback-Leibler neighbourhoods, \textit{e.g.} through remote-contiguity. According to theorem 4.8 of [146] posterior odds or Bayes factors then select the correct model consistently.
Here, we derive necessary and sufficient conditions for the existence of a uniform test sequence, in terms of the uniformity \( \mathcal{U}_\infty \).

**Definition 8.4.** Let \( (\mathcal{P}, \mathcal{U}) \) be a model with uniformity. We say that two model subsets \( B \) and \( V \) are uniformly separated by \( \mathcal{U} \) if there exists an entourage \( U \in \mathcal{U} \) such that for every \( P, Q \in \mathcal{P}, (P, Q) \in U \) implies that either \( P, Q \in B \), or \( P, Q \in V \).

**Theorem 8.4.** Let \( \mathcal{P} \) be a model and let \( B, V \) be model subsets. The following are equivalent:

(i.) there exists a uniform test sequence \( (\phi_n) \) for \( B \) versus \( V \), c.f. (8.5);

(ii.) the indicator \( 1_V : B \cup V \rightarrow \{0, 1\} \) is \( \mathcal{U}_\infty \)-uniformly continuous;

(iii.) the subsets \( B \) and \( V \) are uniformly separated by \( \mathcal{U}_\infty \).

**Proof.** Assume that there exists a uniform test sequence \( (\phi_n) \) for \( B \) versus \( V \). Define the \( \mathcal{U}_\infty \)-uniformly continuous functions \( \psi_n : \mathcal{P} \rightarrow [0, 1], \psi_n(P) = P^* \phi_n \) and note that the difference \( |\psi_n(P) - 1_V(P)| \) goes to zero uniformly over \( B \cup V \). So, for every \( \varepsilon > 0 \) there exist an \( N \geq 1 \) such that for all \( n \geq N \), \( \sup_{P} |\psi_n(P) - 1_V(P)| < \varepsilon / 3 \) and an entourage \( W \in \mathcal{U}_\infty \) such that for all \( (P, Q) \in W \), \( |\psi_n(P) - \psi_n(Q)| < \varepsilon / 3 \). Therefore,

\[
|1_V(P) - 1_V(Q)| \leq |1_V(P) - \psi_n(P)| + |\psi_n(P) - \psi_n(Q)| + |\psi_n(Q) - 1_V(Q)| < \varepsilon,
\]

for all \( (P, Q) \in V \). To show that (ii.) implies (i.), choose \( \psi_n = 1_V \). The equivalence of (ii.) to (iii.) follows directly from the definition of \( \mathcal{U}_\infty \)-uniform continuity of \( 1_V : \mathcal{P} \rightarrow \{0, 1\} \).

To expand on formulation (iii.), \( B \) and \( V \) are uniformly separated by \( \mathcal{U}_\infty \), if and only if, there exist \( J, m \geq 1 \), \( \varepsilon > 0 \) and bounded, measurable functions \( f_1, \ldots, f_J : \mathcal{X}^m \rightarrow [0, 1] \) such that,

\[
\max_{1 \leq j \leq J} |P^m f_1 - Q^m f_1| < \varepsilon,
\]

implies that either \( P, Q \in B \), or \( P, Q \in V \). If the model is \( \mathcal{U}_\infty \)-compact, \( m = 1 \) suffices.

**Proposition 8.2.** Let \( \mathcal{P} \) be a model and let \( B, V \) be model subsets with \( \mathcal{U}_\infty \)-closures \( \mathcal{B} \) and \( \mathcal{V} \). If \( B \) and \( V \) are uniformly separated by \( \mathcal{U}_\infty \), then \( \mathcal{B} \cap \mathcal{V} = \emptyset \). If \( \mathcal{P} \) is relatively \( \mathcal{U}_\infty \)-compact, the converse is also true.

**Proof.** Suppose that there exists a \( P \in \mathcal{B} \cap \mathcal{V} \). Let \( W \in \mathcal{U}_\infty \) be any entourage. There exists an entourage \( W' \in \mathcal{U}_\infty \) such that \( W' \circ W' \subseteq W \). (Recall that \( W' \circ W' \) denotes the collection of all pairs \( (P, Q) \in \mathcal{P} \times \mathcal{P} \) for which there exists an \( R \in \mathcal{P} \) such that \( (P, R) \in W' \) and \( (R, Q) \in W' \); more generally, see [42]!) The sets \( U_1 = \{ P' \in \mathcal{P} : (P, P') \in W' \} \) and \( U_2 = \{ P' \in \mathcal{P} : (P', P) \in W' \} \) are neighbourhoods of the point \( P \), so \( U_1 \cap B \neq \emptyset \) and \( U_2 \cap V \neq \emptyset \). Pick \( P_1 \in U_1 \cap B \) and \( P_2 \in U_2 \cap V \). Then \( (P_1, P_2) \in W' \circ W' \subseteq W \) so that \( W \cap B \times V = \emptyset \), i.e., \( W \) does not separate \( B \) from \( V \) uniformly. Conversely, assume that the \( \mathcal{U}_\infty \)-closure \( \overline{\mathcal{P}} \) of \( \mathcal{P} \) is \( \mathcal{U}_\infty \)-compact and that \( B \) and \( V \) are not uniformly separated by \( \mathcal{U}_\infty \), that is, for every \( W \in \mathcal{U}_\infty \), there exists a pair \( (P_W, Q_W) \in B \times V \cap W \). The collection \( \{(P_W, Q_W) : W \in \mathcal{U}_\infty \} \) forms a net in \( B \times V \). By the compactness of \( \overline{\mathcal{P}} \) there exists a convergent subnet \( (P_W', Q_W') \) with a limit \((P, P)\) on the diagonal of \( \overline{\mathcal{P}} \times \overline{\mathcal{P}} \). So there exists \( P \in \overline{\mathcal{P}} \) and nets \((P'_W) \subseteq B \) and \((Q'_W) \subseteq V \) such that \( P'_W \rightarrow P \) and \( Q'_W \rightarrow P \). This shows that \( P \in \mathcal{B} \cap \mathcal{V} \).
If $\mathcal{P}$ is not compact in $\mathcal{T}_m$, it is possible for two closed subsets to have no points in common, yet fail to be uniformly separated. (For comparison: two closed subsets of (the non-compact sets) $\mathbb{R}^k$ for $k \geq 1$ can have empty intersections but be at distance zero, for example $\{(x,y) \in (0,\infty) \times \mathbb{R} : y \geq 1/x\}$ and $\{(x,y) \in (0,\infty) \times \mathbb{R} : y \leq -1/x\}$ in $\mathbb{R}^2$. Note that this is not possible if we replace $\mathbb{R}^k$ by some compact subset.)

**Example 8.1.** Hellinger tests (from the minimax theorem)

It is shown in [160] that the centre of a total-variational ball cannot be tested with uniform testing power against the interior of said ball and this negative result stays true if we change from an $L_1$- to $L_p$-metrics for $p > 1$.

### 8.4 Pointwise testability

This section focuses on pointwise testability. Between Bayesian testability (which requires only Borel measurability) and uniform testability (which requires no less than uniform separation), pointwise testability is an interesting test-case where the question of testability may find its most natural or balanced answer.

In the first subsection, we consider pointwise testability according to definition 8.2 in models that are not dominated. Subsequent subsections focus on hypotheses that are asymptotically indistinguishable by statistical testing, and on a characterization of testable hypotheses in dominated models.

To start with a situation for which most have more intuition, consider the case of a model in which consistent estimators $\hat{P}_n : \mathcal{X}^n \to \mathcal{P}$ exists, $n \geq 1$. Here $\mathcal{P}$ is a set of single-observation distributions $P$, assumed Hausdorff in some topology $\mathcal{T}$. Consistency says that for every $P \in \mathcal{P}$ and neighbourhood $U$ of $P$, we have $P^n(\hat{P}_n \in U) \to 1$. Given two open hypotheses $B,V \subset \mathcal{P}$ with $B \cap V = \emptyset$, define $\phi_n(X^n) = 1\{\hat{P}_n \in V\}$ and note that for any $P \in B$, $B$ is a neighbourhood of $P$ so $P^n(\phi_n = 1) \to 1$. So $(\phi_n)$ is a pointwise test sequence for $B$ versus $V$. If we restrict attention to $\mathcal{P}' = B \cup V$, $B$ and $V$ are complementary, so that $B$ and $V$ are both clopen sets, i.e. $B,V$ both lie in the first ambiguous class $\Delta_0^1(\mathcal{P}')$. We summarize with the following proposition.

**Proposition 8.3.** If $P \in \mathcal{P}$ can be estimated consistently and $B$ is clopen, there exist pointwise tests for $B$ versus its complement.

So clopeness is sufficient if we can estimate, but is it also necessary? And which topologies on $\mathcal{P}$ are strong enough? Below we shall see that the topology $\mathcal{T}_m$ imposes itself and that, in fact, the requirement that $B$ is both an $F_\sigma$ and a $G_\delta$ (i.e. that $B$ lie in the second ambiguous class $\Delta_0^2(\mathcal{P}')$) is necessary for the existence of pointwise tests, and in complete models, also sufficient.
8 Consistent tests and model selection

8.4.1 Pointwise testability in non-dominated models

As a general, introductory remark, let us first prove that there is no difference between pointwise testability in its “almost-sure”, “in-probability” and “in-expectation” versions.

Proposition 8.4. Let \( \mathcal{P} \) be a model with hypotheses \( B \) and \( V \), \( B \cap V = \emptyset \). The following are equivalent:

1. there exists a test sequence \((\phi_n)\) such that for all \( P \in B \) and \( Q \in V \),
   \[ P^n \phi_n \rightarrow 0, \quad Q^n (1 - \phi_n) \rightarrow 0, \]
2. there exists a test sequence \((\phi_n)\) such that for all \( P \in B \) and \( Q \in V \),
   \[ \phi_n(X^n) \xrightarrow{P} 0, \quad (1 - \phi_n(X^n)) \xrightarrow{Q} 0, \]
3. there exists a test sequence \((\phi_n)\) such that for all \( P \in B \) and \( Q \in V \),
   \[ \phi_n(X^n) \xrightarrow{P-a.s.} 0, \quad (1 - \phi_n(X^n)) \xrightarrow{Q-a.s.} 0. \]

This equivalence has a very useful and immediate implication: one is often interested in testing procedures that have property (D), which is a way to formulate the almost-sure version of testing in proposition 8.4. The construction of almost-sure test sequences is often difficult (see, however, [65]), but their existence can be inferred from the much-easier-to-prove in-probability pointwise testability of \( B \) versus \( V \). This fact can be exploited, for example, in Bayesian model selection, and consequently [146], in frequentist model selection with posteriors, if a suitable prior is used.

Some testing problems do not require analysis at the level of the Le Cam-Schwartz theorem because a test sequence can readily be constructed.

Example 8.2. For fixed, measurable \( D \subset \mathcal{X} \), can we test whether \( \text{supp}(P) \subset D \)? (I)

For a measurable \( D \subset \mathcal{X} \), it is possible to test,

\[ H_0 : P(D) = 1, \quad H_1 : P(D) < 1. \]

Namely, take test functions \( \phi_n : \mathcal{X}^n \rightarrow [0, 1] \), defined by,

\[ \phi_n(X_1, \ldots, X_n) = 1 - \prod_{i=1}^n 1\{X_i \in D\}, \]

Then, for all \( P \) satisfying \( H_0 \), \( \phi_n(X^n) \xrightarrow{P-a.s.} 0 \) and for all \( Q \) satisfying \( H_1 \), \( \phi_n(X^n) \xrightarrow{Q-a.s.} 1 \), which implies testability according to definition 8.2.

The above may seem trivial but it has many applications, for example the following.
Example 8.3. For any random $X \in [-\infty, \infty]$, is tightness of $X$ testable?
Suppose we have a measurable map $X : \mathcal{X} \to [-\infty, \infty]$ and hypotheses,

$$H_0 : X \text{ is tight, } H_1 : X \text{ is not tight.}$$

Define $C = \{ \omega \in \mathcal{X} : |X(\omega)| = \infty \}$ and $D = \mathcal{X} \setminus C$ and apply example 8.2 to conclude that tightness is testable.

Another example concerns the presence of point-masses in the data-distribution.

Example 8.4. Can we test whether a distribution contains any point-masses?
In other words, we require a test for the hypotheses,

$$H_0 : \forall x \in \mathcal{X}, P(\{x\}) = 0, \quad H_1 : \exists x \in \mathcal{X}, P(\{x\}) > 0.$$ 

A suitable test sequence is constructed from ties in the sample,

$$\phi_n(X_1, \ldots, X_n) = 1 - \prod_{i,j=1, i \neq j}^n 1\{X_i \neq X_j\}.$$ 

If there exists an $x \in \mathcal{X}$ such that $P(\{x\}) = p > 0$, then the probabilities of seeing no ties decrease like $(1 - p)^n$, so $P^n \phi_n \to 1$; while if $H_0$ holds, probabilities for ties are zero, $P(X_i = X_j) = 0$ if $i \neq j$, so $P^n \phi_n = 0$.

But in more complicated cases one needs the Le Cam-Schwartz theorem. Without formulating requirements on the model, we focus solely on the testability question itself: a pointwise test sequence $(\phi_n)$ for $B$ versus $V$ exists, if and only if, there exists a sequence of $\mathcal{U}_\infty$-uniformly continuous $\psi_n : B \cup V \to [0, 1]$ such that $\psi_n(P) \to 1_{V}(P)$ for all $P \in B \cup V$. Let us first look at example 8.2 through the Le Cam-Schwartz equivalence.

Example 8.5. For fixed, measurable $D \subset \mathcal{X}$, can we test whether $\text{supp}(P) \subset D$? (II)
Take the hypotheses of example 8.2. Let $B = \{ P \in \mathcal{P} : P(D) = 1 \}$ and $V = \{ P \in \mathcal{P} : P(D) < 1 \}$. Define the function $f : \mathcal{P} \to [0, 1]$ by $f(P) = P(D)$ and the sequence $\psi_n = 1 - f^n$. Then the $\psi_n$ are $\mathcal{U}_\infty$-uniformly continuous and $\psi_n(P) \to 0$ for all $P \in B$, while $\psi_n(P) \to 1$ for all $P \in V$.

Example 8.6. Can we test independence of two events $A$ and $B$?
Let $A$ and $B$ be two measurable subsets of the sample space $\mathcal{X}$ for a single observation. We test the hypotheses,

$$H_0 : P(A \cap B) = P(A)P(B), \quad H_1 : P(A \cap B) \neq P(A)P(B).$$

Consider the three $\mathcal{U}_1$-uniformly continuous functions $f_i : \mathcal{P} \to [0, 1], (i = 1, 2, 3),$

$$f_1(P) = P(A \cap B), \quad f_2(P) = P(A), \quad f_3(P) = P(B),$$

and the uniformly continuous function $g : [0, 1]^3 \to [1, -1], g(x_1, x_2, x_3) = x_1 - x_2x_3$. The composition $h : \mathcal{P} \to [-1, 1], h = g \circ (f_1, f_2, f_3)$ and $|h|$ are $\mathcal{U}_1$-uniformly continuous, and so are the functions $\psi_n = |h|^{1/n}$. Note that $\psi_n(P) = 0$ for all $n \geq 1$ if
$h(P) = 0$, and $\psi_n(P) \to 1$ in all other cases. So independence of events $A$ and $B$ is asymptotically testable.

### 8.4.2 Pointwise non-testability

As is pointed out in [86], conditions and examples for non-testability of hypotheses have been largely lacking for a long time (but see [65] for a notable exception). To better understand potential problems obstructing testability, we focus on necessary conditions for testability and hypotheses that are impossible to test.

Suppose that there exists a pointwise consistent test sequence $(\phi_n)$ for $B$ versus $V$. Defining $\psi_n : \mathcal{P} \to [0, 1]$, 

$$\psi_n(P) = P^\phi_n,$$

it is immediate that the $\psi_n$ are all $\mathcal{U}_\infty$-uniformly continuous and that $\psi_n(P) \to 1$ for every $P$. This implies that the nature of pointwise testable pairs of hypotheses $B, V$ can be described quite precisely.

**Proposition 8.5.** Suppose that there exists a pointwise consistent test sequence $(\phi_n)$ for $B$ versus $V$. Then both $B$ and $V$ are both $G_\delta$- and $F_\sigma$-sets with respect to $T_\infty$ in the subspace $B \cup V$.

**Proof.** Let $\varepsilon < 1/2$ be given, consider the closed sets $B_n = \{ P \in B \cup V : \psi_n(P) \leq \varepsilon \}$ and $V_n = \{ P \in B \cup V : \psi_n(P) \geq 1 - \varepsilon \}$. For every $P \in B$ there exists an $N \geq 1$ such that $P$ lies in the closed set $\cap_{n \geq N} B_n$. So $P$ lies in the $F_\sigma$-set $\cup_{N \geq 1} \cap_{n \geq N} B_n$. Conversely, if $P$ lies in $\cup_{N \geq 1} \cap_{n \geq N} B_n$, then $\psi_n(P) \leq \varepsilon$ for large enough $n$, which implies that $\psi_n(P) \to 0$ because we assume testability, so $P \in B$. Conclude that $B$ is an $F_\sigma$-set. Since the same holds for $V$ by symmetry, the complement of $B$ in $B \cup V$ is also $F_\sigma$, that is, $B$ is also $G_\delta$.

In the language of descriptive set theory, hypotheses that are testable versus their complements in the model belong to the class of ambiguous sets $\Delta^0_2(\mathcal{P})$. Because $\mathcal{P}$ with the $\mathcal{I}_\infty$-topology is not necessarily metrizable, there is no guarantee that $\mathcal{I}_\infty$-open (or -closed) subsets are $F_\sigma$ (or $G_\delta$) in general. However, in Polish models (for examples, see remark 8.3 and corollary 8.5) testability implies completeness.

**Corollary 8.1.** Suppose that $\mathcal{P} = B \cup V$ is Polish in the $\mathcal{I}_\infty$-topology and that $B$ is pointwise testable versus $V$. Then the hypotheses $B$ and $V$ are complete (and hence Polish) subspaces of $\mathcal{P}$.

**Proof.** Hypotheses $B$ and $V$ are both $F_\sigma$ in $B \cup V$, if and only if, they are both $G_\delta$ in $B \cup V$, if and only if, they are both Polish in $B \cup V$. Metrizability and separability (in metrizable spaces) are subspace properties but completeness is not.

Testability of hypotheses in Polish spaces implies that the hypothesized sets $B$ and $V$ themselves are complete. That suggests that pointwise non-testability of a hypothesis can be shown through the Baire category theorem.
Corollary 8.2. Suppose $\mathcal{P}$ is a Baire space in a topology that is $T_\infty$ or finer. If $B$ and $\mathcal{P} \setminus B$ are dense in $\mathcal{P}$, then $B$ is not pointwise testable versus $\mathcal{P} \setminus B$.

Proof. We prove this by contradiction: assume that $B$ is testable versus its complement in $\mathcal{P}$. Then $B$ and its complement $C := \mathcal{P} \setminus B$ are both $G_\delta$-sets, so there exist sequences of open sets $(B_n)$ and $(C_n)$ such that $B = \cap_{n=1}^\infty B_n$ and $C = \cap_{n=1}^\infty C_n$. Because both $B$ and $C$ are dense in the Baire space $\mathcal{P}$, the intersection $D = \cap\{B_n \cap C_n : n \geq 1\}$ is a countable intersection of dense open subsets, so $D$ is dense. However, $B$ and $C$ are disjoint, so

$$\left( \bigcap_{n=1}^\infty B_n \right) \cap \left( \bigcap_{n=1}^\infty C_n \right) = B \cap C = \emptyset,$$

so the intersection $D$ cannot be dense.

Remark 8.1. The condition that $\mathcal{P}$ be a Baire space is not as stringent as it looks: if $C \subset B$ and $W \subset V$, then non-testability of $C$ versus $W$ implies non-testability for $B$ versus $V$. So the above corollary could have been formulated slightly more generally as follows: if $B, V \subset \mathcal{P}$ and $B \cap V = \emptyset$ are given and there exists a Baire subspace $D$ of $\mathcal{P}$ in which both $D \cap B$ and $D \cap V$ are dense, then $B$ is not testable versus $V$. Aside from the remark that the Polish spaces we have discussed are Baire spaces, the Baire property is often applicable in dominated (sub-)models, under the condition of uniform integrability. Namely, because (locally) compact Hausdorff spaces are Baire spaces, and (relative) $T_\infty$-compactness is often an easily accessible property (see the argument leading up to corollary 8.5), finding a Baire sub-problem $D$ in examples is perhaps less demanding than it appears.

Example 8.7. Is Cover’s rational means problem testable?

The above proves that Cover’s rational mean problem has a negative answer. To prove this, first note that $\mathcal{P}$ is dominated and the Dunford-Pettis theorem shows that $\mathcal{P}$ is $T_\infty$-compact (so that $T_\infty = T_1$). There is an injective parametrization $P : [0, 1] \to \mathcal{P}$, with $P_p(\{1\}) = 1 - P_p(\{0\}) = p$: any $p, q \in [0, 1]$, $P_p \neq P_q$ and, given $f : [0, 1] \to [0, 1]$,

$$| (P_p - P_q) f | = | (p - q) (f(1) - f(0)) | \leq | p - q |,$$

so that $P$ is $T_\infty$-continuous and therefore a homeomorphism. Since $[0, 1]$ is a complete metric space, $\mathcal{P}$ is a Baire space for the $T_\infty$-topology. Because both $[0, 1] \cap \mathbb{Q}$ and $[0, 1] \setminus \mathbb{Q}$ are dense in $[0, 1]$, the images $\mathcal{P}_0 := \{ P_p : p \in [0, 1] \cap \mathbb{Q} \}$ and $\mathcal{P}_1 := \{ P_p : p \in [0, 1] \setminus \mathbb{Q} \}$ are $T_\infty$-dense in $\mathcal{P}$. So there does not exist a pointwise test for $p \in [0, 1] \cap \mathbb{Q}$ versus $p \in [0, 1] \setminus \mathbb{Q}$.

Unfortunately, many common statistical assumptions are of this type.

Example 8.8. Is integrability of a random variable $X$, $P[X] < \infty$, testable?

Let $X$ be a real-valued random variable with some distribution in the space $\mathcal{P}$ of all probability distributions on $\mathbb{R}$. When equipped with the total-variation norm or the
Hellinger metric, \((\mathcal{P}, \mathcal{I}_d)\) refines \(\mathcal{I}_w\), is a Polish space and therefore has the Baire property. Define the dichotomy \(\mathcal{P}_0 = \{P \in \mathcal{P} : |P| < \infty\}\), \(\mathcal{P}_1 = \{P \in \mathcal{P} : |P| = \infty\}\). The sets \(\mathcal{P}_0\) and \(\mathcal{P}_1\) are non-empty, so let \(P \in \mathcal{P}_0\) and \(Q \in \mathcal{P}_1\) be given. For any \(0 < \varepsilon < 1\), \(P' = (1-\varepsilon)P + \varepsilon Q\) satisfies \(\|P' - P\| = \varepsilon \|P + Q\| \leq 2\varepsilon\), but \(P' \in \mathcal{P}_1\). Conclude that \(\mathcal{P}_1\) lies \(\mathcal{I}_d\)-dense in \(\mathcal{P}\). Conversely, tightness of \(Q\) implies that for every \(\varepsilon > 0\), there exists a constant \(M > 0\) such that \(|Q(A) - Q(A||X| \leq M)| < \varepsilon\) for all measurable \(A \subset \mathbb{R}\). Since \(Q(|X| \leq M) \in \mathcal{P}_0\), we also see that \(\mathcal{P}_0\) lies \(\mathcal{I}_d\)-dense in \(\mathcal{P}\). So \(\mathcal{P}_0\) cannot be tested versus \(\mathcal{P}_1\).

Since we cannot test for integrability of \(X\), there is no asymptotic, statistical way of finding out whether use of the Law of Large Numbers is justified. In fact, integrability with regard to any unbounded random variable on \(\mathbb{R}\) (e.g. \(P|f(X)| < \infty\) for some \(f : \mathbb{R} \to \mathbb{R}\)) cannot be tested: in particular, square-integrability of \(X\) cannot be tested, so use of the Central Limit Theorem cannot be justified with tail-probability one either, based on an i.i.d. sample.

**Example 8.9.** Can we test whether a random variable \(X\) is compactly supported?

\[
H_0 : \exists K : P(X \in K) = 1, \quad H_1 : \forall K : P(X \in K) < 1
\]

Let all \(P \in \mathcal{B}\) be such that \(P(X \in K) = 1\) for some compact \(K\) and all \(Q \in \mathcal{V}\) such that there is no such \(K\). Then for all \(0 < \varepsilon < 1\), \(P' = (1-\varepsilon)P + \varepsilon Q \in \mathcal{V}\) while \(\|P - P'\| \leq 2\varepsilon\), so \(V\) lies \(\mathcal{I}_d\)-dense in \(\mathcal{B} \cup \mathcal{V}\). Since \(\mathbb{R}\) is a Radon space, for \(\varepsilon > 0\) and any \(Q \in \mathcal{V}\) there exists a compact \(K\) such that \(|Q(A) - Q(A||X| \leq M)| < \varepsilon\) for all \(A\). Therefore, also \(B\) lies \(\mathcal{I}_d\)-dense in \(\mathcal{B} \cup \mathcal{V}\). Since the collection of all probability measures on \(\mathbb{R}\) is completely metrizable in the \(\mathcal{I}_d\) topology, \(\mathcal{B} \cup \mathcal{V}\) is a Baire space and we conclude that there does not exist a pointwise test sequence for \(H_0\) versus \(H_1\).

Cover’s rational mean problem can be called prototypical for non-testability of hypotheses, at least, if we are willing to restrict the issue to models that are Polish for \(\mathcal{I}_w\) (for examples, see remark 8.3 and corollary 8.5). In Polish models we consider the potential testability of hypotheses that correspond to analytic subsets. (A subset \(A\) is analytic if it is the continuous image of a Polish space ([134], sections 7.F, 21.F). The class of all analytic sets is very large; it contains all Borel subsets of \(\mathcal{P}\).) To demonstrate how Cover’s problem makes an appearance when non-testability is in play, we consider analytic subsets \(B\) and \(V\) of a Polish \(\mathcal{P} = \mathcal{B} \cup \mathcal{V}\) that are not both \(F_\sigma\)-sets. Clearly \(B\) is not asymptotically pointwise testable versus \(V\). Hurewicz’s theorem [134] provides crucial insight.

**Theorem 8.5.** (Hurewicz) Let \(\mathcal{P}\) be a Polish space and let \(A\) be analytic in \(\mathcal{P}\). If \(A\) is not \(F_\sigma\), then there exists a Cantor set \(C\) such that \(C \setminus A\) is countably dense in \(C\), homeomorphic to \([0, 1] \cap \mathbb{Q}\), while \(C \cap A\) is closed in \(A\), and homeomorphic to \([0, 1] \setminus \mathbb{Q}\).

On the basis of the (more general) Kechris-Louveau-Woodin theorem [133, 134] we may say the following: suppose that \(\mathcal{P}\) is Polish in the \(\mathcal{I}_w\)-topology. If we have two disjoint analytic model subsets \(B\) and \(V\) that are not both \(F_\sigma\) in \(\mathcal{I}_w\) (which is
necessary for testability), then there exists a sub-testing-problem, in the form of a (Cantor) subset \( C \subset B \cup V \) with a representation on \([0,1]\) in which the testing of \( B \cap C \) versus \( V \cap C \) is represented on \([0,1]\) as Cover’s rational mean problem. So if, for example, we are in a model that has finite total-variational entropy numbers (so that on the separable completion, total-variational and \( T_\infty \) topologies coincide), non-\( F_\sigma \)-ness of any analytic hypotheses can always be reduced to a non-testable Cover sub-problem.

### 8.4.3 Pointwise testability in dominated models

For the following theorem, recall that the testing problem has a (uniform) representation on \( X \), if there exists a \( T_\infty \)-(uniformly-)continuous surjective map \( f : B \cup V \to X \) such that \( f(B) \cap f(V) = \emptyset \).

**Theorem 8.6.** Let a dominated model \( \mathcal{P} \) with hypotheses \( B, V \) \( B \cap V = \emptyset \) be given. The following are equivalent,

1. there exists a pointwise test sequence for \( B \) versus \( V \);
2. the testing problem has a representation \( f : B \cup V \to X \) on a normal space \( X \) and there exist disjoint \( F_\sigma \)-sets \( B', V' \subset X \) such that \( f(B) \subset B', f(V) \subset V' \);
3. the testing problem has a uniform representation \( \psi : B \cup V \to X \) on a separable, metrizable space \( X \) with \( \psi(B), \psi(V) \in \Delta_0^2(X) \).

The proof of this theorem requires some vector-space reasoning. Given the model \( \mathcal{P} \) in the \( T_\infty \)-topology, we define the linear space \( E \) of all bounded, continuous \( f : \mathcal{P} \to \mathbb{R} \) and the linear space \( F \) that is the linear span of the collection of all degenerate (Borel) measures \( \delta_P \) on \( \mathcal{P} \): for any \( \lambda \in F \), there exist \( m \geq 1, \lambda_1, \ldots, \lambda_m \in \mathbb{R} \setminus \{0\} \) and distinct \( P_1, \ldots, P_m \in \mathcal{P} \) such that \( \lambda \) can be written (uniquely) as:

\[
\lambda = \sum_{i=1}^{m} \lambda_i \delta_{P_i}.
\]

(8.10)

Definition of the bi-linear form \( \langle \cdot, \cdot \rangle : E \times F \to \mathbb{R} \),

\[
\langle f, \lambda \rangle = \int_{\mathcal{P}} f(P) d\lambda(P) = \sum_{i=1}^{m} \lambda_i f(P_i),
\]

puts \( E \) and \( F \) in duality. This duality is separating for both \( E \) and \( F \). (Because \( \mathcal{P} \) is a uniform space (hence regular), point-sets are separated by continuous functions.) With the corresponding weak topologies \( \sigma(E, F) \) and \( \sigma(F, E) \), the spaces \( E \) and \( F \) form a dual pair of Hausdorff locally convex spaces (see, [46], Ch. 4). Note that the topology of pointwise convergence for bounded, continuous functions on \( \mathcal{P} \) coincides with \( \sigma(E, F) \). Within \( E \), define,

\[
H = \{ f \in E : 0 \leq f \leq 1, f \mathcal{P}_\infty \text{-unif. cont.} \}.
\]
The bi-polar theorem guarantees that for $H$, the closure $\overline{H}$ equals the bi-polar $H^{\circ\circ}$ which enables the following result.

**Lemma 8.1.** Every $\mathcal{F}_\infty$-continuous $f : \mathcal{P} \to [0, 1]$ lies in the $\sigma(E,F)$-closure $\overline{H}$ of $H$.

**Proof.** According to the bi-polar theorem (see theorem 1 of [46], Ch. II, § 6, No. 3), the polar $H^{\circ\circ} \subset E$ of the polar $H^\circ \subset F$ is equal to the closed convex envelope of $H \cup \{0\}$. Since $H$ is convex and contains $0 \in E$, $H^{\circ\circ} = \overline{H}$. For given $\lambda \in F$ there exists an $m \geq 1$, $\lambda_1, \ldots, \lambda_m \in \mathbb{R} \setminus \{0\}$ and distinct $P_1, \ldots, P_m \in \mathcal{P}$ such that $\lambda$ is written uniquely as $\lambda = \sum_i \lambda_i \delta_{P_i}$. Fix some $m \geq 1$ and distinct $P_1, \ldots, P_m \in \mathcal{P}$ and consider the finite-dimensional subspace of $F$ we obtain when we vary $w = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$. Any $f \in H$ is represented on this subspace only through the values $v = (f(P_1), \ldots, f(P_m)) \in [0,1]^m$ and any $\lambda$ supported on $\{P_1, \ldots, P_m\}$ lies in $H^\circ$ whenever the inner product $\langle v, w \rangle$ in $\mathbb{R}^m$ is greater than or equal to $-1$. Because the cube $[0,1]^m$ is the convex hull of its corner points, we see that if the coefficients $\lambda_1, \ldots, \lambda_m$ are such that

$$\sum_{i \in M} \lambda_i \geq -1,$$ (8.11)

for any finite subset $M$ of $\{1, \ldots, m\}$, then $\lambda \in H^\circ$. Conclude that if we define $L$ to be the subset of all $\lambda \in F$ that satisfy (8.11) when decomposed according to (8.10), then $L \subset H^\circ$. Conversely, let $\lambda \in H^\circ$ be given (again represented in the form $\lambda = \sum_i \lambda_i \delta_{P_i}$). For every $1 \leq i < j \leq m$, define the $\mathcal{B}_\infty$-measurable maps $x^n \mapsto \phi_{ij,n}(x^n)$ to be likelihood ratio tests (with $\mu = P_i + P_j$ and $p_i = dP_i/d\mu$, $p_j = dP_j/d\mu$):

$$\phi_{ij,n}(X^n) = 1\{p^n_i(X^n) < p^n_j(X^n)\}.
$$

Then, because the Hellinger distance $H(P_i, P_j)$ between $P_i$ and $P_j$ is strictly positive,

$$P^n_i \phi_{ij,n} + P^n_j (1 - \phi_{ij,n})$$

$$= \int \left( p^n_i(x^n) 1\{p^n_i(x^n) < p^n_j(x^n)\} + p_j(x^n) 1\{p^n_i(x^n) \geq p^n_j(x^n)\} \right) d\mu^n(x^n)$$

$$\leq \int \sqrt{p^n_i(x^n) p^n_j(x^n)} d\mu^n(x^n) = 1 - \frac{1}{2} \int \left( \sqrt{p^n_i(x^n)} - \sqrt{p^n_j(x^n)} \right)^2 d\mu^n(x^n)$$

$$= 1 - H^2(P^n_i, P^n_j) \leq e^{-nH^2(P_i, P_j)} \to 0.$$ (For the last inequality, see, for example, lemma 2.17 in Strasser [208].) Choose some $0 < \varepsilon < 1/2$ and $N$ large enough such that $P^n_i \phi_{ij,N} < \varepsilon$ and $P^n_j \phi_{ij,N} > 1 - \varepsilon$. Then the function $f_{ij} : \mathcal{P} \to [0, 1]$,

$$f_{ij}(P) = \left( P^n_i \phi_{ij,N} - P^n_j \phi_{ij,N} \right) / \left( P^n_i \phi_{ij,N} - P^n_j \phi_{ij,N} \right) \vee 1,$$

is uniformly continuous with respect to $\mathcal{U}_N$ (which coarsens $\mathcal{U}_\infty$) and satisfies $f_{ij}(P_i) = 0$, $f_{ij}(P_j) = 1$, so $f_{ij}$ lies in $H$ and separates $P_i$ and $P_j$. For every $M \subset \{1, \ldots, m\}$, we can construct an $f \in H$ from the collection $\{f_{ij} : 1 \leq i < j \leq m\}$,
such that,

$$\langle f, \lambda \rangle = \sum_{i \in M} \lambda_i,$$

so that \( H^o \subset L \) as well. Conclude that \( H^o = L \). Next, consider the polar \( H^{oo} \subset E \) of \( H^o \); let \( \lambda \in H^o \) and \( f : \mathcal{P} \to [0,1] \) in \( E \) be given. Reasoning like before we see that \( \langle f, \lambda \rangle \) can be replaced by an inner product \( \langle v, w \rangle \) in \( \mathbb{R}^m \) and that \( \langle f, \lambda \rangle \geq -1 \) because the coefficients \( \lambda_1, \ldots, \lambda_m \) satisfy (8.11). Conclude that \( f \in H^{oo} \).

So any \( \mathcal{T}_w \)-continuous \( f : \mathcal{P} \to [0,1] \) is approximated arbitrarily closely by \( \mathcal{U}_w \)-uniformly continuous functions, with respect to any of the semi-norms that define \( \sigma(E,F) \). Although, for every continuous \( f \), this implies the existence of nets of uniform functions that converge to \( f \), nothing is implied regarding the existence of a convergent sequence of uniform functions. For that step, the conclusion of the next lemma is sufficient, however.

**Lemma 8.2.** If \( \mathcal{P} \) is dominated, \( \overline{\mathcal{P}} \) is separable and metrizable with respect to \( \sigma(E,F) \).

**Proof.** Because we assume that the \( \sigma \)-algebra \( \mathcal{B} \) is countably generated, Strasser’s lemma 4.1 [208] says that \( \mathcal{P} \) is separable with respect to the total-variational topology. This implies that \( \mathcal{P} \) is also separable in the \( \mathcal{T}_w \)-topology (because \( \mathcal{T}_d \) refines \( \mathcal{T}_w \), but see also theorems 4.4 and 21.3 in [208]). As the linear span of a set with countable dense subset, \( F \) (has a total set and) is separable with respect to \( \sigma(E,F) \).

And as a consequence of that, \( E \) is first-countable at zero with respect to \( \sigma(E,F) \). The total-variational norm \( \| \cdot \|_{TV} \) makes \( F \) a normed space, with continuous dual \( F' \), and \( F' \) can be equipped with the (weak-star) topology \( \sigma(F',F) \). If we define, for every bounded \( \mathcal{T}_w \)-continuous \( f : \mathcal{P} \to \mathbb{R} \), the linear map \( g_f : F \to \mathbb{R} \),

$$g_f(\lambda) = \int_{\mathcal{P}} f(P) d\lambda(P) = \langle f, \lambda \rangle,$$  

(8.12)

then, with \( 0 \leq |f| \leq \|f\| = \sup_{P \in \mathcal{P}} |f(P)| \),

$$|g_f(\lambda)| \leq \|f\| \|\lambda\|_{TV},$$

so \( g_f \) lies in \( F' \), for every \( f \in E \). This map is one-to-one and a \( \sigma(E,F) \)-to-\( \sigma(F',F) \) homeomorphism between \( E \) and \( \mathcal{E} = \{ g_f : f \in E \} \subset F' \), and we conclude that \( \mathcal{E} \) is first-countable at zero. Also, every norm-bounded set in \( E \) (and in particular the set \( \overline{\mathcal{E}} \)) is mapped to a norm-bounded subset of \( \mathcal{E} \) (denoted \( G \) in the case of \( \mathcal{E} \)) by (8.12), with norm-bounded closure \( \overline{\mathcal{G}} \) with respect to the norm-topology on \( F' \). Then, according to cor. 2 of [46], Ch. III, §3, No. 4, \( \overline{\mathcal{G}} \) is a compact, metrizable space for \( \sigma(F',F) \), which implies that \( G \) is separable and metrizable with respect to \( \sigma(F',F) \), which is equivalent to separability and metrizability of \( \overline{\mathcal{E}} \) for \( \sigma(E,F) \).

Now a diagonalization argument suffices to draw the conclusion that if \( B \) and \( V \) are separated by a sequence of continuous functions (i.e. there exist continuous \( \psi_n : B \cup V \to [0,1] \) such that \( \psi_n \to 1_V \)), then \( B \) is pointwise testable versus \( V \). The
representation in terms of $F_{\sigma}$-sets on a normal space guarantees the existence of the $\psi_n$ through Urysohn’s lemma.

**Proof.** (of theorem 8.6)
Assume condition (ii). The disjoint sets $B'$ and $V'$ can be written as countable unions of closed sets and because $X$ is a normal space there exists a sequence of continuous $g_n : X \to [0, 1], (n \geq 1)$ such that for each $x \in B'$ (resp. $y \in V'$), there is an $N \geq 1$ such that $g_n(x) = 0$ (resp. $g_n(y) = 1$) for all $n \geq N$. Composition with $f : B \cup V \to X$ gives rise to a sequence of $\mathcal{F}_\sigma$-continuous $\psi_n = g_n \circ f : B \cup V \to [0, 1]$ such that $\psi_n(P) \to 1_V(P)$ for all $P \in B \cup V$. For each $n \geq 1$, lemma 8.1 asserts that $\psi_n$ lies in $\overline{H}$ and, according to lemma 8.2, $\overline{H}$ is metrizable for $\sigma(E, F)$, which implies the existence of a consistent pointwise test $P$ versus $V$, i.e. condition (i) follows from condition (ii).

Next, assume condition (i), that $(\phi_n)$ is a pointwise test sequence for $B$ versus $V$. Define the $\mathcal{U}_\sigma$-uniformly continuous maps $\psi_n : B \cup V \to [0, 1], \psi_n(P) = P^\epsilon \phi_n$, and the mapping $\psi : B \cup V \to \prod_{n \geq 1} [0, 1]. \psi(P) = (\psi_n(P) : n \geq 1)$. The map $\psi$ is $\mathcal{U}_\sigma$-uniformly continuous and the image $X = \psi(B \cup V)$ in the (separable, metrizable) product space $\prod_{n \geq 1} [0, 1]$ is separable and metrizable. Next, we reason similar to proposition 8.5: let $0 < \epsilon < 1/2$ be given and consider the closed product sets $c_n, w_n \subset \prod_{n \geq 1} [0, 1],$

$$c_n = [0, 1] \times \ldots \times [0, 1] \times [0, \epsilon] \times [0, 1] \times \ldots,$$

$$w_n = [0, 1] \times \ldots \times [0, 1] \times [\epsilon, 1] \times [0, 1] \times \ldots,$$

(with the $\epsilon$-dependent intervals as the $n$-th factors) and the sets $b_N = \cap_{n \geq N} (c_n \cap X), v_N = \cap_{n \geq N} (w_n \cap X)$ for all $N \geq 1$ which are closed in the subspace $X$. Note that for any $P \in B$ (resp. any $Q \in V$), there exists an $N \geq 1$ such that $\psi(P) \in b_N$ (resp. $\psi(Q) \in v_N$), so $\psi(B)$ is a subset of the $F_\sigma$-set $\cup_{N \geq N} b_N$ in $X$ (resp. $\psi(V)$ is a subset of the $F_\sigma$-set $\cup_{N \geq N} v_N$ in $X$). Conversely, if $x \in \cup_{N \geq N} b_N$ (resp. $y \in \cup_{N \geq N} v_N$), there exists a $P \in B \cup V$ such that $x = \psi(P)$ (resp. $y = \psi(P)$) and $\lim_n \psi_n(P) < 1/2$ (resp. $\lim_n \psi_n(P) > 1/2$), which means that $P \in B$ (resp. $P \in V$), i.e. $\cup_{N \geq N} b_N \subset \psi(B)$ (resp. $\cup_{N \geq N} v_N \subset \psi(V)$). So $\psi(B) = X \setminus \psi(V) \in \Delta_2^0(X).$ Condition (iii) follows from condition (i). Condition (ii) follows from condition (iii) because metrizable spaces are normal spaces.

**Corollary 8.3.** Suppose that $\mathcal{P}$ is dominated and there exist disjoint $F_\sigma$-sets $B', V'$ in the completion $\hat{\mathcal{P}}$ (for $\mathcal{U}_\sigma$) such that $B \subset B', V \subset V'$. Then $B$ is pointwise testable versus $V$.

**Proof.** Since $\mathcal{P}$ is pre-compact for $\mathcal{U}_\sigma$, the completion $\hat{\mathcal{P}}$ is compact (and hence normal) and the canonical embedding $\mathcal{P} \to \hat{\mathcal{P}}$ is continuous. Formulation (ii) of theorem 8.6 is then satisfied, and we conclude formulation (i).

**Remark 8.2.** Based on corollary 8.3, it is tempting to conclude that testability must be equivalent to the existence of disjoint $F_\sigma$-sets $B'', V''$ such that $B \subset B''$ and $V \subset V''.

232 8 Consistent tests and model selection
in the original model $\mathcal{P}$. But corollary 8.3 requires more: the existence of disjoint $F_\sigma$-sets $B, V$ in $\mathcal{P}$ cannot be guaranteed from the existence of disjoint $B', V'$ that are $F_\sigma$ in $\mathcal{P}$. For the same reason, the converse of corollary 8.3 does not follow from corollary 8.5. This observation does allow for the following re-formulation, however: suppose that $\mathcal{P}$ is dominated and complete for $\mathcal{U}_\infty$ with disjoint subsets $B, V$. Then $B$ is pointwise testable versus $V$, if and only if, there exist disjoint $F_\sigma$-sets $B', V' \subset \mathcal{P}$ such that $B \subset B', V \subset V'$.

Although perhaps pleasantly succinct from a mathematical perspective, corollary 8.3 is not practical unless the model can easily be shown to be complete for $\mathcal{U}_\infty$. More common, for example, are models that describe a (possibly non-parametric) family of Lebesgue densities as a metric space, where the metric is related in some way (e.g. through inequalities) to the Hellinger or total-variational metrics. For the following corollary, we think of the model $\mathcal{P}$ as a metric space with a metric $d$ that (is equal to or) refines the total-variational metric (e.g. for all $P, Q \in \mathcal{P}$, $\|P - Q\| \leq f(d(P, Q))$ for some strictly increasing $f : [0, \infty) \to [0, \infty)$). The argument below gives an explanation for the ubiquity in the mathematical statistics literature of the assumption that the model has finite metric entropy numbers (e.g. for all $\epsilon > 0$, the covering number $N(\epsilon, \mathcal{P}, d) < \infty$).

**Corollary 8.4.** Suppose that $\mathcal{P}$ is dominated and totally bounded with respect to the total-variational metric. Then $B, V \subset \mathcal{P}$, $B \cap V = \emptyset$ are pointwise testable, if and only if, $B, V$ are $F_\sigma$-sets for the total-variational topology in $B \cup V$.

**Proof.** The closure $\overline{\mathcal{P}}$ of $\mathcal{P}$ in the vector space of all finite, signed measures on $(\mathcal{X}, \mathcal{B})$ with respect to $\mathcal{T}_V$ is compact. Because $\mathcal{T}_V$ refines $\mathcal{T}_\infty$, the identity $i : \overline{\mathcal{P}} \to \mathcal{P}$ is a $\mathcal{T}_V$-to-$\mathcal{T}_\infty$ homeomorphism. The inverse $i^{-1}$ is $\mathcal{T}_\infty$-to-$\mathcal{T}_V$ continuous (and so is its restriction to $B \cup V$). Since the subspace $B \cup V$ remains metrizable for $\mathcal{T}_V$, $i^{-1} : B \cup V \to B \cup V$ is a representation of the testing problem on a normal space.

**Remark 8.3.** To appreciate the role that metric entropy numbers play here, consider a model $\mathcal{P}$ of Lebesgue densities $p : [0, 1] \to \mathbb{R}$ and equip it with the $L^\infty$-norm, $d(P, Q) = \|P - Q\|_\infty$, then certainly, $\|P - Q\|_{TV} \leq d(P, Q)$. Hypotheses that involve $d$, like testing for $\|\cdot\|_\infty$-neighbourhoods of a fixed $Q \in \mathcal{P}$,

$$H_0 : \|p - q\|_\infty < \delta, \quad H_1 : \|p - q\|_\infty \geq \delta,$$

are, in principle, not testable because these hypotheses are $F_\sigma$-sets in a topology that is strictly stronger than $\mathcal{T}_\infty$. But if the model is totally bounded with respect to $d$, the proof of corollary 8.4 extends because $d$ refines total variation. That renders the hypotheses of (8.13) testable, because now they correspond to $F_\sigma$ sets in the total-variational and $\mathcal{T}_\infty$ topologies.

Because $\mathcal{T}_\infty$ refines Prohorov’s weak topology $\mathcal{T}_C$, we may also weaken the model topology to $\mathcal{T}_C$ by imposing $\mathcal{T}_\infty$-compactness. (Relative $\mathcal{T}_\infty$-compactness is a weaker requirement than relative compactness in total-variation.) The results of Er- makov (2017) [86] are formulated under this assumption and the main result of Dembo and Peres (1994) [64] relates to ours through the same construction.
Corollary 8.5. Suppose that $\mathcal{P}$ is dominated by a probability measure, with a uniformly integrable family of densities. Then $B, V \subset \mathcal{P}$, $B \cap V = \emptyset$ are pointwise testable, if and only if, $B, V$ are $F_\sigma$-sets for Prokhorov’s weak topology in $B \cup V$.

Proof. Consider the model $\mathcal{P}$ as a subspace of the vector space of all finite, signed measures on $(\mathcal{X}, \mathcal{B})$ equipped with the $\mathcal{U}_1$-uniformity, and denote by $\mathcal{T}_1$ the $\mathcal{T}_1$-closure of $\mathcal{P}$. Because $\mathcal{P}$ is dominated by a probability measure $Q$, $\mathcal{P}$ is dominated by $Q$ (see, e.g., lemma 4.3 and theorem 4.8 in [208]). Clearly, the resulting family $\mathcal{T}_1(Q)$ of $Q$-densities in $L^1(Q)$ is the weak closure of $\mathcal{P}_Q$. In fact, embedding $L^1(Q)$ in the space of all finite, signed measures on $(\mathcal{X}, \mathcal{B})$ canonically, $\mathcal{T}_1(Q)$ with the weak topology and $\mathcal{T}_1$ are homeomorphic. By assumption, $\mathcal{T}_1(Q)$ is relatively weakly compact, so $\mathcal{T}_1$ is weakly compact and $\mathcal{P}$ is $\mathcal{T}_1$-compact. It is shown in the proof of lemma 3 of section 17.5 of Le Cam (1986) [162] (in the somewhat broader context of theorem 6 of appendix 8 in [162]) that weak convergence of a net $f_\alpha \to f$ in $L^1(Q)$ implies weak convergence of product densities $f_\alpha^n \to f^n$ weakly in $L^1(Q^n)$, as a result of the Dunford-Pettis theorem (see also lemma 3.8 in [208]).

Suppose that $f_\alpha$ is an arbitrary net in $\mathcal{T}_1$, then there exists a convergent subnet $f_\beta \to f$ in $\mathcal{T}_1(Q)$ which implies that $f_\beta^n \to f^n$ weakly in $L^1(Q^n)$. That means that the set $\{f^n \in L^1(Q^n) : f \in \mathcal{T}_1(Q)\}$ is weakly compact and $\mathcal{T}_1$ is compact with respect to $\mathcal{T}_1$, for all $n \geq 1$. Because compact spaces have unique uniformities compatible with their topologies, $\mathcal{U}_n = \mathcal{U}_1$ for all $n \geq 1$ and consequently $\mathcal{U}_\infty = \mathcal{U}_1 = \mathcal{U}_C$ on $\mathcal{T}_1$. Therefore, the identity $i : \mathcal{T}_1 \to \mathcal{T}_1$ is a $\mathcal{U}_n$-to-$\mathcal{U}_C$ homeomorphism of uniform spaces. Since $\mathcal{T}_1$ is metrizable for $\mathcal{T}_C$, so is the subspace $B \cup V$, which means that $i : B \cup V \to B \cup V$ is a (uniform) representation of the testing problem on a normal space. Theorem 8.6 then asserts the existence of a pointwise test of $B$ versus $V$.

Corollary 8.5 is related to theorem 2 in Dembo and Peres (1994) [64], which says that in a dominated model, a test sequence for $B$ versus $V$ exists if $B$ and $V$ are contained in disjoint $F_\sigma$-sets for the $\mathcal{T}_C$-topology, while the converse holds true whenever $\int (dP/dQ)^p dQ < \infty$ for some $p > 1$ and all $P \in \mathcal{P}$ (which implies uniform integrability). Ermakov formulates the following strengthening of the Dembo-Peres result.

Corollary 8.6. (Ermakov (2014), theorem 3.2)

Suppose that $\mathcal{P}$ is dominated by a probability measure, with a uniformly integrable family of densities. Then $B, V \subset \mathcal{P}$, $B \cap V = \emptyset$ are pointwise testable, if and only if, there exist sequences $(B_m), (V_m)$ with $\cup_m B_m = B$ and $\cup_m V_m = V$, and uniform test sequences $(\phi_{m,n} : \mathcal{X} \to [0, 1] : n \geq 1), (m \geq 1)$, such that,

$$\sup_{P \in B_m} P^n \phi_{m,n} + \sup_{Q \in V_m} Q^n (1 - \phi_{m,n}) \to 0,$$

for all $m \geq 1$.

Proof. Assume that $B$ is pointwise testable versus $V$; according to proposition 8.5, there exist sequences $(B_m), (V_m), (m \geq 1)$ of sets that are closed in the subspace $B \cup V$. Since $B \cup V \subset \mathcal{P}$ is relatively $\mathcal{T}_\infty$-compact, $B_m$ and $V_m$ are uniformly separated by
The above proof of Ermakov’s theorem relies in a crucial way on (relative) compactness with respect to $\mathcal{T}_m$, because the all-important existence assertion in the proof follows from proposition 8.2.

Example 8.10. Are Cantor sets of the right topological type to be testable?

Cover’s rational mean problem of hypotheses (8.1), concerning the parameter $p \in [0,1]$ of and i.i.d. sequence $X_1, X_2, \ldots$ of coin-flips, may also be posed with other hypotheses such as those of (8.3): can we test whether $p$ lies in the Cantor set $B = C \subset [0,1]$ or in its complement $V = [0,1] \setminus C$? As we have seen in example 8.7, the map $[0,1] \to \mathcal{P} : p \mapsto P_p$ is a $\mathcal{T}_m$-homeomorphism. In particular, this implies $\mathcal{P}$ is compact and metrizable for $\mathcal{T}_m$. Because $B$ is closed, $B$ is $F_{\sigma}$ in $[0,1]$ and so is its image in $\mathcal{P}$. Because open sets in metrizable spaces are $F_{\sigma}$, $V$ and its image in $\mathcal{P}$ are $F_{\sigma}$. We may now use theorem 8.6 or corollary 8.3 to conclude that $C$ is testable versus its complement. More broadly, any (non-empty) topological space is homeomorphic to $C$, if and only if, it is perfect, compact, totally disconnected and metrizable. So the above concrete example represents a whole class of testing problems, those in which one of the hypotheses satisfies said characteristic topological properties as a subspace of a model $\mathcal{P} = B \cup V$ that is metrizable with respect to $\mathcal{T}_m$.

8.5 Bayesian test sequences

First of all, the existence of a Bayesian test sequence is linked directly to behaviour of the posterior itself. In the following, $\mathcal{P}$ is a model for i.i.d. data $X^n$ taking values in a measurable space $(\mathcal{X}^n, \mathcal{B}^n)$. Assume that $X^n \sim P^n$, for some $P \in \mathcal{P}$ and all $n \geq 1$. Assume also that $\mathcal{P}$ has a $\sigma$-algebra $\mathcal{G}$ such that $P \mapsto P^n(A)$ is measurable for all $n \geq 1$ and $A \in \mathcal{B}^n$, and a prior $\Pi : \mathcal{G} \to [0,1]$.

Theorem 8.7. Let a model $(\mathcal{P}, \mathcal{G}, \Pi)$ with hypotheses $B, V \in \mathcal{G}$ be given, with $\Pi(B) > 0, \Pi(V) > 0$. The following are equivalent,

i. there exists a Bayesian test sequence for $B$ versus $V$,

ii. there are test functions $\phi_n : \mathcal{X}^n \to [0,1]$ such that for $\Pi$-almost-all $P \in B, Q \in V$,

$$P^n\phi_n \to 0, \quad Q^n(1 - \phi_n) \to 0,$$

iii. there are test functions $\phi_n : \mathcal{X}^n \to [0,1]$ such that,

$$\int_B P^n \phi_n d\Pi(P) + \int_V Q^n(1 - \phi_n) d\Pi(Q) \to 0,$$
iv. for $\Pi$-almost-all $P \in B, Q \in V$, 

$$
\Pi(V|X^n) \xrightarrow{P.a.s.} 0, \quad \Pi(B|X^n) \xrightarrow{Q} 0.
$$

Proof. Because $0 \leq \phi_n \leq 1$, i. implies ii.; by dominated convergence, ii. implies iii.; iii. leads to iv. through Martingale convergence and the inequality (see lemma 2.2 in [146]),

$$
\int_B P^n \Pi(V|X^n) d\Pi(P) \leq \int_B P^n \phi_n d\Pi(P) + \int_V Q^n (1 - \phi_n) d\Pi(Q),
$$

which holds for all $n \geq 1$ and any test sequence $(\phi_n)$; iv. gives i. when we set $\phi_n = \Pi(V|X^n)$. For details, see theorem 2.4 in [146].

An almost-sure version of definition 8.7 is also equivalent, through direct application of proposition 8.4, pointwise in a set of prior mass one. The interpretation of this theorem is gratifying to supporters of the likelihood principle and pure Bayesians: distinctions between model subsets are Bayesian testable, if and only if, they are picked up by the posterior asymptotically and the posterior itself can be viewed as the test function.

Breiman, Le Cam and Schwartz (1964) [47] provide a careful measurability argument to explain the essence of Doob’s consistency theorem. The astonishing generality of Doob’s theorem comes from the measure-theoretical (rather than topological) answer to questions related to posterior convergence. (Although the original reference for these notions is [47], a more complete exposé is found in Le Cam (1986) [162].)

**Definition 8.5.** Let $\mathcal{P}$ be a model with prior $\Pi$. An event $B \in \mathcal{B}(\infty)$ is called a $\Pi$-zero-one set, if $P^\infty(B) = P^\infty(B)^2$, for $\Pi$-almost-all $P \in \mathcal{P}$. A model subset $G \in \mathcal{G}$ is called a $\Pi$-one set if there exists a $\Pi$-zero-one set $B$ such that $G = \{P \in \mathcal{P}: P^\infty(B) = 1\}$.

The collection of all $\Pi$-one sets forms a sub-$\sigma$-algebra of $\mathcal{G}$, which we denote by $\mathcal{G}_1$. Let $\mathcal{G}_0$ denote the initial $\sigma$-algebra for the collection $\{P \mapsto P(A): A \in \mathcal{B}\}$ (which coincides with the initial $\sigma$-algebra for the collection $\{P \mapsto P^\infty(A): A \in \mathcal{B}(\infty)\}$, see lemma 3.10 in [208]). Then $\mathcal{G}_0$ is contained in the Borel $\sigma$-algebra for $\mathcal{F}_1$. In order to make the next argument, we assume that the domain of the prior contains $\mathcal{G}_0$, for example if the prior is Borel for $\mathcal{F}_1, \mathcal{F}_\infty$ or total-variation. Asymptotic posterior convergence is then fully specified by the following observation.

**Proposition 8.6.** Let $\mathcal{P}$ be a model with prior $\Pi$ that contains $\mathcal{G}_0$. Let $V$ be a $\Pi$-one set. Then,

$$
\Pi(V|X^n) \xrightarrow{P.a.s.} 1_V(P),
$$

for $\Pi$-almost-all $P \in \mathcal{P}$.

Proof. Define the products $\Omega_n = \mathcal{P} \times \mathcal{B}(\infty)$ with full product $\sigma$-algebras $\sigma(\mathcal{G} \times \mathcal{B}(\infty))$ and sub-$\sigma$-algebras $\mathcal{F}_n = \{\emptyset, \mathcal{G}, \mathcal{P}\} \times \mathcal{B}(\infty)$. The product $\Omega = \mathcal{P} \times \mathcal{B}(\infty)$ with full product $\sigma$-algebra $\sigma(\mathcal{G} \times \mathcal{B}(\infty))$ has a sub-$\sigma$-algebra $\mathcal{F}_\infty = \{\emptyset, \mathcal{G}, \mathcal{P}\} \times \mathcal{B}(\infty)$, and the
f filtration \( \{ \mathcal{F}_n : n \geq 1 \} \) has limit \( \mathcal{F}_\infty \). Given a prior \( \Pi \) on \((\mathcal{P}, \mathcal{G})\), a joint distribution \( S : \mathcal{F} \to [0,1] \) on \( \Omega \) is fixed by defining \( S(A \times B) = \int_A P^\omega(B) d\Pi(P) \) for \( A \in \mathcal{G} \) and \( B \in \mathcal{B}^\omega \) (which requires measurability of \( P \mapsto P^\omega(B) \) for \( B \in \mathcal{B}^\omega \)). For any \( \mathcal{F} \)-measurable \( g : \Omega \to [0,1] \) the conditional expectations \( \{ E_S[g|\mathcal{F}_n] : n \geq 1 \} \) form a martingale. If, with slight abuse of notation, we maintain \( 1_V \) for the function \( \Omega \to [0,1] : (P, x_\omega) \mapsto 1_V(P) \), we observe that the posteriors \( \Pi(V|X^n) = E[1_V|\mathcal{F}_n] \) form a martingale relative to \( S \). According to Doob’s martingale convergence theorem there exists an \( \mathcal{F}_\infty \)-measurable random variable \( f_\nu \) such that \( \Pi(V|X^n) \to f_\nu \), \( S \)-almost-surely. Since \( V \) is a \( \Pi \)-one set, there exists an event \( B \in \mathcal{B}^\omega \) such that \( 1_V(P) = 1_B(x_\omega) \), \( S \)-almost-surely. Hence,

\[
\Pi(V|X^n) = E_S[1_V|\mathcal{F}_n] = E_S[1_B|\mathcal{F}_n] \to E_S[1_B|\mathcal{F}_\infty] = 1_B = 1_V,
\]

\( S \)-almost-surely, which amounts to, \( P \)-almost-surely for \( \Pi \)-almost-all \( P \) (by Fubini’s theorem).

The remaining question, then, is whether the \( \sigma \)-algebra of \( \Pi \)-one sets is large enough to be interesting. The answer is given in proposition 2 of section 17.7 in Le Cam (1986) [162]: if the model is a Hausdorff space with Radon prior \( \Pi \) and the \( \sigma \)-field \( \mathcal{B} \) on \( \mathcal{X} \) is countably generated, then \( \mathcal{G} = \mathcal{G}_1 \). This implies Doob’s consistency theorem (since any prior on a Polish space is Radon) and more, c.f. the corollary to proposition 2 of section 17.7 in [162] (beaware of some typos and omissions in the proofs). We summarize and conclude as follows.

**Theorem 8.8.** Let \((\mathcal{P}, \mathcal{G})\) be a measurable model, with a prior \( \Pi \) that is a Radon measure, and hypotheses \( B, V \). There is a Bayesian test sequence for \( B \) versus \( V \), if and only if, \( B, V \) are \( \mathcal{G} \)-measurable.

**Proof.** In order for the definition of Bayesian testing to make sense, it is necessary that \( B \) and \( V \) are measurable. Conversely, if \( B \) is measurable and \( V \subset \mathcal{P} \setminus \mathcal{B} \), then \( \phi_n(X^n) = \Pi(\mathcal{P} \setminus B|X^n) \) is a Bayesian test sequence for \( B \) versus \( V \).

**Example 8.11.** Is Cover’s rational mean problem Bayesian testable?

Let’s revisit Cover’s rational mean problem to illustrate the Bayesian answer: consider priors \( \Pi_B \) and \( \Pi_V \) for \( B = [0,1] \cap \mathbb{Q} \) and \( V = [0,1] \setminus \mathbb{Q} \) such that \( \Pi_B(B) = 1 \) and \( \Pi_V(V) = 1 \), (for example, enumerate \( [0,1] \cap \mathbb{Q} = \{ q_i : i \geq 1 \} \) and define, for every measurable \( F \subset [0,1] \),

\[
\Pi_B(F) = \sum_{i \geq 1} 2^{-i} 1_F(q_i). \tag{8.15}
\]

For \( \Pi_V \) we may simply choose Lebesgue measure. Set \( \Pi = \frac{1}{2} \Pi_B + \frac{1}{2} \Pi_V \) on \([0,1] \). As we have seen in example 8.7, the compact space \([0,1] \) is homeomorphic to the model \( \mathcal{P} = \{ P_p : p \in [0,1] \} \) with the \( \mathcal{I}_\infty \)-topology through \( p \mapsto P_p \). Therefore, \( \mathcal{P} \) is Polish for \( \mathcal{I}_\infty \), which implies that \( \Pi \) is Radon. Since \( B \) is Borel measurable in \([0,1] \), the corresponding subset of \( \mathcal{P} \) is Borel measurable for \( \mathcal{I}_\infty \), implying Bayesian testability of \( B \) versus \( V \). Proposition 8.6 even strengthens that to,

\[
\Pi(p \in V|X^n) \xrightarrow{P-\text{a.s.}} 0, \quad \Pi(p \in V|X^n) \xrightarrow{P-\text{a.s.}} 1,
\]
for \( q \in [0, 1] \cap \mathbb{Q} \) and \( r \in [0, 1] \setminus \mathbb{Q} \). So the tests \( \phi(X^n) = \Pi(p \in V | X^n) \) (or the indicators for posterior odds of proposition 8.8) have property (D), albeit with a \( \Pi \)-null-set of exceptions. Indeed corollary 8.2 and example 8.7 establish that this \( \Pi \)-null-set is non-empty. So Cover’s rational mean problem does have a Bayesian type solution. (It appears [56] that D. Blackwell made Cover aware of a Bayesian approach leading to a solution of the rational mean problem but failed to convince him fully of the validity of his alternative.)

To conclude this section, we provide an unexpected frequentist consequence of the Bayesian considerations of this section.

**Theorem 8.9.** Let \( \mathcal{P} \) be a model that is countable. Any \( B, V \subset \mathcal{P} \) with \( B \cap V = \emptyset \) are pointwise testable.

**Proof.** For any two \( P, Q \in \mathcal{P} \) there exists a measurable \( 0 \leq f \leq 1 \) such that \( Pf \neq Qf \), so \( \mathcal{T}_1 \) is the discrete topology on \( \mathcal{P} \) (and so is \( \mathcal{T}_0 \)). Any countable discrete space is Polish and the corresponding Borel \( \sigma \)-algebra is the power set of \( \mathcal{P} \). Pick any (Borel) prior \( \Pi \) on \( \mathcal{P} \) such that \( \Pi(\{P\}) > 0 \) for all \( P \in \mathcal{P} \). Any \( V \) is Bayesian testable versus any disjoint \( B \) under \( \Pi \) and the test functions \( \phi_n(X_1, \ldots, X_n) = \Pi(V | X_1, \ldots, X_n) \) form a Bayesian test sequence for \( B \) versus \( V \). Because the only null-set of the prior is \( \emptyset \), Bayesian test sequences under \( \Pi \) are also pointwise test sequences.

That means that the example of hypotheses (8.2) has full validity as a frequentist procedure.

**Corollary 8.7.** *(Dembo and Peres (1994))*

Regarding the parameter \( p \in [0, 1] \) for i.i.d.-Bernoulli-\( p \) distributed \( X_1, X_2, \ldots, \), there exists a pointwise test sequence that distinguishes,

\[
H_0 : p \in [0, 1] \cap \mathbb{Q}, \quad H_1 : p \in [0, 1] \cap \sqrt{2} + \mathbb{Q},
\]

asymptotically.

### 8.6 Bayesian testing power and model selection for frequentists

Proposition 8.6 settles the Bayesian question, but with Bayesian tests, more is possible. In the frequentist, constructive answer to the testability question, we shall require control over the power of the tests. The following proposition from [146] formulates a general upper bound based on barycentres. (Denote the density for the local prior predictive distribution \( P_{\Pi}^{B} \) with respect to \( \mu_n = P_{\Pi}^{B} + P_{\Pi}^{V} \) by \( p_{B,n} \), and similar for \( P_{\Pi}^{V} \).)

**Proposition 8.7.** Let \( (\mathcal{P}, \mathcal{G}) \) be a model with priors \( (\Pi_n) \) and two measurable model subsets \( B, V \) with \( \Pi(B), \Pi(V) > 0 \). For every \( n \geq 1 \), there exists a \( \phi_n : \mathcal{X}^n \to [0, 1] \) such that,
\begin{align}
\int_B P^n \phi_n d\Pi(P) + \int_Q Q^n (1 - \phi_n) d\Pi(Q) & \leq \int (\Pi(B) p_{B,n}(x))^{\alpha} \left(\Pi(V) p_{V,n}(x)\right)^{1-\alpha} d\mu_n(x), \\
\text{(8.16)}
\end{align}

for any $0 \leq \alpha \leq 1$.

**Proof.** See proposition 2.6 in [146].

The following demonstrates that a sequence of tests based on posterior odds (or Bayes factors) is optimal, and thus obeys any upper bound for Bayesian testing power, including that of proposition 8.7 and the exponential bounds that follow from uniformly testable hypotheses and proposition 8.1.

**Proposition 8.8.** Let $(\mathcal{P}, \mathcal{G})$ be a model with priors $(\Pi_n)$ and measurable model subsets $B, V$. For every $n \geq 1$, the test $\phi_n(X^n) = 1\{X^n : \Pi(V|X^n) \geq \Pi(B|X^n)\}$ based on posterior odds has optimal Bayesian testing power.

**Proof.** Consider the decision-theoretic problem of setting the optimal $\phi \in [0,1]$ for picking $B$ or $V$ based on the data, with loss $\ell: \mathcal{P} \times [0,1] \rightarrow [0,1]$, 

$$
\ell(P, \phi) = \begin{cases} 
0, & \text{if } P \not\in B \cup V, \\
|\phi - 1_V(P)|, & \text{if } P \in B \cup V.
\end{cases}
$$

Data-driven decisions $\phi_n(X^n)$ for all $n \geq 1$ are test functions. The Bayesian risk functions,

$$
r_n(\phi_n, \Pi) = \int_{\mathcal{P}} P^n \ell(P, \phi_n) d\Pi(P),
$$

equal the Bayesian testing power,

$$
r_n(\phi_n, \Pi) = \int_B P^n |\phi_n - 1_V(P)| d\Pi(P) + \int_Q Q^n |\phi_n - 1_V(Q)| d\Pi(Q)
= \int_B P^n \phi_n d\Pi(P) + \int_Q Q^n (1 - \phi_n) d\Pi(Q),
$$

for all $n \geq 1$. Bayes’s rule implies that if, for all $n \geq 1$ and $P^n$-almost-all $x^n \in \mathcal{X}^n$, $\phi_n(x^n)$ is the minimizer,

$$
\int_{\mathcal{P}} \ell(P, \phi_n(x^n)) d\Pi(P|X^n = x^n) = \inf_{\psi \in [0,1]} \int_{\mathcal{P}} \ell(P, \psi) d\Pi(P|X^n = x^n),
$$

then $\phi_n(x^n)$ optimizes Bayesian testing power:

$$
r_n(\phi_n, \Pi) = \inf_{\psi_n} r_n(\psi_n, \Pi),
$$

(whence the infimum runs over all possible choices $\psi_n: \mathcal{X}^n \rightarrow [0,1]$ for the $n$-th test function). To conclude, note that,
\[
\int_\mathcal{P} \ell(P, \psi_n(x^n)) d\Pi(P|X^n = x^n)
= \int_B \psi_n(x^n) d\Pi(P|X^n = x^n) + \int_V (1 - \psi_n(x^n)) d\Pi(Q|X^n = x^n)
= \psi_n(x^n) \Pi(B|X^n = x^n) + (1 - \psi_n(x^n)) \Pi(V|X^n = x^n),
\]
is minimal if we choose \(\psi_n(x^n) = 1\{x^n : \Pi(V|X^n) \geq \Pi(B|X^n)\}\).

We appeal to a theorem from [146] to make the final step in the proof that the existence of sufficiently powerful Bayesian tests, in combination with the requirement of remote contiguity (see definition 3.1 in [146]) of the local prior predictive distributions \(P_\Pi^n\) with respect to the true distribution of the data \(P^n\), implies that posteriors select the correct underlying hypothesis with probability growing to one.

**Theorem 8.10.** For all \(n \geq 1\), let the model be a measurable space \((\mathcal{P}, \mathcal{G})\) with priors \(\Pi_n : \mathcal{G} \to [0, 1]\). Consider disjoint, measurable \(B, V \subset \Theta\) with \(\Pi_n(B), \Pi_n(V) > 0\) such that,

i. There exist Bayesian tests for \(B\) versus \(V\) of power \(a_n \downarrow 0\),
\[
\int_B P^n \phi_n d\Pi_n(P) + \int_V Q^n (1 - \phi_n) d\Pi_n(Q) = o(a_n),
\]

ii. For every \(P \in B, P^n \preceq a_n^{-1} P^n_\Pi(B)\), and for every \(Q \in V, Q^n \preceq a_n^{-1} P^n_\Pi(V)\).

Then the indicators \(\phi_n(X^n) = 1\{X^n : \Pi(V|X^n) \geq \Pi(B|X^n)\}\) for posterior odds form a pointwise test sequence for \(B\) versus \(V\).

So if we can find a sequence of priors for which, (a.) we can prove the existence of a suitably powerful Bayesian test sequence, and which, (b.) induces remote contiguity at the right rate, the resulting posterior forms a constructive means (through posterior odds) to achieve consistent frequentist model selection.

### 8.7 Conclusions and discussion

Theorem 8.6 does not leave the model choice completely free, a dominated model is required. This restriction, however annoying, is of an essential nature because it is ultimately due to the sequential nature of the i.i.d. experiment. The shortest way to explain this is as follows. Using the Le Cam-Schwartz theorem to prove the existence of tests, the most peculiar aspect of the conditions is the fact that a sequence of uniformly continuous functions is required. In the context of weak topologies, which are not first-countable in general, requiring existence of a convergent net or filter is natural but requiring existence of convergent sequences poses a considerable extra burden. If the weak topology in question happens to be metrizable, like Prokhorov’s weak topology and the \(\sigma(E, F)\)-topology on the norm-bounded subset \(\overline{H}\) of lemma 8.2, first-countability is recovered and sequential convergence coincides with net convergence, but \(\mathcal{Z}_\infty\) is not first-countable in general. Earlier work
[64, 181, 86] solves this problem by requiring domination and uniform integrability, using $\mathcal{I}_w$-compactness c.f. the Dunford-Pettis theorem to equate $\mathcal{I}_w$ to Prokhorov’s weak topology. In our argument, the problem is lifted by the (admittedly only sufficient) condition that the model is dominated. Given that the reason for this restriction (see the proof of theorem 8.3) is the sequential nature of i.i.d. experiments, it seems unlikely that there is a formulation of theorem 8.6 for non-dominated models of the same or very similar form.

### 8.7.1 Model assumptions

“There are statistical questions that I shouldn’t even be thinking about... I can’t afford to waste my time like that.”

### 8.7.2 Model selection

Let $\mathcal{P}$ be a model for i.i.d. data $X^n \sim P^n$, ($n \geq 1$), consisting of $M \geq 1$ disjoint sub-models, $\mathcal{P} = \mathcal{P}_1 \cup \ldots \cup \mathcal{P}_M$. Assume that $P \in \mathcal{P}$. The simplest form that the model-selection question takes, is to require asymptotically consistent selection of the sub-model $\mathcal{P}_i$ such that $P \in \mathcal{P}_i$.

### 8.8 Exercises [EMPTY]
Chapter 9
Application: non-parametric errors-in-variables regression

To demonstrate in a typical way how the methods presented in chapter 6 are applied in practice, we consider the asymptotic behaviour of the posterior distribution for the errors-in-variables model. The model describes measurements consisting of paired observations \((X, Y)\) that are represented in terms of an unobserved \(Z\). The random variable \(Z\) is related to \(X\) directly and to \(Y\) through a regression function \(f\), both perturbed by Gaussian errors. We assume that \(Z\) falls into a (known) bounded subset of the real line with probability one, but otherwise leave its distribution unconstrained. In the semi-parametric literature, the regression function comes from a parametric (see Taupin (2001) [212]), or even linear (see, e.g. Anderson (1984) [6]) class of functions. In the following, we broaden that assumption to non-parametric regression classes, discussing the errors-in-variables problem also for Lipschitz and smooth functions.

Hence, the formulation we use involves two non-parametric components, the distribution of \(Z\) and the regression function \(f\). We give Hellinger rates of convergence for the posterior distribution of the errors-in-variables density in non-parametric and parametric regression classes, using the posterior rate-of-convergence theorem 6.6 (or rather, a version based on the Hellinger metric entropy, c.f. Ghosal et al. (2000) [101]). Conditions that bound the rate of convergence can be decomposed into two terms, one for each of the non-parametric components of the model. The rate is then determined by the term that dominates the bound.

9.1 Errors-in-variables regression

When data is observed in pairs \((X, Y) \in \mathbb{R}^2\) and there is reason to assume that there is some unknown functional relation \(f : \mathbb{R} \to \mathbb{R}\) between \(X\) and \(Y\), observed with an additive regression error \(e \in \mathbb{R}\), the most straightforward model is,

\[
Y = f(X) + e. \tag{9.1}
\]
Estimation then occurs in a family of possible regression functions $f$ based on a sample $(X_i, Y_i), i \geq 1$, usually assuming that the $e_i$ form an i.i.d. sample from some error distribution, typically with expectation equal to 0 (and commonly a known normal distribution). However popular, this model suffers from a serious shortcoming: it assumes that $X$ has been observed with infinite precision, while very often there is some uncertainty in the observation of $X$. This uncertainty causes what is known as attenuation bias: because the observed $X$ are only a noisy reflection of some unobserved quantity that determines the value of $Y$ through $f$, the calculation is “blurred” horizontally. In the common case of an unknown linear $f(x) = ax + b$, estimation of $a$ in the model (9.1) is biased towards 0, attenuating the regression function and making it resemble a constant function more appears reasonable graphically.

To prevent this, the errors-in-variables model studies a version of the regression model that takes the error in observed $X$ into account explicitly: pairs $(X,Y)$ are assumed to be distributed as,

$$
X = Z + e_1, \\
Y = f(Z) + e_2,
$$

(9.2)

where $(e_1,e_2)$ and $Z$ are independent and $f : \mathbb{R} \to \mathbb{R}$ belongs to a family of regression functions. Usually, the distribution of the errors $(e_1,e_2)$ is assumed to be known up to a (finite-dimensional) parameter $\sigma$ whereas the distribution $F$ of $Z$ is completely unknown in the most general case. The primary interest lies in estimation of the regression function $f$ from a i.i.d. sample of pairs $(X_1,Y_1), (X_2,Y_2), \ldots, (X_n,Y_n)$ in the presence of the nuisance parameter $F$. Applications include all situations in which a functional dependence between measurements with errors is to be established.

The primary difference between errors-in-variables and ordinary regression using a set of design points $x_1, \ldots, x_n$, is the stochastic nature of the variable $X$. Regarding $X$, the variable $e_1$ is referred to as the “random error”, whereas the variability of $Z$ is said to be the “systematic error” (Anderson (1984) [6]). Kendall and Stuart (1979) [135] distinguish between the “functional” errors-in-variables problem, in which $Z$ is non-stochastic, taking on the values of ‘design points’ $z_1, \ldots, z_n$, and the “structural” errors-in-variables problem, in which $Z$ is stochastic. Best known is linear errors-in-variables regression, in which $f$ is assumed to depend linearly on $z$ (see, e.g. [6] for an extensive overview of the literature). Efficient estimators for the parameters of $f$ have been constructed by Bickel and Ritov (1987) [28], Bickel et al. (1998) [29] and Van der Vaart (1988, 1996) [216, 217]. Errors-in-variables regression involving a parametric family of non-linear regression functions has been analysed by Taupin and others (see Taupin (2001) [212] and references therein). In Fan and Troung (1993) [88], the rate of convergence (in a weighted $L_2$-sense) of Nadaraya-Watson-type kernel estimators for the conditional expectation of $Y$ given $Z$ (and hence for the regression function) are considered using deconvolution methods.

In this chapter we analyse the structural errors-in-variables problem for non-parametric families of regression functions in a Bayesian setting; we consider the behaviour of posterior distributions for the parameter $(\sigma,f,F)$ in the asymptotic
9.1 Errors-in-variables regression

It is stressed that in this formulation, the errors-in-variables problem has two non-parametric components, one being the distribution of the underlying variable $Z$ and the other the regression function. The emphasis lies on the interplay between these two non-parametric aspects of the model, as illustrated by their respective contributions to the rate of convergence (see, e.g., theorems 9.3 and 9.4).

9.1.1 Definition of the EIV model

We assume throughout this chapter that there is some known constant $A > 0$ such that $Z \in [-A, A]$ with probability one. Furthermore, we assume (unless indicated otherwise) that the errors $e_1$ and $e_2$ are independent and distributed according to the same normal distribution $\Phi_{\sigma}$ on $\mathbb{R}$ with mean zero and variance $\sigma^2$ (i.e., a special case of restricted Gaussian errors in the terminology of [28]). Writing $\phi_{\sigma}$ for the normal density of both $e_1$ and $e_2$, the model consists of a family of distributions for the observations $(X, Y)$, parametrized by $(\sigma, f, F) \in I \times \mathcal{F} \times D$, where it is assumed that:

(a) $I$ is a closed interval in the positive reals, bounded away from zero and infinity, i.e., $I = [\sigma, \overline{\sigma}] \subset (0, \infty)$.
(b) $D$ is the collection of all probability distributions on the compact symmetric interval $[-A, A]$, parametrized by all corresponding Stieltjes functions $F$.
(c) $\mathcal{F} \subseteq C_b[-A, A] \subset C[-A, A]$ is a bounded family of continuous regression functions $f : [-A, A] \rightarrow [-B, B]$. We shall distinguish various cases by further requirements, including equicontinuity, Lipschitz- and smoothness-bounds. Also considered is the parametric case, in which $\mathcal{F}$ is parametrized by a subset of $\mathbb{R}^k$.

For all $(\sigma, f, F) \in I \times \mathcal{F} \times D$, we define the following convoluted density for the distribution of observed pair $(X, Y)$:

$$p_{\sigma, f, F}(x, y) = \int_{\mathbb{R}} \phi_{\sigma}(x - z) \phi_{\sigma}(y - f(z)) dF(z),$$

(9.3)

for all $(x, y) \in \mathbb{R}^2$.

It is stressed that when we speak of the errors-in-variables model $\mathcal{P}$, we refer to the collection of probability measures $P_{\sigma, f, F}$ on $\mathbb{R}^2$ defined by the Lebesgue-densities parametrized in the above display (rather than the parameter space $I \times \mathcal{F} \times D$). In many cases we regard $\mathcal{P}$ as a metric space, using either the Hellinger metric or $L_1(\mu)$-norm. As far as the parameter space is concerned, the model may not be identifiable: if, for given $F \in D$, two regression functions $f, g \in \mathcal{F}$ differ only on a set of $F$-measure zero, the corresponding densities $p_{\sigma, f, F}$ and $p_{\sigma, g, F}$ are equal on all of $\mathbb{R}^2$ (for all $\sigma \in I$). Determination of the true regression function $f_0$ based on an i.i.d. $P_0$-distributed sample can therefore be done only $F_0$-almost-everywhere (where $P_0 = P_{\sigma_0, f_0, F_0}$). Ultimately, the results we give are based on the Hellinger distance, which, in the present circumstances, gives rise to a semi-metric on the parameter space $I \times \mathcal{F} \times D$ for the same reason. The ‘well-known’ identifiability problems in
the linear errors-in-variables model (see e.g. Reiersøl (1950) [193]) arising due to interchangability of Gaussian components of the distribution of $Z$ with the error-distribution (see also [6] and [28]) do not occur in our considerations, because we assume the distribution of $Z$ to be compactly supported.

### 9.1.2 Posterior concentration theorem

Conditions for the theorem on Bayesian rates of convergence are again formulated in terms of the specific Kullback-Leibler neighbourhoods (6.23) of $P_0 \in \mathcal{P}$. Recall the Ghosh-Ghosal-van der Vaart theorem, which we write here with the help of entropy condition (6.31), where $N(\varepsilon, \mathcal{P}, H)$ denote the covering numbers with respect to the Hellinger metric on $\mathcal{P}$, i.e. the minimal number of Hellinger balls of radius $\varepsilon > 0$ needed to cover $\mathcal{P}$.

**Theorem 9.1.** Let $\mathcal{P}$ be a model and assume that the sample $U_1, U_2, \ldots$ is i.i.d. $P_0$-distributed for some $P_0 \in \mathcal{P}$. For a given prior $\Pi$, suppose that there exists a sequence of strictly positive numbers $\varepsilon_n$ with $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$ and constants $R_1, R_2 > 0$, such that:

\[
\Pi(B(\varepsilon_n; P_0)) \geq e^{-R_1 n\varepsilon_n^2}, \\
\log N(\varepsilon_n, \mathcal{P}, H) \leq R_2 n\varepsilon_n^2,
\]

for all large enough $n$. Then, for every sufficiently large constant $M$, the posterior distribution satisfies:

\[
\Pi_n(P \in \mathcal{P} : H(P, P_0) \geq M\varepsilon_n \mid U_1, \ldots, U_n) \to 0,
\]

as $n \to \infty$, in $P_0$-expectation.

The assumption that the model is well-specified, i.e. $P_0 \in \mathcal{P}$, can be relaxed. In Kleijn and Van der Vaart (2003) [142], the above theorem is given in the case of a misspecified model. We do not give misspecified versions of the results, although we believe that the conditions of the necessary theorems in [142] are met in the model we consider.

### 9.2 Rates of posterior convergence in function spaces

#### 9.2.1 Lipschitz and smoothness classes

We consider regression classes $\mathcal{F}$ contained within the class $C_B[-A, A]$ of all continuous functions $f : [A, A] \to \mathbb{R}$ bounded by a (known) constant $B > 0$. At the very least, we also require equicontinuity of $\mathcal{F}$, which guarantees compactness in the
topology of the uniform norm $\| \cdot \|$ according to the Arzelà-Ascoli theorem. Consequently, covering numbers $N(\varepsilon, \mathcal{F}, \| \cdot \|)$ are finite and an important part of the argument rests on bounds on these covering numbers we establish later. We distinguish several non-parametric and parametric examples of such classes below, but remark that other regression classes for which bounds on covering numbers exist, can also be used.

(i) $\text{Lip}_M(\alpha)$ (for some $M > 0$ and $0 < \alpha \leq 1$), the class of all Lipschitz functions $f \in \mathcal{C}_B[-A,A]$ with constant $M$ and exponent $\alpha$, i.e.

$$|f(z) - f(z')| \leq M|z - z'|^\alpha,$$

for all $z, z' \in [-A,A]$.

(ii) $D_{\alpha,M}(q)$ (for some $0 < \alpha \leq 1$, $M > 0$ and an integer $q \geq 1$), the class of all $q$ times differentiable functions $f \in \mathcal{C}_B[-A,A]$ for which the $q$-th derivative $f^{(q)}$ belongs to $\text{Lip}_M(\alpha)$.

(iii) $\mathcal{F}_\Theta$, a parametric class of regression functions which forms a subset of $\text{Lip}_M(\alpha)$ for some $\alpha \in (0,1]$ and $M > 0$. We assume that there exists a bounded, open subset $\Theta \subset \mathbb{R}^k$ for some $k \geq 1$ such that $\mathcal{F}_\Theta = \{f_\theta : \theta \in \Theta\}$. Furthermore, we assume that the map $\theta \mapsto f_\theta$ is Lipschitz-continuous, i.e. there exist constants $L > 0$ and $\rho \in (0,1]$ such that for all $\theta_1, \theta_2 \in \Theta$:

$$\|f_{\theta_1} - f_{\theta_2}\| \leq L\|\theta_1 - \theta_2\|^\rho.$$  

Often, it is more convenient to unify cases (i) and (ii) above, by considering the family of classes $\mathcal{C}_{\beta,L}[-A,A]$ defined as follows. For given $\beta > 0$ and $L > 0$, we define $\beta$ to be the greatest integer such that $\beta < \beta$ and we consider, for suitable functions $f : [-A,A] \to \mathbb{R}$, the norm:

$$\|f\|_{\beta} = \max_{k \leq \beta} \|f^{(k)}\| + \sup_{z_1, z_2} \frac{|f^{(\beta)}(z_1) - f^{(\beta)}(z_2)|}{|z_1 - z_2|^\alpha},$$

where the supremum is taken over all pairs $(z_1, z_2) \in [-A,A]^2$ such that $z_1 \neq z_2$. The class $\mathcal{C}_{\beta,L}[-A,A]$ is then taken to be the collection of all continuous $f : [-A,A] \to \mathbb{R}$ for which $\|f\|_{\beta} \leq L$. Note that for $0 < \beta \leq 1$, $\beta = 0$ and $\mathcal{C}_{\beta,L}[-A,A]$ is a Lipschitz class bounded by $L$; if $\beta > 1$, differentiability of a certain order is implied, as well as boundedness of all derivatives and a Lipschitz property for the highest derivative.

### 9.2.2 Competing entropy bounds

As indicated in subsection 9.1.2, the Hellinger rate of convergence $\varepsilon_n$ is bounded by two conditions, one related to the small-$\varepsilon$ behaviour of the (Hellinger) entropy of the model, the other by the small-$\varepsilon$ behaviour of the prior mass in Kullback-Leibler neighbourhoods of the form (6.23). The first condition is considered in section 9.3:
Theorem 9.3 says that the Hellinger covering number of the errors-in-variables model has an upper bound that consists of two terms, one resulting from the $(\sigma, F)$-part of the model and the other from minimal covering of the regression class:

$$
\log N(\epsilon, \mathcal{P}, H) \leq L_0 \left( \frac{1}{\epsilon} \right)^3 + \log N(L\epsilon, \mathcal{F}, \|\cdot\|),
$$

(9.9)

for small $\epsilon > 0$ and some constants $L, L_0 > 0$. If the regression class $\mathcal{F}$ is ’small’ enough, in the sense that the first term in the entropy bound displayed above dominates in the limit $\epsilon \to 0$, the candidate rates of convergence $\epsilon_n$ are parametric up to a logarithmic factor.

Lemma 9.1. If there exists a constant $L_1 > 0$ such that:

$$
\log N(\epsilon, \mathcal{F}, \|\cdot\|) \leq L_1 \left( \frac{1}{\epsilon} \right)^3,
$$

(9.10)

for small enough $\epsilon > 0$, then the entropy condition (9.5) is satisfied by the sequence:

$$
\epsilon_n = n^{-1/2}(\log n)^{3/2},
$$

(9.11)

for large enough $n$.

Proof. Under the above assumption, $\log N(\epsilon, \mathcal{P}, H)$ is upper bounded by the first term in (9.9) with a larger choice for the constant. Note that the sequence $\epsilon_n$ as defined in (9.11) satisfies $\epsilon_n \downarrow 0$ and $n\epsilon_n^2 \to \infty$. Also note that $\epsilon_n \geq 1/n$ for large enough $n$, so that for some $L > 0$,

$$
\log N(\epsilon_n, \mathcal{F}, \|\cdot\|) \leq \log N(1/n, \mathcal{F}, \|\cdot\|) \leq L(\log n)^3,
$$

and $n\epsilon_n^2 = (\log n)^3$, which proves that $\epsilon_n$ satisfies (9.5).

It is also possible that the small-$\epsilon$ behaviour of the errors-in-variables entropy is dominated by the covering numbers of the regression class. In that case the r.h.s. of (9.9) is replaced by a single term proportional to $\log N(L\epsilon, \mathcal{F}, \|\cdot\|)$ for small enough $\epsilon$. If the regression functions constitute a Lipschitz or smoothness class, lemma 9.13 gives the appropriate upper bound for the entropy, leading to the following candidate rates of convergence.

Lemma 9.2. For an errors-in-variables model $\mathcal{P}$ based on a regression class $C_{\beta, M}[\!-A, A]$, the entropy condition (9.5) is satisfied by the sequence:

$$
\epsilon_n = n^{-\beta/(2\beta+1)},
$$

(9.12)

for large enough $n$.

Proof. As argued above, the Hellinger entropy of the errors-in-variables model is upper-bounded as follows:
9.2 Rates of posterior convergence in function spaces

\[ \log N(\varepsilon, \mathcal{P}, H) \leq \frac{K}{\varepsilon^{1/\beta}}, \]

for some constant \( K > 0 \) and small enough \( \varepsilon \). The sequence \( \varepsilon_n \) satisfies \( \varepsilon_n \downarrow 0 \) and \( n\varepsilon_n^2 \rightarrow \infty \). Furthermore, note that:

\[ \log N(\varepsilon_n, \mathcal{P}, H) \leq K n^{1/(2\beta+1)} = Kn \cdot n^{-\frac{2\beta}{2\beta+1}} = Kn\varepsilon_n^2, \]

for large enough \( n \).

### 9.2.3 Competing lower bounds on prior mass

Similar reasoning applies to condition (9.4) for the small-\( \varepsilon \) behaviour of the prior mass of Kullback-Leibler neighbourhoods of the form (6.23). Section 9.4 discusses the necessary lemmas in detail. We define priors \( \Pi_I, \Pi_F \) and \( \Pi_D \) on the parametrizing spaces \( I, \mathcal{F} \) and \( D \) respectively and choose the prior \( \Pi \) on the model \( \mathcal{P} \) as induced by their product under the map \( (\sigma, f, F) \mapsto P_{\sigma, f, F} \) (which is measurable, as shown in lemma 9.10). The prior \( \Pi_I \) is chosen as a probability measure on \( I \) with continuous and strictly positive density with respect to the Lebesgue measure on \( I \). Priors for the various regression classes discussed in the beginning of this section are discussed in subsection 9.5.2. The prior \( \Pi_D \) on \( D \) is based on a Dirichlet process with base measure \( \alpha \) which has a continuous and strictly positive density on all of \([ -A, A ]\).

As with the covering numbers discussed above, we find (see theorem 9.4) that (the logarithm of) the prior mass of Kullback-Leibler neighbourhoods is lower bounded by two terms, one originating from the prior on the regression class and the other from the priors on the remaining parameters in the model:

\[ \log \Pi \left( B(\delta \log(1/\delta); P_0) \right) \geq -c \left( \log \frac{1}{\delta} \right)^3 + \log \Pi_{\mathcal{F}} \left( f \in \mathcal{F} : \|f - f_0\| \leq \delta \right), \]

(9.13)

for some constants \( K, c > 0 \) and small enough \( \delta > 0 \). If the prior mass in \( \mathcal{F} \) around the true regression function \( f_0 \) does not decrease too quickly with decreasing \( \delta \), the bound that dominates (9.13) is proportional to the first term on the r.h.s., which leads to near-parametric candidate rates of convergence.

**Lemma 9.3.** If there exists a constant \( c' > 0 \) such that:

\[ \log \Pi_{\mathcal{F}} \left( f \in \mathcal{F} : \|f - f_0\| \leq \varepsilon \right) \geq -c' \left( \log \frac{1}{\varepsilon} \right)^3, \]

(9.14)

for small enough \( \varepsilon > 0 \), then the prior-mass condition (9.4) is satisfied by the sequence (9.11) for large enough \( n \).

**Proof.** Condition (9.14) implies that (9.13) holds with the lower bound on the r.h.s. replaced by only its first term with a larger choice for the constant \( c \). The substitution
Application: non-parametric errors-in-variables regression

\[ \varepsilon = K \delta \log(1/\delta) \] leads to a constant and a \( \log \log(1/\delta) \) correction, both of which are dominated by \( \log(1/\delta) \) for small enough \( \delta \). (See the proof of lemma 9.4, where a similar step is made.) It follows that:

\[
\log \Pi \left( B(\varepsilon; P_0) \right) \geq -c'' \left( \log \frac{1}{\varepsilon} \right)^3,
\]

for some constant \( c'' > 0 \) and small enough \( \varepsilon \). The remainder of the proof is identical to that of lemma 9.1.

However, it is also possible that the prior mass around \( f_0 \) in the regression class decreases more quickly than (9.14). In that case the lower bound on the r.h.s. of (9.13) is determined by the prior on \( \mathcal{F} \). The following lemma assumes a so-called net-prior on the regression class \( \mathcal{F} \), a construction that is explained in subsection 9.5.2.

**Lemma 9.4.** For an errors-in-variables model \( \mathcal{P} \) based on a regression class \( C_{\beta,M}[-A,A] \) with a net-prior \( \Pi \), the prior-mass condition (9.4) is satisfied by the sequence:

\[
\varepsilon_n = n^{-\beta/(2\beta+1)} \left( \log n \right)^{1/2},
\]

(9.15)

for large enough \( n \).

**Proof.**

Given \( \beta \), the prior mass in neighbourhoods of the true regression function \( f_0 \) for a net prior \( \Pi \) is lower bounded by the expression on the r.h.s. in (9.39). Since this term dominates in the r.h.s. of (9.13) for small \( \delta \), the prior mass of Kullback-Leibler neighbourhoods of \( P_0 \) in \( \mathcal{P} \) satisfies the following lower bound:

\[
\log \Pi \left( B(K\delta \log(1/\delta); P_0) \right) \geq -L \frac{1}{\delta^{1/\beta}},
\]

for some constants \( K, L > 0 \) and small enough \( \delta \). Define \( \varepsilon = K \delta \log(1/\delta) \) and note that, for small enough \( \delta \):

\[
\frac{1}{\varepsilon^{1/\beta}} \left( \log \frac{1}{\varepsilon} \right)^{1/\beta} = K^{-1/\beta} \frac{1}{\delta^{1/\beta}} \left( \log \frac{1}{\delta} \right)^{-1/\beta} \left( \log \frac{1}{\delta} - \log K - \log \log \frac{1}{\delta} \right)^{1/\beta} 
\geq K^{-1/\beta} \frac{1}{\delta^{1/\beta}} \left( \log \frac{1}{\delta} \right)^{-1/\beta} \left( \frac{1}{2} \log \frac{1}{\delta} \right)^{1/\beta} 
\geq \left( \frac{1}{2} \right)^{1/\beta} K^{-1/\beta} \frac{1}{\delta^{1/\beta}}.
\]

For the first inequality in the above display, we have used that \( \log K \leq \log \log \frac{1}{\delta} \leq \frac{1}{4} \log \frac{1}{\delta} \) (for small enough \( \delta \)). We see that there exists a constant \( L' > 0 \), such that, for small enough \( \varepsilon > 0 \):

\[
\log \Pi \left( B(\varepsilon; P_0) \right) \geq -L' \frac{1}{\varepsilon^{1/\beta}} \left( \log \frac{1}{\varepsilon} \right)^{1/\beta}.
\]
9.2 Rates of posterior convergence in function spaces

The sequence $\varepsilon_n$ satisfies $\varepsilon_n \downarrow 0$ and $n\varepsilon_n^2 \to \infty$. Define the sequence $a_n = n^{-\beta/(2\beta+1)}$ and note that $\varepsilon_n \geq a_n$ (for large enough $n$) so that for some constant $R > 0$:

$$\log \Pi(B(\varepsilon_n; P_0)) \geq \log \Pi(B(a_n; P_0)) \geq -L' \frac{1}{a_n^{1/\beta}} \left(\log \frac{1}{a_n}\right)^{1/\beta}$$

$$= -Rn^{\frac{1}{2\beta+1}} (\log n)^{\frac{1}{\beta}} = -R\varepsilon_n^2,$$

for large enough $n$.

9.2.4 Various rates of posterior convergence

In the case of a parametric regression class ($\mathcal{F}_\Theta$ as defined under case (iii) in the beginning of this section) and a prior on $\Theta$ with strictly positive and continuous density, the conditions of lemmas 9.1 and 9.3 are satisfied. From lemma 9.15, we know that in the case of a parametric class of regression functions, covering numbers satisfy (9.10). Furthermore, from lemma 9.17, we know that for a parametric class, the prior mass in neighbourhoods of $f_0$ satisfies (9.14). The resulting conclusion for the rate of convergence in parametric regression classes is given in the theorem below.

We summarize the main results in the following theorem by stating the rates of convergence for the classes defined in the beginning of this section. The proof consists of combination of the preceding lemmas.

**Theorem 9.2.** For the specified regression classes, the assertion of theorem 9.1 holds with the following rates of convergence.

(i) If $\mathcal{F} = \text{Lip}_M(\alpha)$ (for some $\alpha \in (0, 1]$ and $M > 0$) with a net prior, the prior-mass condition for neighbourhoods of $f_0$ in the regression class determines the rate, given by the sequence $\varepsilon_n$ defined in lemma 9.4 with $\beta = \alpha$:

$$\varepsilon_n = n^{-\frac{\alpha}{2\alpha+1}} (\log n)^{\frac{1}{2\alpha+1}}.$$

(ii) If $\mathcal{F} = D_{\alpha,M}(q)$ (for some $M > 0$ and integer $q \geq 1$) with a net prior, the prior-mass condition for neighbourhoods of $f_0$ again determines the rate, given by the sequence $\varepsilon_n$ defined in lemma 9.4 with $\beta = q + \alpha$:

$$\varepsilon_n = n^{-\frac{q+\alpha}{2q+2\alpha-1}} (\log n)^{\frac{1}{2q+2\alpha-1}}.$$

(iii) If $\mathcal{F} = \mathcal{F}_\Theta$ is a parametric class with a prior that has a continuous and strictly positive density throughout $\Theta$, the rate is determined by the posterior convergence with regard to the parameter $F$ and is given by:

$$\varepsilon_n = n^{-1/2} (\log n)^{3/2}.$$
Concerning the parametric rate of convergence, it is stressed that this rate applies to the full, non-parametric problem and can not be compared with semi-parametric rates for estimation of the parameter $\theta$ in the presence of the nuisance parameter $F$. With regard to the logarithmic corrections to the powers of $n$ in the expressions for the rate of convergence in Lipschitz- and smoothness-classes, we note that they originate from (the proof of) lemma 9.4: the logarithm is introduced by the transition from $\delta$ to $\varepsilon$, which compensates for the logarithmic correction in the extent of the Kullback-Leibler neighbourhoods $B(K\delta\log(1/\delta); P_0)$. When considering near-parametric rates (as in lemmas 9.1 and 9.3), logarithmic corrections of this kind do not influence the calculation, but they do play a role in non-parametric regression. It is possible that these logarithmic corrections to the rate can be omitted, the proof depending on a version of theorem 9.1 along the lines of theorem 2.4 of Ghosal et al. (2000) [101], in which the prior-mass condition is replaced by a more complicated, but less demanding bound on a ratio of prior masses. Note that the rate (9.15) approaches that given in (9.12) for large values of $\beta$, i.e. for regression classes with a high degree of differentiability.

Regarding classes with a high degree of differentiability, one might expect that suitably restricted classes of analytic regression functions would allow for convergence at the rate (9.15) in the limit $\beta \to \infty$, i.e. $1/\sqrt{n}$. However, in that case (9.9) and (9.13) are dominated by the contribution from the parameter $F \in D$, so the expected result would be the parametric rate of convergence given above, i.e. $1/\sqrt{n}$ with logarithmic correction of the order $(\log n)^{3/2}$.

### 9.3 Model entropy

One of the two primary conditions in theorems on non-parametric Bayesian rates of convergence (see, e.g. theorem 9.1), is an upper-bound on the covering numbers with respect to a metric on the model, in our case the Hellinger metric. In this section, we relate the Hellinger metric entropy of the model to entropy numbers of the three parametrizing spaces, i.e. $I$, $\mathcal{F}$ and $D$. Due to technical reasons (see subsection 9.3.3, which contains the proofs of all lemmas in this section), we can and shall express most results in terms of the $L_1(\mu)$-norm rather than the Hellinger metric, demonstrating in the (proof of) theorem 9.3 that this does not influence the entropy calculation.

#### 9.3.1 Nets in parametrizing spaces

We start the discussion by considering the $L_1(\mu)$-distance between densities in the model that differ only in one of the three parameters $(\sigma, f, F)$, the goal being the definition of an $\varepsilon$-net over $\mathcal{P}$ from $\varepsilon$-nets over the spaces $I$, $\mathcal{F}$ and $D$ separately.
With the following lemma, we indicate the possibility of generalizing the discussion that follows to situations in which less is known about the error distribution, by a bound on the $L_1(\mu)$-difference under variation of the parameter for the error distribution. For the next lemma only, we define $\{\psi_\sigma : \sigma \in \Sigma\}$ to be a family of Lebesgue densities of probability distributions on $\mathbb{R}^2$, parametrized by $\sigma$ in some (parametric or non-parametric) set $\Sigma$. The densities $p_{\sigma,f,F}$ are still given by a convolution c.f. (9.3) (because we maintain the assumption of independence of $Z$ and $(e,f)$).

**Lemma 9.5.** For every $f \in \mathcal{F}$ and $F \in D$,

$$\|p_{\sigma,f,F} - p_{\tau,f,F}\|_{1,\mu} \leq \|\psi_\sigma - \psi_\tau\|_{1,\mu},$$

for all $\sigma, \tau \in \Sigma$.

Specializing back to the situation of interest, we find the following lemma.

**Lemma 9.6.** In the case of equally distributed, independent normal errors $(e_1,e_2)$ with mean zero and equal but unknown variance in the interval $[\sigma, \overline{\sigma}]$:

$$\|\psi_\sigma - \psi_\tau\|_{1,\mu} \leq 4|\sigma - \tau|.$$

Similar inequalities can be derived for other parametric families of kernels, for instance the Laplace kernel. In the case of a non-parametric family of error distributions, it may be necessary to derive a (sharper) bound, based on the Hellinger distance between $p_{\sigma,f,F}$ and $p_{\tau,f,F}$. This generalized approach is not pursued here and the rest of this chapter relies on the assumption that the errors $(e_1,e_2)$ are as in the above lemma.

Next we consider the dependence of densities in the model on the regression function $f$.

**Lemma 9.7.** There exists a constant $K > 0$ such that for all $\sigma \in I$ and all $F \in D[-A,A]$:

$$\|p_{\sigma,f,F} - p_{\sigma,g,F}\|_{1,\mu} \leq K\|f - g\|_{1,F},$$

(9.16)

for all $f, g \in \mathcal{F}$.

The bound depends on the distribution $F$ for the underlying random variable $Z$ and proves the claim we made earlier, concerning identifiability of the regression function only up to null-sets of the distribution $F$. To derive a bound that is independent of $F$, we note that for all $F \in D$ and all $f, g \in C[-A,A]$:

$$\|f - g\|_{1,F} \leq \sup\{|f-g|(z) : z \in [-A,A]\} = \|f - g\|,$$

(9.17)

the right side being finite as a result of continuity of $f$ and $g$ and compactness of the interval $[-A,A]$. Note that we cannot simply equate the uniform norm $\|\cdot\|$ in (9.17) to the $L_\infty$-norm because the Lebesgue measure on $[-A,A]$ does not dominate all $F \in D$.

The bound $H^2(P,Q) \leq \|p - q\|_{1,\mu}$ suggests that metric entropy numbers for the Hellinger distance can safely be upper-bounded by those for the $L_1(\mu)$-norm. In
For all \( \sigma \in I, f, g \in \mathcal{F} \) and \( F \in D \):

\[
H(P_{\sigma,f,F}, P_{\sigma,g,F}) \leq \frac{1}{2\sigma} \left( \int_{[-A,A]} (f(z) - g(z))^2 \, dF(z) \right)^{1/2}.
\]

Although useful, the above bound depends on the particular values of \( \sigma, F \), which is undesirable in situations below. The lower bound for the interval \( I \) and the uniform bound on \( f - g \) serve to prove a bound on the Hellinger distance proportional to the uniform norm (as opposed to its square-root) of the difference between regression parameters.

Corollary 9.1. There exists a constant \( L > 0 \) such that for all \( \sigma \in I, f, g \in \mathcal{F} \) and \( F \in D \):

\[
H(P_{\sigma,f,F}, P_{\sigma,g,F}) \leq L\|f - g\|.
\]  \tag{9.18}

The above two lemmas and the fact that approximation in the uniform norm of subclasses of bounded continuous functions on closed intervals is well-understood, strongly suggests that the class of regression functions is to be endowed with the uniform norm to find nets. We do this in subsection 9.5.1 for the regression classes mentioned earlier.

To bound the contribution of the parameter \( F \) to the covering numbers of the model, we approximate \( F \) by a discrete distribution \( F' \) with a number of support points that is bounded by the approximation error in \( L_1(\mu) \). Note that the number of support points needed depends on a power of \( \log(1/\epsilon) \), so that a sharper bound in terms of the Hellinger distance is not necessary (see above).

Lemma 9.9. There exist constants \( C, C' > 0 \) such that for all \( (\sigma, f) \in I \times \mathcal{F} \) and \( F \in D \), there is a discrete \( F' \) on \([-A,A]\) with less than \( C(\log(1/\epsilon))^2 \) support points such that

\[
\|p_{\sigma,f,F} - p_{\sigma,f,F'}\|_{1,\mu} \leq C'\epsilon.
\]

We stress that the particular choice \( F' \) depends on the regression function \( f \). The above lemma implies that the set \( D_\epsilon \) of all discrete \( F \in D \) with less than \( C(\log(1/\epsilon))^2 \) support points parametrizes an \( \epsilon \)-net over \( \mathcal{P} \). For any fixed pair \( (\sigma, f) \in I \times \mathcal{F} \), the \( \epsilon \)-net parametrized by \( D_\epsilon \) is a \( 2\epsilon \)-net over the submodel \( \mathcal{P}_{\sigma,f} = \{ p_{\sigma,f,F} \in \mathcal{P} : F \in D \} \) so that

\[
N(\epsilon, \mathcal{P}_{\sigma,f}, \|\cdot\|_{1,\mu}) \leq N(2\epsilon, \{ p_{\sigma,f,F} \in \mathcal{P} : F \in D_\epsilon \}, \|\cdot\|_{1,\mu}).
\]
The direct nature of the above approximation (as opposed to the procedure for the parameters $\sigma$ and $f$, where we first bound by a norm on the parametrizing variable and then calculate the entropy in the parametrizing space) circumvents the notoriously difficult dependence of mixture densities on their mixing distribution, responsible for the (logarithmically) slow rate of convergence in deconvolution problems. Indeed, problems of this nature plague the method of Fan and Truong (1993) [88], which is based on a kernel-estimate for $F$ and leads to a Nadaraya-Watson-type of estimator for the regression function. Here we are only interested in covering the model $\mathcal{P}$, which allows us to by-pass the deconvolution problem by means of the above lemma.

### 9.3.2 Metric entropy of the errors-in-variables model

This subsection is devoted entirely to the following theorem, which uses the lemmas of the previous subsection to calculate the Hellinger entropy of the errors-in-variables model $\mathcal{P}$.

**Theorem 9.3.** Suppose that the regression family $\mathcal{F}$ is one of those specified in the beginning of section 9.2). Then there exist constants $L, L' > 0$ such that the Hellinger covering numbers of the model $\mathcal{P}$ satisfy:

$$\log N(\varepsilon, \mathcal{P}, H) \leq L' \left( \log \frac{1}{\varepsilon} \right)^3 + \log N(L\varepsilon, \mathcal{F}, \|\|),$$

(9.19)

for small enough $\varepsilon$.

**Proof.** If the class of regression functions $\mathcal{F}$ is a Lipschitz-class with exponent in $(0, 1)$, we set $\alpha$ equal to that exponent. In other cases we set $\alpha = 1$.

Let $\varepsilon > 0$ be given, fix some $\sigma \in I, f \in \mathcal{F}$. According to lemma (9.9) the collection $\mathcal{P}_{\sigma,f}$ of all $p_{\sigma,f,F'}$ where $F'$ is a discrete distribution in $D$ with at most $N_\varepsilon = \alpha^2 C (\log (1/\varepsilon))^2$ support points, forms an $\varepsilon^\alpha$-net over $\mathcal{P}_{\sigma,f}$ with respect to the $L_1(\mu)$-norm. Therefore any $\varepsilon^{\alpha}$-net $\mathcal{Q}_{\sigma,f}^\varepsilon$ over $\mathcal{P}_{\sigma,f}$ is a $2\varepsilon^{\alpha}$-net over $\mathcal{P}_{\sigma,f}$.

Let $\mathcal{S}_\varepsilon$ be a minimal $\varepsilon^\alpha$-net for the simplex with $\ell_1$-norm in $\mathbb{R}^{N_\varepsilon}$. As is shown by lemma A.4 in Ghosal and Van der Vaart (2001) [?], the order of $\mathcal{S}_\varepsilon$ does not exceed $(5/\varepsilon^{\alpha})^{N_\varepsilon}$. Next we define the grid $G_\varepsilon = \{0, \pm \varepsilon, \pm 2\varepsilon, \ldots\} \subset [-A, A]$ and $\mathcal{Q}_{\sigma,f}^\varepsilon$ as the collection of all distributions on $[-A, A]$ obtained by distributing the weights in a vector from $\mathcal{S}_\varepsilon$ over the points in $G_\varepsilon$. We project an arbitrary $p_{\sigma,f,F'}$ in $\mathcal{P}_{\sigma,f}$ onto $\mathcal{Q}_{\sigma,f}^\varepsilon$ in two steps: given that $F' = \sum_{i=1}^{N_\varepsilon} \lambda_i \delta_{z_i}$, for some set of $N_\varepsilon$ points $z_i \in [-A, A]$ and non-negative weights such that $\sum \lambda_i = 1$, we first project the vector $\lambda$ onto a vector in $\mathcal{S}_\varepsilon$ and second, shift the resulting masses to the closest point in $G_\varepsilon$. One easily sees that the first step leads to a new distribution $F''$ such that:

$$\|p_{\sigma,f,F'} - p_{\sigma,f,F''}\|_{1,\mu} \leq \varepsilon.$$
As for the second step, in which \( F'' = \sum_{i=1}^{N_k} \lambda_i' \delta_{\zeta_i} \) is ‘shifted’ to a new distribution \( F'' = \sum_{i=1}^{N_k} \lambda_i' \delta_{\zeta_i} \) such that \( |z_i - z_i'| \leq \epsilon \), we note that:

\[
|p_{\sigma,f,F''} - p_{\sigma,f,F'''}| \leq \sum_{i=1}^{N_k} \lambda_i' \left( |\varphi_\sigma(x - z_i) - \varphi_\sigma(x - z_i')| \varphi_\sigma(y - f(z_i)) + |\varphi_\sigma(y - f(z_i)) - \varphi_\sigma(y - f(z_i'))| \varphi_\sigma(x - z_i') \right),
\]

which implies that the \( L_1(\mu) \)-difference satisfies:

\[
\|p_{\sigma,f,F''} - p_{\sigma,f,F'''}\|_{1,\mu} \leq \sum_{i=1}^{N_k} \lambda_i' \left( \int |\varphi_\sigma(x - z_i) - \varphi_\sigma(x - z_i')| \, dx + \int |\varphi_\sigma(y - f(z_i)) - \varphi_\sigma(y - f(z_i'))| \, dy \right).
\]

By assumption, the family of regression functions satisfies (9.7), which is used to establish that there exists a constant \( K > 0 \) such that

\[
\|p_{\sigma,f,F''} - p_{\sigma,f,F'''}\|_{1,\mu} \leq Ke^{\alpha}.
\]

(for small enough \( \epsilon \)), along the same lines as the proof of lemma 9.7. Summarizing, we assert that for some constant \( K_3 > 0 \), \( \mathcal{D}_{\sigma,f}^\epsilon \) is a \( K_3^2\epsilon^{a_1} \)-net over \( \mathcal{P}_{\sigma,f} \). There exist an \( \epsilon^{a_2} \)-net \( I_\epsilon \) over \( I \) (with norm equal to absolute differences) and an \( \epsilon^{a_1/2} \)-net \( \mathcal{F}_\epsilon \) over \( \mathcal{F} \) in the uniform norm. (The order of \( \mathcal{F}_\epsilon \) is bounded in lemmas 9.13 and 9.15.)

By virtue of the triangle inequality and with the help of lemma 9.5 and corollary 9.1, we find that constants \( K_1, K_2 > 0 \) exist such that:

\[
H(p_{\sigma,F}, p_{\tau,g,F'}) \leq H(p_{\sigma,F}, p_{\tau,F}) + H(p_{\tau,F}, p_{\tau,g,F}) + H(p_{\tau,g,F}, p_{\tau,g,F'})
\]

\[
\leq \|p_{\sigma,F} - p_{\tau,F}\|_{1,\mu}^{1/2} + K_1 \|f - g\| + \|p_{\tau,g,F} - p_{\tau,g,F'}\|_{1,\mu}^{1/2} + K_2 \|\tau - \tau'\|^{1/2} + \|p_{\tau,g,F} - p_{\tau,g,F'}\|_{1,\mu}^{1/2},
\]

for all \( \sigma, \tau \in I_\epsilon, f \in \mathcal{F}, g \in \mathcal{F}_\epsilon \) and \( F, F' \in D \). For every fixed pair \( (\tau, g) \in I_\epsilon \times \mathcal{F}_\epsilon \), we define the \( K_3^2\epsilon^{a_1} \)-net \( \mathcal{D}_{\tau,g}^\epsilon \) like above and choose \( F' \) in the above display so that \( p_{\tau,g,F'} \) lies in \( \mathcal{D}_{\tau,g}^\epsilon \) and approximates \( p_{\tau,g,F} \) to within \( L_1(\mu) \)-distance proportional to \( \epsilon^{a_2} \). This shows that the set:

\[
\mathcal{D}_\epsilon = \bigcup \{ \mathcal{D}_{\tau,g}^\epsilon : \tau \in I_\epsilon, g \in \mathcal{F}_\epsilon \},
\]

forms a \( Ke^{a_1/2} \)-net over \( \mathcal{P} \) with respect to the Hellinger distance, where \( K = K_1 + K_2 + K_3 \). The order of this net can be calculated and forms an upper bound for the Hellinger covering number of the model.
\[
\log N(Ke^{\alpha/2}, \mathcal{P}, H) \leq \log N(e^{\alpha}, I, \| \cdot \|) + \log N(e^{\alpha/2}, \mathcal{F}, \| \cdot \|) + \log N(D_{\tau,q}^E),
\]
where \(N(D_{\tau,q}^E)\) denotes the uniform bound on the number of points in the nets \(D_{\tau,q}^E\), given by:
\[
\log N(D_{\tau,q}^E) = L'' \left( \log \frac{1}{\epsilon} \right)^3,
\]
for some constant \(L'' > 0\) as is easily checked from the above. Moreover, the covering numbers for the finite-dimensional, bounded space \(I\) satisfy, for some constant \(L'' > 0\):
\[
\log N(e^{\alpha}, I, \| \cdot \|) \leq L'' \log \frac{1}{\epsilon}.
\]
(Note that in the two displays above, any exponent for \(\epsilon\) (e.g. \(\alpha/2\)) is absorbed in the constants \(L'\) and \(L''\)). Note that for small enough \(\epsilon\), the contribution from the mixing parameter \(F\) dominates that of the parameter \(\sigma\). Eventually, we find the bound:
\[
\log N(\epsilon, \mathcal{P}, H) \leq L' \left( \log \frac{1}{\epsilon} \right)^3 + \log N(L\epsilon, \mathcal{F}, \| \cdot \|),
\]
for small enough \(\epsilon > 0\) and some \(L', L'' > 0\).

### 9.3.3 Proofs of several lemmas

**Proof.** Proof of lemma 9.5 Fix \(f \in \mathcal{F}\) and \(F \in D\), let \(\sigma, \tau \in \Sigma\) be given. Consider the \(L_1(\mu)\) difference:
\[
\|p_{\sigma,f,F} - p_{\tau,f,F}\|_1,\mu \leq \int_{\mathbb{R}} \int_{\mathbb{R}^2} \left| \psi_\sigma(x,z,y-f(z)) - \psi_\tau(x-z,y-f(z)) \right| d\mu(x,y) dF(z),
\]
by Fubini’s theorem. Translation invariance of the Lebesgue measure and the domain of integration \(\mathbb{R}^2\) make it possible to translate over \((z,f(z))\) to render the inner integral independent of \(z\) and integrate with respect to \(F\) with the following result:
\[
\|p_{\sigma,f,F} - p_{\tau,f,F}\|_1,\mu \leq \int_{\mathbb{R}^2} \left| \psi_\sigma(x,y) - \psi_\tau(x,y) \right| d\mu(x,y),
\]
thus leading to an upper bound that is independent of both \(f\) and \(F\).

**Proof.** Proof of lemma 9.6 The \(L_1(\mu)\)-difference of the densities \(\psi_\sigma\) and \(\psi_\tau\) equals the total-variational difference between the corresponding distributions \(\Psi_\sigma\) and \(\Psi_\tau\) and can be expressed in terms of the event \(\{\psi_\sigma > \psi_\tau\}\) as follows:
\[
\|\psi_\sigma - \psi_\tau\|_1,\mu = 2 \left( \Psi_\sigma(\psi_\sigma > \psi_\tau) - \Psi_\tau(\psi_\sigma > \psi_\tau) \right).
\]
In the case of normally and equally distributed, independent errors \((e_1,e_2)\) the kernel is \(\psi_\sigma(x,y) = \varphi_\sigma(x)\varphi_\sigma(y)\), with \(\sigma \in I\). Assuming that \(\sigma < \tau\), the event in question is a ball in \(\mathbb{R}^2\) of radius \(r_0\) centred at the origin (and its complement if \(\sigma > \tau\), where
where the regression class

where we have used the upper and lower bounds for the interval $I$.

**Proof.** Proof of lemma 9.7 Let $\sigma \in I$, $F \in D[-A,A]$ and $f, g \in \mathcal{F}$ be given. Since the $x$-dependence of the densities $p_{\sigma,f,F}$ and $p_{\sigma,g,F}$ is identical and can be integrated out, the $L_1(\mu)$-difference can be upper-bounded as follows:

$$\|p_{\sigma,f,F} - p_{\sigma,g,F}\|_{1,\mu} \leq \int_{\mathbb{R}} \int_{\mathbb{R}} \left| \varphi_\sigma(y-f(z)) - \varphi_\sigma(y-g(z)) \right| dydF(z).$$

Fix a $y \in \mathbb{R}$ and $z \in [-A,A]$. We note:

$$\left| \varphi_\sigma(y-f(z)) - \varphi_\sigma(y-g(z)) \right| \leq \int_{y-f(z)}^{y-g(z)} \varphi_\sigma'(u) du \leq \sup_{u \in J} |\varphi_\sigma'(u)| |f(z) - g(z)|,$$

where $J = [y-f(z) \vee g(z), y-f(z) \wedge g(z)]$. The uniform bound on the functions in the regression class $\mathcal{F}$ guarantees that $J \subset J' = [y-B,y+B]$. If $y \geq 2B$, then $y-B \geq \frac{1}{2}y \geq B > 0$, so if, in addition, $\frac{1}{2}y \geq \sigma$, we see that for all $u \in J'$, $u \geq \frac{1}{2}y \geq \sigma$, thus restricting $u$ to the region in which the derivative of the normal density decreases monotonously:

$$\varphi_\sigma'(u) \leq \varphi_\sigma'(\frac{1}{2}y).$$

Symmetry of the normal density allows us to draw the same conclusion if $y$ lies below $-2B$ and $-2\sigma$. Using the explicit form of the normal density and the constant $T = 2(B \vee \sigma)$, we derive the following upper bound on the supremum:

$$\sup \left\{ |\varphi_\sigma'(u)| : u \in J \right\} \leq Ks(y),$$

where the function $s$ is given by:

$$s(y) = \begin{cases} |y| \varphi_\sigma(y), & \text{if } |y| \geq T, \\ ||\varphi_\sigma'||_\infty, & \text{if } |y| < T. \end{cases}$$

Note that $s$ does not depend on the values of the parameters. Therefore:

$$\|p_{\sigma,f,F} - p_{\sigma,g,F}\|_{1,\mu} \leq \int_{\mathbb{R}} \int_{\mathbb{R}} Ks(y)|f(z) - g(z)| dydF(z).$$

Since the integral over $s(y)$ is finite, the asserted bound follows.

**Proof.** (Proof of lemma 9.8) Consider a binary experiment $E_1 = (\mathbb{R}^3, \mathcal{G}(3), \{P, Q\})$, giving two possible distributions $P, Q$ for the triplet $(X, Y, Z)$ that describes the errors-in-variables model (c.f. (9.2)). The map $T$ that projects by $T(X, Y, Z) = (X, Y)$ leads to another binary experiment $E_2 = (\mathbb{R}^2, \mathcal{G}(2), \{P^T, Q^T\})$ which is less infor-
mative than $E_1$. (The phrase “less informative” is defined in the sense of Le Cam, i.e., for every test function $\phi_2$ in $E_2$, there exists a test function $\phi_1$ in $E_1$ such that $P\phi_1 \leq P^T\phi_2$ and $Q\phi_1 \geq Q^T\phi_2$ (see, for instance, Strasser (1985) [208], definition 15.1.) This property follows from the fact that $\sigma(X,Y) \subset \mathcal{B}^{[2]}$ is such that $T^{-1} (\sigma(X,Y)) \subset \sigma(X,Y,Z) \subset \mathcal{B}^{[3]}$, which makes it possible to identify every test function in $E_2$ with a test function in $E_1$, while there may exist test functions on $\mathbb{R}^3$ that are not measurable with respect to $T^{-1}(\sigma(X,Y))$. Corollary 17.3 in Strasser (1985) [208] asserts that the Hellinger distance decreases when we make the transition from a binary experiment to a less informative binary experiment, so we see that:

$$H(P^T,Q^T) \leq H(P,Q).$$

(9.20)

In the case at hand, we choose $P^T = P_{\sigma,f,F}$ and $Q^T = P_{\sigma,g,F}$. From the definition of the errors-in-variables model (9.2), we obtain the conditional laws:

$$\mathcal{L}_P(X,Y \mid Z) = N(Z,\sigma^2) \times N(f(Z),\sigma^2),$$

$$\mathcal{L}_Q(X,Y \mid Z) = N(Z,\sigma^2) \times N(g(Z),\sigma^2),$$

and, of course, $\mathcal{L}_P(Z) = \mathcal{L}_Q(Z) = F$. It follows that:

$$H^2(P,Q) = \int_{\mathbb{R}^1} (dP^{1/2} - dQ^{1/2})^2$$

$$= \int_{\mathbb{R}^1} \varphi_\sigma(x-z) \left( \varphi_\sigma(y-f(z))^{1/2} - \varphi_\sigma(y-g(z))^{1/2} \right)^2 dF(z) dx dy$$

$$= \int_{[-A,A]} H^2(N(f(z),\sigma^2),N(g(z),\sigma^2)) dF(z),$$

by Fubini’s theorem. A straightforward calculation shows that:

$$H^2(N(f(z),\sigma^2),N(g(z),\sigma^2)) = 2 \left( 1 - e^{-\frac{1}{2}(f(z)-g(z))^2/(2\sigma^2)} \right) \leq \frac{1}{4\sigma^2} (f(z)-g(z))^2,$$

where we use that $1 - e^{-x} \leq x$ for all $x \geq 0$. Upon combination of the above two displays and (9.20), we obtain:

$$H^2(P_{\sigma,f,F},P_{\sigma,g,F}) \leq \frac{1}{4\sigma^2} \int_{[-A,A]} (f(z)-g(z))^2 dF(z),$$

which proves the assertion.

**Proof.** Proof of lemma 9.9 Let $\varepsilon > 0$, $\sigma \in I$, $f \in \mathcal{F}$ be given, fix $M \geq 2A \vee 2B$ and $k \geq 1$. A Taylor-expansion up to order $k-1$ of the exponential in the normal density demonstrates that:
Thus choosing \( \psi \) functions \(|\psi_1| > 1\) such that:

\[
\frac{1}{\sigma \sqrt{2\pi}} \sum_{j=0}^{k-1} \frac{1}{j!} \left( \frac{1}{2} \right)^{j} \left( \frac{x-z}{\sigma} \right)^{2j} \leq \frac{1}{\sigma \sqrt{2\pi}} k! \left( \frac{1}{2} \right)^{k} \left( \frac{x-z}{\sigma} \right)^{2k},
\]

where we have used that \( k! \geq k^k e^{-k} \). Similarly, we obtain:

\[
\frac{1}{\sigma \sqrt{2\pi}} \sum_{j=0}^{k-1} \frac{1}{j!} \left( \frac{1}{2} \right)^{j} \left( \frac{y-f(z)}{\sigma} \right)^{2j} \leq \frac{1}{\sigma \sqrt{2\pi}} \left( \frac{e}{2k} \right)^{k} \left( \frac{y-f(z)}{\sigma} \right)^{2k}.
\]

Considering \(|x|, |y| \leq M\) and using that \( \sigma \geq \sigma > 0 \), we see that there exists a constant \( C_1 > 0 \) (independent of \( \sigma \) and \( f \)) such that both residuals of the last two displays are bounded above by \( (C_1 M^2) k^k \). So for all \( x, y \) like above,

\[
\left| p_{\sigma,f,F} - p_{\sigma,f,F'} \right|(x,y)
\leq \frac{1}{2 \pi \sigma^2} \left| \int \sum_{j=0}^{k-1} \frac{1}{j!} \left( \frac{1}{2} \right)^{j} \left( \frac{x-z}{\sigma} \right)^{2j} \left( \frac{y-f(z)}{\sigma} \right)^{2j} d(F-F')(z) \right| (9.21)
\]

\[
+ \left( \frac{C_1 M^2}{k} \right)^k + \left( \frac{C_1 M^2}{k} \right)^k.
\]

Lemma A.1 in Ghosal and Van der Vaart (2001) [1] asserts that there exists a discrete distribution \( F' \) on \([-A, A]\) with at most \((k^2+1)\) support points such that for all functions \( \psi_{f,j}(z) = z^{2j} f(j)(z) \) the \( F \) and \( F' \)-expectations coincide, i.e.:

\[
\int_{[-A, A]} \psi_{f,j} dF = \int_{[-A, A]} \psi_{f,j} dF'.
\]

Thus choosing \( F' \), the first term in (9.21) vanishes and we see that (for large enough \( k \)):

\[
\sup_{|x| \leq M} \left| p_{\sigma,f,F} - p_{\sigma,f,F'} \right|(x,y) \leq 5 \left( \frac{C_1 M^2}{k} \right)^k (9.22)
\]

For points \((x, y)\) outside \([-M, M] \times [-M, M]\), we note that there exists a constant \( C_2 > 0 \) such that for all \(|x| \geq 2A, |y| \geq 2B\):

\[
\varphi_\sigma(x-z) \leq \varphi_\sigma \left( \frac{1}{2} \right) \leq C_2 \varphi_\sigma \left( \frac{1}{2} \right),
\]

\[
\varphi_\sigma(y-f(z)) \leq \varphi_\sigma \left( \frac{1}{2} \right) \leq C_2 \varphi_\sigma \left( \frac{1}{2} \right),
\]

\((C_2 = \| \varphi_\sigma \|_\infty / \| \varphi_\sigma \|_\infty \) will do). Since \( M \geq 2A \vee 2B \), there exists a constants \( C_3, C_4 > 0 \) such that:

\[
\left| p_{\sigma,f,F} - p_{\sigma,f,F'} \right|(x,y)
\leq \frac{1}{\sigma \sqrt{2\pi}} \sum_{j=0}^{k-1} \frac{1}{j!} \left( \frac{1}{2} \right)^{j} \left( \frac{x-z}{\sigma} \right)^{2j} \left( \frac{y-f(z)}{\sigma} \right)^{2j} d(F-F')(z)
\]
where we have used Fubini’s theorem and translation invariance of Lebesgue measure in the second step and the fact that \( \psi_\sigma'(x) = -\left(x/\sigma^2\right)\psi_\sigma(x) \) in the last. Now, let \( \epsilon > 0 \) be given. We decompose the domain of integration for the \( L_1(\mu) \)-difference between \( p_{\sigma,f,F} \) and \( p_{\sigma,f,F'} \) into the region where \( |x| \vee |y| \leq M \) and its complement. Using the uniform bound (9.22) on the region bounded by \( M \) and (9.23) for the tails, we find that there is a constant \( D_1 \) such that:

\[
\|p_{\sigma,f,F} - p_{\sigma,f,F'}\|_{1,\mu} \leq D_1 \left( M^2 \left( \frac{C_1 M^2}{k} \right)^k + \epsilon^{-C_4 M^2} \right). \tag{9.24}
\]

In order to bound the r.h.s. by \( \epsilon \) we fix \( M \) in terms of \( \epsilon \):

\[
M = \sqrt{\log \frac{1}{\epsilon}},
\]

and note that the lower bound \( M \geq 2A \vee 2B \) is satisfied for small enough \( \epsilon \). Upon substitution, the first term in (9.24) leads to \( (D_1/C_4)D_2^k \epsilon^{(k+1)\log \log \frac{1}{\epsilon}} e^{-k \log k} \) (where \( D_2 = C_1/C_4 \)), so that the choice:

\[
k \geq D_3 \log \frac{1}{\epsilon},
\]

(for some large \( D_3 > D_2 \)) suffices to upper bound the \( L_1(\mu) \)-difference appropriately. The smallest integer \( k \) above the indicated bound serves as the minimal number of support points needed.

Note that the \( f \)-dependence of the functions \( \psi_{f,ij} \) carries over to the choice for \( F' \), which is therefore \( f \)-dependent as well.

### 9.4 Model prior

Assume that the model is well-specified and denote by \( P_0 \in \mathcal{P} \) (corresponding to some, not necessarily unique, \( \sigma_0 \in \mathcal{I} \), \( f_0 \in \mathcal{F} \) and \( F_0 \in \mathcal{D} \) the true distribution underlying the i.i.d. sample. We define a prior \( \Pi \) on \( \mathcal{P} \) by defining priors on the parameter spaces \( \mathcal{I} \), \( \mathcal{F} \) and \( \mathcal{D} \) and taking \( \Pi \) equal to the probability measure induced by the map \( (\sigma,f,F) \mapsto P_{\sigma,f,F} \) on \( \mathcal{I} \times \mathcal{F} \times \mathcal{D} \) with product-measure to \( \mathcal{P} \). The prior on \( \mathcal{I} \) is denoted \( \Pi_I \) and is assumed to have a density \( \pi_I \), continuous and strictly
positive at $\sigma_0$. The prior $\Pi_\mathcal{F}$ on $\mathcal{F}$ is specified differently for each of the classes defined in the beginning of section 9.2, but all have as their domain the Borel $\sigma$-algebra generated by the norm topology on $C[-A,A]$. The definition of these priors is postponed to subsection 9.5.2. The prior $\Pi_D$ on $D$ is based on a Dirichlet process with base measure $\alpha$ which has a continuous and strictly positive density on all of $[−A,A]$. The domain of $\Pi_D$ is the Borel $\sigma$-algebra generated by the topology of weak convergence.

The fact that these priors are defined on the product of the parameter spaces rather than the errors-in-variables model $\mathcal{P}$ itself, necessitates a lemma asserting appropriate measurability. So before we discuss the properties of priors, we show that the map $\hat{p}$ that takes parameters $(\sigma, f, F)$ into densities $p_{\sigma, f, F}$ (c.f. (9.3)) is measurable.

**Lemma 9.10.** Endow $I$ and $\mathcal{F}$ with their norm topology and $D$ with the topology of weak convergence. Then the map $\hat{p} : I \times \mathcal{F} \times D \to L_1(\mu)$ is continuous in the product topology.

**Proof.** The space $D$ with the topology of weak convergence is metric, so the product topology on $I \times \mathcal{F} \times D$ is a metric topology as well. Let $(\sigma_n, f_n, F_n)$ be a sequence, converging to some point $(\sigma, f, F)$ in $I \times \mathcal{F} \times D$ as $n \to \infty$. As a result of the triangle inequality and lemmas 9.5–9.7, the $L_1(\mu)$-distance satisfies:

$$\|p_{\sigma_n, f_n, F_n} - p_{\sigma, f, F}\|_{1,\mu} \leq K_1|\sigma_n - \sigma| + K_2\|f_n - f\| + \|p_{\sigma, f, F_n} - p_{\sigma, f, F}\|_{1,\mu}, \quad (9.25)$$

for some constants $K_1, K_2 > 0$. Since $F_n$ converges to $F$ weakly, the continuity of the regression function $f$, combined with the continuity and boundedness of the Gaussian kernel and the portmanteau lemma guarantee that

$$\int_{[-A,A]} \varphi_\sigma(x-z) \varphi_\sigma(y-f(z)) dF_n(z) \to \int_{[-A,A]} \varphi_\sigma(x-z) \varphi_\sigma(y-f(z)) dF(z),$$

as $n \to \infty$ for all $(x, y) \in \mathbb{R}^2$. Using the $(\mu$-integrable) upper-envelope for the model $\mathcal{P}$ and dominated convergence, we see that

$$\|p_{\sigma, f, F_n} - p_{\sigma, f, F}\|_{1,\mu} \to 0,$$

and hence the r.h.s. of (9.25) goes to zero. We conclude that $\hat{p}$ is continuous in the product topology.

Note that the $L_1(\mu)$- and Hellinger topologies on the model $\mathcal{P}$ are equivalent, so that the above lemma implies continuity of $\hat{p}$ in the Hellinger topology. Hence $\hat{p}^{-1}$ is a well-defined map between the Borel $\sigma$-algebras of the model with the Hellinger topology and the product $I \times \mathcal{F} \times D$.

The following lemma establishes that the prior-mass condition (9.4) can be analysed for the regression class and the parameter space for $(\sigma, F)$ separately. Lower bounds for the prior mass in appropriate neighbourhoods of the point $(\sigma_0, F_0)$ are incorporated immediately.
Theorem 9.4. Suppose that the regression family \( \mathcal{F} \) is one of those specified in the beginning of section 9.2. Assume that the prior \( \Pi \) on \( \mathcal{P} \) is of the product form indicated above. Then there exist constants \( K, c, C > 0 \) such that:

\[
\Pi(B(\sqrt{K} \delta \log(1/\delta); P_0)) \geq C \exp\left(-c(\log(1/\delta))^3\right) \Pi_{\mathcal{F}}(f \in \mathcal{F} : \|f - f_0\| \leq \delta),
\]

for small enough \( \delta \).

Proof. If the class of regression functions \( \mathcal{F} \) is a Lipschitz-class with exponent in \((0,1)\), we set \( \alpha \) equal to that exponent. In other cases we set \( \alpha = 1 \).

Let \( \varepsilon > 0 \) be given. By lemma 9.9 there exists a discrete \( F_0' \) in \( D \) with at most \( N_\varepsilon = C(\log(1/\varepsilon))^2 \) support points \( z_1, \ldots, z_{N_\varepsilon} \) of the form \( F_0' = \sum_{i=1}^{N_\varepsilon} p_i \delta_{z_i} \) with \( \sum_{i=1}^{N_\varepsilon} p_i = 1 \), such that:

\[
\|p_{\sigma_{0},f_0,F_0} - p_{\sigma_{0},f_0,F_0'}\|_{1,\mu} \leq C' \varepsilon^\alpha,
\]

for some constant \( C' > 0 \). Although the assertion of lemma 9.9 is stronger, we include the power of \( \varepsilon \) because we assume (without loss of generality) that the set of support points for \( F_0' \) is \( 2\varepsilon \)-separated. If this is not the case, take a maximal \( 2\varepsilon \)-separated subset and shift the masses of other support points of \( F_0' \) to points in the chosen subset within distance \( 2\varepsilon \), to obtain a new discrete distribution \( F_0'' \). Arguing as in the proof of theorem 9.3, we see that the corresponding change in \( L_1(\mu) \)-distance between \( p_{\sigma_{0},f_0,F_0'} \) and \( p_{\sigma_{0},f_0,F_0''} \) is upper-bounded by a multiple of \( \varepsilon^\alpha \), since the family of regression functions satisfies (9.7) by assumption. The distribution function \( F_0'' \) so obtained may then replace \( F_0' \). By lemma 9.11, there exists a constant \( K_3 > 0 \) such that for all \( F \in D \):

\[
\|p_{\sigma_{0},f_0,F} - p_{\sigma_{0},f_0,F_0'}\|_{1,\mu} \leq K_3 \left( \varepsilon^\alpha + \sum_{i=1}^{N_\varepsilon} |F[z_i] - \varepsilon, z_i + \varepsilon| - p_i \right).
\]

Let \((\sigma, f, F)\) be a point in the parameter space of the model. The Hellinger distance between \( p_{\sigma,f,F} \) and \( p_{\sigma_{0},f_0,F_0} \) is upper-bounded as follows (for constants \( K_1, K_2 > 0 \):

\[
H(P_{\sigma,f,F}, P_{\sigma_{0},f_0,F_0}) \leq H(P_{\sigma_{0},f,F}, P_{\sigma_{0},f_0,F}) + H(P_{\sigma_{0},f,F}, P_{\sigma_0,f_0,F_0}) + H(P_{\sigma_0,f_0,F}, P_{\sigma_0,f_0,F_0})
\]

\[
\leq \left\| p_{\sigma,f,F} - p_{\sigma_0,f,F} \right\|_{1,\mu}^{1/2} + H(P_{\sigma_0,f,F}, P_{\sigma_0,f_0,F}) + \left\| p_{\sigma_0,f_0,F} - p_{\sigma_0,f_0,F_0} \right\|_{1,\mu}^{1/2}
\]

\[
\leq K_1 |\sigma - \sigma_0|^{1/2} + K_2 \|f - f_0\| + \left( \left\| p_{\sigma_0,f_0,F} - p_{\sigma_0,f_0,F_0} \right\|_{1,\mu} + \left\| p_{\sigma_0,f_0,F_0} - p_{\sigma_0,f_0,F_0} \right\|_{1,\mu} \right)^{1/2},
\]

where we have used lemmas 9.5, 9.6 and corollary 9.1. Moreover, we see that there exists a constant \( K_4 > 0 \) such that for small enough \( \eta > 0 \) and \( P \in \mathcal{P} \) such that \( H(P, P_0) \leq \eta \):
\(-P_0 \log \frac{p}{p_0} \vee P_0 \left( \log \frac{p}{p_0} \right)^2 \leq K_\delta^2 \eta^2 \left( \log \frac{1}{\eta} \right)^2\),

as a result of lemma 9.12. Combining the last two displays and using definition (6.23), we find that, for some constants \(K_\delta, K_\alpha > 0\), the following inclusions hold:

\[
\left\{ (\sigma, f, F) \in I \times \mathcal{F} \times D : \right. \\
|\sigma - \sigma_0|^{1/2} \leq \varepsilon^a, \quad \|f - f_0\| \leq \varepsilon^{a/2}, \quad \sum_{j=1}^{N_\epsilon} |F[z_j - \varepsilon, z_j + \varepsilon] - p_j| \leq \varepsilon^a \}
\]

\[
\subset \left\{ (\sigma, f, F) \in I \times \mathcal{F} \times D : H(P_{\sigma, f, F}, P_0) \leq K_\delta \varepsilon^{a/2} \right\}
\]

\[
\left\{ P \in \mathcal{F} : H(P, P_0) \leq K_\delta \varepsilon^{a/2} \right\} \subset B(K_\delta \varepsilon^{a/2} \log(1/\varepsilon); P_0),
\]

(9.27)

for small enough \(\varepsilon\) and with the notation \(p_0\) for the density of \(P_0\) (\(p_0 = p_{\sigma_0, f_0, P_0}\)).

Using the fact that the prior measure of the rectangle set on the l.h.s. of the first inclusion above factorizes, we find that:

\[
\Pi \left( B(K_\delta \varepsilon^{a/2} \log(1/\varepsilon); P_0) \right) \geq \Pi_I \left( \sigma \in I : |\sigma - \sigma_0|^{1/2} \leq \varepsilon^a \right) \Pi_{\mathcal{F}} \left( f \in \mathcal{F} : \|f - f_0\| \leq \varepsilon^{a/2} \right) 
\times \Pi_D \left( F \in D : \sum_{j=1}^{N_\epsilon} |F[z_j - \varepsilon, z_j + \varepsilon] - p_j| \leq \varepsilon^a \right).
\]

Note that \(\varepsilon^a \geq \varepsilon\) for small enough \(\varepsilon\), so that

\[
\Pi_D \left( \sum_{j=1}^{N_\epsilon} |F[z_j - \varepsilon, z_j + \varepsilon] - p_j| \leq \varepsilon^a \right) \geq \Pi_D \left( \sum_{j=1}^{N_\epsilon} |F[z_j - \varepsilon, z_j + \varepsilon] - p_j| \leq \varepsilon \right).
\]

According to lemma 6.1 in Ghosal et al. (2000) [101] (also given as lemma A.2 in Ghosal and Van der Vaart (2001) [2]), there are constants \(C', c' > 0\) such that

\[
\Pi_D \left( \sum_{j=1}^{N_\epsilon} |F[z_j - \varepsilon, z_j + \varepsilon] - p_j| \leq \varepsilon \right) \geq C \exp(-c' N_\epsilon \log(1/\varepsilon)) \geq C' \exp(-c' C (\log(1/\varepsilon))^3).
\]

Furthermore, continuity and strict positivity of the density of the prior \(\Pi_I\) imply that (see the proof of lemma 9.17):

\[
\Pi_I (\sigma \in I : |\sigma - \sigma_0| \leq \varepsilon^a) \geq \pi_1 \varepsilon^a = \pi_1 \exp(-\alpha \log(1/\varepsilon)),
\]

for some constant \(\pi_1 > 0\). Note that the exponent on the r.h.s. falls above all multiples of \(-\log(1/\varepsilon)^3\) for small enough \(\varepsilon\). Substitution of \(\delta = \varepsilon^{a/2}\) leads to the conclusion that there exist constants \(K, c, C > 0\) such that:

\[
\Pi \left( B(K \delta \log(1/\delta); P_0) \right) \geq C \exp\left(-c (\log(1/\delta))^3\right) \Pi_{\mathcal{F}} \left( f \in \mathcal{F} : \|f - f_0\| \leq \delta \right),
\]
for small enough $\delta$.

If the model is not identifiable in the parameter space $I \times \mathcal{F} \times D$, the above conditions are more stringent than necessary. The point $(\sigma_0, f_0, F_0)$ may not be the only one that is mapped to $P_0$, so the first inclusion in (9.27) may discount parts of the parameter space that also contribute to the Kullback-Leibler neighbourhoods $B(\varepsilon; P_0)$. However, the methods we use to lower-bound the prior mass rely on uniformity in the sense that neighbourhoods of every point in the parameter space receive a certain minimal fraction of the total prior mass. Therefore, identifiability issues do not affect the argument.

### 9.4.1 Lemmas

In the following lemma, it is assumed that the regression class $\mathcal{F}$ is one of those specified in the beginning of section 9.2. If the class of regression functions is a Lipschitz-class with exponent in $(0,1)$, we set $\alpha$ equal to that exponent. In other cases we set $\alpha = 1$.

**Lemma 9.11.** Let $\varepsilon > 0$ be given and let $F' = \sum_{i=1}^{N} p_i \delta_{z_i}$ be a convex combination of point-masses, where the set $\{z_i : i = 1,\ldots,N\}$ is $2\varepsilon$-separated. Then there exists a constant $K > 0$ such that for all $\sigma \in I$, $f \in \mathcal{F}$ and all $F \in D$:

$$
\|p_{\sigma,f,F} - p_{\sigma,f,F'}\|_{1,\mu} \leq K \left( \varepsilon^\alpha + \sum_{i=1}^{N} |F[z_i - \varepsilon, z_i + \varepsilon] - p_i| \right),
$$

for small enough $\delta$.

**Proof.** Let $F$ be given. We partition the real line by $\mathbb{R} = \cup_i A_i \cup B$, with $B = \cap_i B_i$, where

$$
A_i = \{ z : |z - z_i| \leq \varepsilon \}, \quad B_i = \{ z : |z - z_i| > \varepsilon \},
$$

and decompose the absolute difference between $p_{\sigma,f,F}$ and $p_{\sigma,f,F'}$ accordingly:

$$
|p_{\sigma,f,F} - p_{\sigma,f,F'}|(x,y) = \left| \int_{\mathbb{R}} \phi_\sigma(x-z) \phi_\sigma(y-f(z)) d(F-F')(z) \right|
$$

$$
= \left| \sum_{i=1}^{N} \int_{A_i} \phi_\sigma(x-z) \phi_\sigma(y-f(z)) d(F-F')(z) + \int_{B} \phi_\sigma(x-z) \phi_\sigma(y-f(z)) dF(z) \right|,
$$

for all $(x,y) \in \mathbb{R}^2$. Integrating this expression over $\mathbb{R}^2$, we find that the $L_1(\mu)$-difference is bounded as follows:

$$
\|p_{\sigma,f,F} - p_{\sigma,f,F'}\|_{1,\mu} \leq \sum_{i=1}^{N} |F[z_i - \varepsilon, z_i + \varepsilon] - p_i| + \int_{\bigcap_i B_i} F d\mu(x,y) dF(z),
$$

$$
+ \sum_{i=1}^{N} \int_{A_i} \int_{\mathbb{R}^2} |\phi_\sigma(x-z) \phi_\sigma(y-f(z)) - \phi_\sigma(x-z_i) \phi_\sigma(y-f(z_i))| d\mu(x,y) dF(z),
$$

for small enough $\delta$. 


We shall prove that for a suitable choice of $\delta$:

\[
\Phi(x-z)\Phi(y-f(z)) - \Phi(x-z_i)\Phi(y-f(z_i)) \leq |\Phi(x-z) - \Phi(x-z_i)|\Phi(y-f(z)) + |\Phi(y-f(z)) - \Phi(y-f(z_i))|\Phi(x-z_i),
\]

and argue as in the proof of lemma 9.7, to see that the integrand is bounded by a multiple of $|z-z_i|^\alpha$ for small enough $\varepsilon$. Noting that the intervals $[z_i - \varepsilon, z_i + \varepsilon]$ are disjoint due to $2\varepsilon$-separation of the set $\{z_i : i = 1, \ldots, N\}$, we see that there exists a constant $L' > 0$ such that

\[
\|p_{f,F} - p_{f,F'}\|_{1,\mu} \leq L' \varepsilon^\alpha + \sum_{i=1}^{N} |F[z_i - \varepsilon, z_i + \varepsilon] - p_i| + F \left( \bigcap_{i=1}^{N} B_i \right).
\]

Furthermore, by De Morgan’s law and the disjointness of the intervals $[z_i - \varepsilon, z_i + \varepsilon]$:

\[
F \left( \bigcap_{i=1}^{N} \{z : |z - z_i| > \varepsilon\} \right) = 1 - F \left( \bigcup_{i=1}^{N} \{z : |z - z_i| \leq \varepsilon\} \right) = \sum_{i=1}^{N} p_i - \sum_{i=1}^{N} F[z_i - \varepsilon, z_i + \varepsilon] \leq \sum_{i=1}^{N} |F[z_i - \varepsilon, z_i + \varepsilon] - p_i|,
\]

which proves the assertion.

**Lemma 9.12.** Let $P, Q \in \mathcal{P}$ be given. There exists a constant $K > 0$ such that for small enough $H(P, Q)$:

\[
\int p \log \frac{p}{q} \, d\mu \leq K^2 H^2(P, Q) \left( \log \frac{1}{H(P, Q)} \right)^2,
\]

\[
\int p \left( \log \frac{p}{q} \right)^2 \, d\mu \leq K^2 H^2(P, Q) \left( \log \frac{1}{H(P, Q)} \right)^2.
\]

(9.28)

The constant $K$ does not depend on $P, Q$.

**Proof.** Fix $\delta \in (0, 1]$ and consider the integral:

\[
M^2_\delta = \int p \left( \frac{p}{q} \right)^\delta \, d\mu.
\]

We shall prove that for a suitable choice of $\delta$, $M^2_\delta < \infty$. Since all densities involved are bounded away from zero and infinity on compacta, we consider only the domain $O = \mathbb{R}^2 \setminus [-C, C] \times [-C, C]$, for some large constant $C \geq A \vee B$. Note that:

\[
\int_O p \left( \frac{p}{q} \right)^\delta \, d\mu \leq \int_O U \left( \frac{U}{L} \right)^\delta \, d\mu,
\]
where \((L, U)\) forms an envelope for the model. This envelope follows from the fact that the regression densities (9.3) fall in the class of mixture densities obtained by mixing the normal kernel \(\phi_{\sigma}(x)\phi_{\sigma}(y)\) on \(\mathbb{R}^2\) by means of a two-dimensional distribution that places all its mass in the rectangle \([-A, A] \times [-B, B]\). There exists a lower bound for this envelope which factorizes into \(x\)- and \(y\)-envelopes \((L_X, U_X)\) and \((L_Y, U_Y)\) that are constant on sets that include \([-A, A]\) and \([-B, B]\) respectively and have Gaussian tails. The domain \(O\) can therefore be partitioned into four subdomains in which either \(x\) or \(y\) is bounded and four subdomains in which both coordinates are unbounded. Reflection-symmetries of the envelope functions suffice to demonstrate that integrals of \(U(U/L)^{\delta}\) can be expressed as products of trivially finite factors and integrals of the form:

\[
\int_{L}^{\infty} U_X(x) \left( \frac{U_X}{L_X} \right)^{\delta} (x) \, d\mu(x), \quad \int_{L}^{\infty} U_Y(y) \left( \frac{U_Y}{L_Y} \right)^{\delta} (y) \, d\mu(y).
\]

For large enough \(C\), the envelope functions \(L_X(x)\) and \(U_X(x)\) are equal to multiples of \(\phi_{\sigma}(x+A)\) and \(\phi_{\sigma}(x-A)\) on the domain \((C, \infty)\) and hence, for some constants \(c, K > 0\):

\[
\int_{L}^{\infty} U_X(x) \left( \frac{U_X}{L_X} \right)^{\delta} (x) \, d\mu(x) \leq K \int_{L}^{\infty} e^{c\delta^2} \phi_{\sigma}(x-A) \, dx,
\]

which is finite for small enough \(\delta > 0\). Similarly, one can prove finiteness of the integrals over \(y\). This proves that the condition for theorem 5 in Wong and Shen (1995) \([2]\) is satisfied. Note that the choice for \(\delta\) is independent of \(p, q\). Furthermore, the value of \(M_{\delta}\) can be upper-bounded independent of \(p, q\), as is apparent from the above. Hence, for small enough \(\eta > 0\), (9.28) holds.

### 9.5 Regression classes

Theorems 9.3 and 9.4 demonstrate that both the entropy and prior-mass conditions in theorem 9.1 can be decomposed in a term that pertains to the regression function \(f\) and a term pertaining to the parameters \((\sigma, F)\). This makes it possible to consider entropy and prior-mass restricted to the regression class separately.

In the first subsection, we state a bound on the metric entropy of the classes \(C_{B,M}[-A,A]\) due to Kolmogorov, who derived it shortly after his introduction of the concept of covering numbers. This bound is used in the second subsection to demonstrate that so-called net priors can be used for non-parametric regression classes in this situation. Also discussed is an alternative approach, that uses (adapted versions of) Jackson’s approximation theorem. Up to a logarithmic correction, the second approach reproduces Kolmogorov’s bound for the metric entropy, but upon application in the form of so-called sieve priors, the resulting lower bounds for the prior mass in neighbourhoods of the true regression function are sub-optimal in a more grave manner. Nevertheless, we indulge in an explanation of the second approach, because it provides a good example of the methods and subtleties of Bayesian procedures in
non-parametric problems. We also give the necessary bounds on the entropy and prior mass of parametric regression classes.

### 9.5.1 Covering numbers of regression classes

The usefulness of bounds (9.17) and (9.18) indicates that the class of regression functions parametrizing the model is best chosen within the (Banach-)space \( C[-A,A] \) of continuous functions on the closed interval \([-A,A]\) with the uniform norm \( \| \cdot \| \). According to the Weierstrass approximation, polynomials are dense in \( C[-A,A] \); bounded families of polynomials can therefore be used to approximate regression families \( \mathcal{F} \) as characterised in point (c) at the beginning of subsection 9.1.1. The Ascoli-Arzelà theorem asserts that if, in addition, \( \mathcal{F} \) is equicontinuous, it is relatively compact. Hence bounded, equicontinuous families \( \mathcal{F} \) are totally bounded in the norm-topology, rendering covering numbers finite,

\[
N(\varepsilon, \mathcal{F}, \| \cdot \|) < \infty,
\]

for all \( \varepsilon > 0 \). However, since we are interested in rates of convergence, finiteness of covering numbers is not enough and a more detailed analysis of the behaviour of \( N(\varepsilon, \mathcal{F}, \| \cdot \|) \) for small \( \varepsilon \) is needed. We reproduce here a result due to Kolmogorov and Tikhomirov (1961) [147] (in a version as presented in Van der Vaart and Wellner (1996) [218]), that gives the required bound:

**Lemma 9.13.** Let \( \beta > 0, M > 0 \) be given. There exists a constant \( K \) depending only on \( \beta \) and \( A \), such that:

\[
\log N(\varepsilon, C_{\beta,M}[-A,A], \| \cdot \|) \leq K \left( \frac{1}{\varepsilon} \right)^{1/\beta},
\]

for all \( \varepsilon > 0 \).

The proof of this lemma is a special version of the proof of theorem 2.7.1 in [218], which consists of a fairly technical approximation by polynomials. To improve our understanding of the above result, we briefly digress on an approach that is based on Jackson’s approximation theorem.

Fix an \( n \geq 1 \); Jackson’s approximation theorem (see Jackson (1930) [?]) says that if \( f \in \text{Lip}_M(\alpha) \), there exists an \( n \)-th order polynomial \( p_n \) such that:

\[
\| f - p_n \| \leq \frac{K}{n^\alpha},
\]

where \( K > 0 \) is a constant that depends only on \( A \) and \( M \). Moreover, if \( f \in D_{\alpha,M}(q) \), there exists a polynomial \( p_n \) of degree \( n \) such that:

\[
\| f - p_n \| \leq \frac{K'}{n^{q+\alpha}},
\]
where $K' > 0$ is a constant that depends on $A$, $q$, $\alpha$ and $M$. Indeed, in its most general formulation, Jackson’s theorem applies to arbitrary continuous functions $f$, relating the degree of approximation to the modulus of continuity. As such, it provides a more precise version of Weierstrass’ theorem.

The class of all $n$-th degree polynomials is larger than needed for the purpose of defining nets over the bounded regression classes we are interested in. Let $B > 0$ denote the constant that bounds all functions in $\mathcal{F}$. With given $\gamma > 0$, define $P'_n = \{ p \in P_n : \| p \| \leq (1 + \gamma)B \}$. By virtue of the triangle inequality, any polynomial used to approximate $f$ as in (9.31) or (9.32) satisfies a bound slightly above and arbitrarily close to $B$ with increasing $n$. Hence, for large enough $n$, $P'_n$ is a $L/n^\beta$-net over $C_{\beta,M}[-A,A]$, where $L > 0$ is a constant that depends only on the constants defining the regression class. For these finite-dimensional, bounded subsets of $C[-A,A]$, the order of suitable nets can be calculated. The upper-bound for the metric entropy of Lipschitz and smoothness classes based on Jackson’s theorem takes the following form.

**Lemma 9.14.** Let $\beta > 0$ and $M > 0$ be given. There exists a constant $K' > 0$ such that:

$$\log N(\epsilon, C_{\beta,M}[-A,A], \| \cdot \|) \leq K' \left( \frac{1}{\epsilon} \right)^{1/\beta} \log \frac{1}{\epsilon},$$

for small enough $\epsilon > 0$.

**Proof.** Let $\epsilon > 0$ be given and choose $n$ to be the smallest integer satisfying $n^\beta \geq 1/\epsilon$. Define $P''_n = \{ p \in P_n : \| p \| \leq L \}$ for some $L > B$. As argued after (9.32), there is a uniformly bounded set $P'_n$ of polynomials of degree $n$ that forms an $\epsilon$-net over $C_{\beta,M}[-A,A]$. If $n$ is chosen large enough, $P'_n$ is a proper subset of $P''_n$. To calculate an upper bound for the covering number of $P'_n$, let $\delta > 0$ be given and let $p_1, \ldots, p_D$ be a (maximal) set of $\delta$-separated polynomials in $P'_n$, where $D$ is the packing number $D(\delta, P'_n, \| \cdot \|)$. Note that the balls $B_i = \{ p \in P'_n : \| p - p_i \| < \frac{1}{2} \delta \}$, $i = 1, \ldots, D$, do not intersect. If $\delta$ is chosen small enough, $B_i \subset P'_n$. The linear map $\hat{p} : \mathbb{R}^{n+1} \to P_n$ that takes a vector $(a_0, \ldots, a_n)$ into the polynomial $\sum_{m=0}^{n} a_m z^m$ is Borel measurable and is used to define the sets $C_i = \hat{p}^{-1}(B_i)$. Note that the sets $C_i$ are obtained from $C = \hat{p}^{-1}(P''_n)$ by rescaling and translation for all $i$. By the same argument as used in the proof of lemma 9.33, we conclude that there is a constant $L$ such that the packing number satisfies:

$$D(\delta, P'_n, \| \cdot \|) \leq \left( \frac{L}{\delta} \right)^{n+1},$$

for small enough $\delta > 0$, which serves as an upper bound for the covering number as well. Choosing $\delta$ equal to a suitable multiple of $n^{-\beta}$ for large enough $n$, we find a constant $K' > 0$ and a net over $C_{\beta,M}[-A,A]$ in $P_n$ of order bounded by $(n^\beta)^{n+1}$. The triangle inequality then guarantees the existence of a slightly less dense net over $C_{\beta,M}[-A,A]$ inside $C_{\beta,M}[-A,A]$ of the same order. We conclude that there exists a constant $K'' > 0$ such that:

$$\log N(\epsilon, C_{\beta,M}[-A,A], \| \cdot \|) \leq K'' n \log n^\beta,$$
for large enough $n$, which leads to the stated bound upon substitution of the relation between $\varepsilon$ and $n$.

The power of $\varepsilon$ in the bound asserted by the above lemma is that of lemma 9.13. The logarithmic correction can be traced back to the $n$-dependence of the radius of the covering balls $B_i$, i.e. the necessity of using finer and finer nets over $P'_n$ to match the $n$-dependence in the degree of approximation. Therefore, there is no obvious way of adapting the above proof to eliminate the $\log(1/\varepsilon)$-factor and Kolmogorov’s approach gives a strictly smaller bound on the entropy. However, the above illustrates the origin of the $\beta$-dependence in the power of $\varepsilon$ more clearly.

For parametric classes (as given under (iii) in the beginning of section 9.2), the entropy is bounded in the following lemma.

**Lemma 9.15.** For a parametric class $\mathcal{F}_\Theta$, there exists a constant $K > 0$ such that the metric entropy is bounded as follows:

$$\log N(\varepsilon, \mathcal{F}_\Theta, \| \cdot \|) \leq K \log \frac{1}{\varepsilon},$$

(9.33)

for small enough $\varepsilon > 0$.

**Proof.** Since, by assumption, $\Theta \subset \mathbb{R}^k$ is bounded by some constant $M' > 0$, the covering numbers of $\Theta$ are upper-bounded by the covering numbers of the ball $B(M', 0) \subset \mathbb{R}^k$ of radius $M'$ centred on 0. Let $\delta > 0$ be given. Since covering numbers are bounded by packing numbers, we see that:

$$N(\varepsilon, \Theta, \| \cdot \|_{\mathbb{R}^k}) \leq D(\delta, B(M', 0), \| \cdot \|_{\mathbb{R}^k}).$$

Let $\theta_1, \ldots, \theta_D$ (with $D = D(\delta, B(M', 0), \| \cdot \|_{\mathbb{R}^k})$) be a maximal $\delta$-separated subset of $B(M', 0)$. The balls $B_i = B\left(\frac{1}{2}\delta, \theta_i\right)$ do not intersect and are all contained in the ball $B(M' + \frac{1}{2}\delta, 0)$ by virtue of the triangle inequality. Therefore, the sum of the volumes of the balls $B_i$ (which are all equal and proportional to $(\frac{1}{2}\delta)^k$, due to translation invariance and scaling behaviour of the Lebesgue measure) lies below the volume of the ball $B(M' + \frac{1}{2}\delta, 0)$. We conclude that:

$$D(\delta, B(M', 0), \| \cdot \|_{\mathbb{R}^k})(\frac{1}{2}\delta)^k \leq (M' + \frac{1}{2}\delta)^k.$$

Assuming that $\delta < 2M'$, we see that:

$$D(\delta, B(M', 0), \| \cdot \|_{\mathbb{R}^k}) \leq \left(\frac{4M'}{\delta}\right)^k.$$

(9.34)

Next, note that due to (9.8), any $\delta$-net over $\Theta$ leads to a $L\delta^p$-net over the regression class $\mathcal{F}_\Theta$, whence we see that:

$$N(L\delta^p, \mathcal{F}_\Theta, \| \cdot \|) \leq N(\delta, \Theta, \| \cdot \|_{\mathbb{R}^k}).$$

(9.35)

Let $\varepsilon > 0$ be given and choose $\delta = (\varepsilon/L)^{1/p}$. Combining (9.34) and (9.35), we find that there exists a constant $K > 0$ such that:
\[ \log N(\varepsilon, \mathcal{F}_\Theta, \| \cdot \|) \leq K \log \frac{1}{\varepsilon} \]

for small enough \( \varepsilon \).

These bounds on the small-\( \varepsilon \) behaviour of the entropy are incorporated in the calculation of bounds for the entropy of the errors-in-variables model through theorem 9.3.

### 9.5.2 Priors on regression classes

This subsection is devoted to the definition of a suitable prior \( \Pi_{\mathcal{F}} \) on the regression class \( \mathcal{F} \). The challenge is to show that \( \Pi_{\mathcal{F}} \) places ‘enough’ mass in small neighbourhoods of any point in the regression class. More specifically, a lower bound is needed for the prior mass of neighbourhoods of the (unknown) regression function \( f_0 \in \mathcal{F} \):

\[ \Pi_{\mathcal{F}} \left( f \in \mathcal{F} : \| f - f_0 \| \leq \delta \right), \]  \[ (9.36) \]

for small enough \( \delta > 0 \) (refer to theorem 9.4).

Jackson’s theorem suggests that a natural definition of a prior on \( \mathcal{F} \) entails the placement of prior mass on all (finite-dimensional) linear spaces of \( n \)-th degree polynomials \( P_n \) on \([-A, A]\), since their union is dense in \( C[-A, A] \) and therefore also in \( \mathcal{F} \). Fix the regression class \( \mathcal{F} \). For all \( n \geq 1 \) we define:

\[ \mathcal{F}_n = \mathcal{F} \cap P_n, \]

i.e. the subsets of \( n \)-th degree polynomials in the regression class. Note that \( \mathcal{F}_n \subset \mathcal{F}_{n+1} \) for all \( n \), and that \( \mathcal{F} \) lies in the closure of their union. The linear map \( \hat{\rho} : \mathbb{R}^{n+1} \rightarrow P_n \) that takes a vector \((a_0, \ldots, a_n)\) into the polynomial \( \sum_{m=0}^{n} a_m z^m \) can be used to define a subset \( \hat{\rho}^{-1}(\mathcal{F}_n) \subset \mathbb{R}^k \) with Lebesgue measure strictly above zero. Normalizing the Lebesgue measure to 1 on \( \hat{\rho}^{-1}(\mathcal{F}_n) \), the inverse map \( \hat{\rho}^{-1} \) serves to define a probability measure \( \Pi_n \) on \( \mathcal{F}_n \). Any sequence \((b_n)_{n \geq 0}\) such that \( b_n \geq 0 \) and \( \sum_{n=0}^{\infty} b_n = 1 \), may be used to define a prior \( \Pi_{\mathcal{F}} \) by the infinite convex combination:

\[ \Pi_{\mathcal{F}}(A) = \sum_{n=0}^{\infty} b_n \Pi_n(A) = \sum_{n=0}^{\infty} b_n \Pi_n(A \cap \mathcal{F}_n), \]  \[ (9.37) \]

for all \( A \) in the Borel \( \sigma \)-algebra generated by the norm topology on \( \mathcal{F} \). Following Huang [122], we refer to priors obtained in this manner as sieve priors.

With a sieve prior, a proof of (9.4) amounts to showing that neighbourhoods of \( f_0 \) have intersections with the sets \( \mathcal{F}_n \) and that the sum of the masses of these intersections is large enough. Obviously, Jackson’s approximation provides a useful way to assert that balls centred on \( f_0 \) intersect with all \( P_n' \) from a certain minimal \( n \) onward. However, as is apparent from (9.36), this is not sufficient, because the
relevant neighbourhoods are restricted to the regression class $\mathcal{F}$. One would have to show that these restricted neighbourhoods intersect with the sets $\mathcal{F}_n$.

Jackson’s theorem does not assert anything concerning Lipschitz-bounds of the approximating polynomial or derivatives thereof. The assertion that $p_n$ approximates $f$ in uniform norm leaves room for very sharp fluctuations of $p_n$ on small scales, even though it stays within a bracket of the form $[f - K/n^\beta, f + K/n^\beta]$. It is therefore possible that $p_n$ lies far outside $\mathcal{F}_n$, rendering neighbourhoods of $p_n$ in $P_n$ unfit for the purpose. Although it is possible to adapt Jackson’s theorem in such a way that the approximating polynomials satisfy a Lipschitz condition that is arbitrarily close to that of the regression class, this adaptation comes at a price with regard to the degree of approximation. As it turns out, this price leads to substantial corrections for the rate of convergence and ultimately to sub-optimality (with respect to the power of $\varepsilon$ rather than logarithmically). That is not to say that sieve priors are in any sense sub-optimal. (Indeed, sieve priors have been used with considerable success in certain situations; for an interesting example, see the developments in adaptive Bayesian estimation, for instance in Huang [122].) The calculation underlying the claims made above merely shows that the construction via adapted versions of Jackson’s theorem does not lead to optimal results, leaving the possibility that a sieve prior satisfies (9.4) open. What it does show, however, is that this may be very hard to demonstrate.

Therefore, we define the prior on the regression class in a different fashion (first proposed in Le Cam (197X) [161], based on ideas from Le Cam (1973) [160]), based on the upper bounds for covering numbers obtained in the previous subsection. Let the regression class $\mathcal{F}$ be a bounded, equicontinuous subset of $C[-A,A]$, so that the covering numbers $N(\varepsilon, \mathcal{F}, \|\cdot\|)$ are finite for all $\varepsilon > 0$. Let $(a_m)_{m \geq 1}$ be a monotonically decreasing sequence, satisfying $a_m > 0$ (for all $m \geq 1$), and $a_m \downarrow 0$. For every $m \geq 1$, there exists an $a_m$-net $\{f_i \in \mathcal{F} : i = 1, \ldots, N_m\}$ over $\mathcal{F}$, where $N_m = N(a_m, \mathcal{F}, \|\cdot\|)$. We define, for every $m \geq 1$, a discrete probability measure $\Pi_m$ that distributes its mass uniformly over the set $\{f_i : i = 1, \ldots, N_m\}$:

$$\Pi_m = \sum_{i=1}^{N_m} \frac{1}{N_m} \delta_{f_i}.$$  

Any sequence $(b_n)_{n \geq 0}$ such that $b_n \geq 0$ and $\sum_{m=0}^{\infty} b_n = 1$, may be used to define a prior $\Pi_{\mathcal{F}}$ on $\mathcal{F}$ by the infinite convex combination:

$$\Pi_{\mathcal{F}}(A) = \sum_{m=0}^{\infty} b_n \Pi_m(A), \quad (9.38)$$

for all $A$ in the Borel $\sigma$-algebra generated by the norm topology on $\mathcal{F}$. Priors defined in this manner are referred to as a net priors and resemble those defined in Ghosal, Ghosh and Ramamoorthi (1997) [99], (see also, Ghosal et al. (2000) [101]).

Note that for all $m \geq 1$ and every $f \in \mathcal{F}$, there is an $f_i$ satisfying $\|f - f_i\| \leq a_m$. So for every $f_0 \in \mathcal{F}$ and all $\delta > 0$, we have:
\[ \Pi_m(\|f - f_0\| \leq \delta) \geq \frac{1}{N_m}, \]

if \( a_m \leq \delta \), i.e. for all \( m \) large enough. This means that the priors \( \Pi_m \) satisfy lower bounds for the mass in neighbourhoods of points in the regression class, that are inversely related to upper bounds satisfied by the covering numbers. As is demonstrated below, choices for the sequences \( a_m \) and \( b_m \) exist such that this property carries over to a prior of the form (9.38).

**Lemma 9.16.** Let \( \beta > 0 \) and \( M > 0 \) be given and define \( \mathcal{F} \) to be the class \( C_{\beta, M}[-A, A] \). There exists a net prior \( \Pi_{\mathcal{F}} \) and a constant \( K > 0 \) such that

\[ \log \Pi_{\mathcal{F}}(f \in \mathcal{F} : \|f - f_0\| \leq \delta) \geq -K \frac{1}{\delta^{1/\beta}}, \tag{9.39} \]

for small enough \( \delta \).

**Proof.** Define, for all \( m \geq 1 \), \( a_m = m - \beta \). Then the covering number \( N_m \) satisfies, for some constant \( K' > 0 \):

\[ \log N_m = \log N(a_m, \mathcal{F}, \| \cdot \|) \leq K'a_m^{-1/\beta} = K'm, \]

according to lemma 9.13. Let \( \delta > 0 \) be given and choose the sequence \( b_m = (1/2)^m \). Let \( M \) be an integer such that:

\[ \frac{1}{\delta^{1/\beta}} \leq M \leq \frac{1}{\delta^{1/\beta}} + 1. \]

Then for all \( m \geq M \), \( a_m \leq \delta \) and, due to the inequality (9.38), the net prior \( \Pi_{\mathcal{F}} \) satisfies:

\[ \Pi_{\mathcal{F}}(f \in \mathcal{F} : \|f - f_0\| \leq \delta) \geq \sum_{m \geq M} b_m \Pi_m(\|f - f_0\| \leq \delta) \geq \sum_{m \geq M} \left( \frac{e^{-K'}}{2} \right)^m \geq \frac{1}{2} e^{-K'M} \geq \frac{1}{2} e^{-K'(\delta^{-1/\beta} + 1)} \geq \frac{1}{2} e^{-2K'\delta^{-1/\beta}}, \tag{9.40} \]

for small enough \( \delta \).

For parametric classes, the prior mass in neighbourhoods of \( f_0 \) is lower-bounded in the following lemma.

**Lemma 9.17.** Assume that the regression class \( \mathcal{F} \) is parametric: \( \mathcal{F} = \mathcal{F}_\Theta \). Any prior \( \Pi_\Theta \) on \( \Theta \) induces a prior \( \Pi_{\mathcal{F}} \) with the Borel \( \sigma \)-algebra generated by the topology of the norm \( \| \cdot \| \) as its domain. Furthermore, if \( \Pi_\Theta \) is dominated by the Lebesgue measure and has a density that is strictly positive at \( \theta_0 \), then there exists a constant \( R > 0 \) such that the prior mass in neighbourhoods of \( f_0 \) is bounded as follows:

\[ \log \Pi_{\mathcal{F}}(f \in \mathcal{F} : \|f - f_0\| \leq \varepsilon) \geq -R \log \frac{1}{\varepsilon}, \tag{9.41} \]
for small enough $\varepsilon > 0$.

Proof. The Lipschitz condition (9.8) ensures that the map $\hat{f} : \Theta \rightarrow \mathcal{F}_\Theta : \theta \mapsto f_\theta$ is continuous, implying measurability with respect to the corresponding Borel $\sigma$-algebras. So composition of $\Pi_\Theta$ with $\hat{f}^{-1}$ induces a suitable prior on $\mathcal{F}_\Theta$. As for the second assertion, let $\delta > 0$ be given. Since $\Pi_\Theta$ has a continuous Lebesgue density $\pi : \Theta \rightarrow \mathbb{R}$ that satisfies $\pi(\theta_0) > 0$ by assumption and since $\theta_0$ is internal to $\Theta$, there exists an open neighbourhood $U \subset \Theta$ of $\theta_0$ and a constant $\pi_1 > 0$ such that $\pi(\theta) \geq \pi_1$ for all $\theta \in U$. Therefore, for all balls $B(\delta, \theta_0) \subset U$ (i.e. for small enough $\delta > 0$), we have:

$$\Pi_\Theta(B(\delta, \theta_0)) = \int_{B(\delta, \theta_0)} \pi(\theta) d\theta \geq V_k \pi_1 \delta^k,$$

where $V_k$ is the Lebesgue measure of the unit ball in $\mathbb{R}^k$. Note that due to property (9.8),

$$\{ \theta \in \Theta : \|\theta - \theta_0\| \leq \delta \} \subset \{ \theta \in \Theta : \|f_\theta - f_0\| \leq L\delta^\rho \},$$

so that, for given $\varepsilon > 0$ and the choice $\delta = (\varepsilon/L)^{1/\rho}$:

$$\log \Pi_{\mathcal{F}}(f \in \mathcal{F} : \|f - f_0\| \leq \varepsilon) \geq \log \Pi_\Theta(\theta \in \Theta : \|\theta - \theta_0\| \leq (\varepsilon/L)^{1/\rho})$$

$$\geq \log(V_k \pi_1 (\varepsilon/L)^{k/\rho}) \geq -R \log \frac{1}{\varepsilon},$$

for some constant $R > 0$ and small enough $\varepsilon$.

The bounds on the small-$\varepsilon$ behaviour of prior mass presented in this subsection are incorporated in the calculation of bounds for the prior mass of Kullback-Leibler neighbourhoods $B(\varepsilon; P_0)$ through theorem 9.4.

9.6 Exercises [EMPTY]
Chapter 10
Application: community detection in the planted bi-section model

To demonstrate how the methods presented in chapter 7 are applied in practice, we consider a sparse stochastic block model, focusing on the questions of community recovery, detection and uncertainty quantification.

10.1 Communities in random graphs

The stochastic block model is a generalization of the Erdős-Rényi random graph model [85] where one studies a version $X^n$ of the complete graph between $n$ vertices under percolation, with edge probability $p_n \in [0, 1]$. Stochastic block models [120] are similar but concern random graphs with vertices that belong to one of several classes and edge probabilities that depend on those classes. If we think of the graph $X^n$ as data and the class assignments of the vertices as unobserved, an interesting statistical challenge presents itself regarding estimation of (and other forms of inference on) the vertices’ class assignments, a task referred to as community detection [107]. The stochastic block model and its generalizations have applications in physics, biology, sociology, image processing, genetics, medicine, logistics, etcetera and are widely employed as a canonical model to study clustering and community detection [91, 1] [229] even state that, “Community detection for the stochastic block model is probably the most studied topic in network analysis.”

In an asymptotic sense, one may wonder under which conditions on edge probabilities community detection can be done in a ‘statistically consistent’ way as the number of vertices $n$ grows, that is, whether it is possible to estimate the true class assignments correctly (or correctly for a fraction of the vertices that goes to one) with high probability. Note that already in the Erdős-Rényi model, asymptotic behaviour is very rich: consider, for example, connectedness of the Erdős-Rényi graphs $X^n$ with a sequence of edge probabilities $(p_n)$ that becomes sparse: $p_n \to 0$. A sharp transition exists: $p_n \geq (A/n) \log n$ leads to a connected graph with high probability, if and only if $A > 1$. In even more sparse circumstances, there is a single giant component (a connected component of size $O(n)$) with high probability,
if and only if $p_n \geq C/n$ with $C > 1$. Below the $1/n$-threshold, the graph $X_n$ fragments into many disconnected sub-graphs of order no larger than $O(\log(n))$ with high probability. At the boundaries $1/n$ and $\log(n)/n$, the Erdős-Rényi graph is said to undergo phase transitions [39], from the fragmented phase to the sparse Kesten-Stigum phase, and then to the less sparse Chernoff-Hellinger phase.

Here and in [2, 171, 178], the community detection problem is studied in the simplest context, that of the so-called planted bi-section model, which is a stochastic block model with two classes, each of $n$ vertices and edge probabilities $p_n$ (within-class) and $q_n$ (between-class). A famous sufficient condition for so-called exact recovery of the class assignments in the planted bi-section model comes from [79]: if there exists a constant $A > 0$ such that,

$$p_n - q_n \geq A \frac{\log n}{n},$$

(10.1)

then community detection by simple minimization of the number of edges between estimated classes constitutes an estimator for the class assignments that concentrates on the true class assignment with high probability. Restriction (10.1) is expressed most naturally in the Chernoff-Hellinger phase and excludes the Kesten-Stigum phase.

Detection of the class assignment poses the weaker requirement that the fraction of vertices that are classified correctly goes to one with high probability (see definition 10.1): it was conjectured in [61, 62] to be possible in block models, if, with $c_n$ and $d_n$ such that $p_n = c_n/n$ and $q_n = d_n/n$.

$$(c_n - d_n)^2 > 2(c_n + d_n).$$

(10.2)

Essentially Decelle et al. argue that in the Kesten-Stigum phase random graphs like that of the planted bi-section model allow estimation of the underlying class assignment, only if their distribution is sufficiently dissimilar from that of an Erdős-Rényi graph ($p_n = q_n$ makes class assignments indistinguishable, and correspondingly, small values for $p_n - q_n$ define a regime in which inference on the class assignment is relatively difficult, see [7]). An analogous claim in the Chernoff-Hellinger phase was first considered more rigorously in [171] and later confirmed, both from a probabilistic/statistical perspective in [176, 178], and independently from an information theoretic perspective in [2]. Defining $a_n$ and $b_n$ by $np_n = a_n\log(n)$ and $nq_n = b_n\log(n)$ and assuming that $C^{-1} \leq a_n, b_n \leq C$ for all but finitely many $n \geq 1$, the class assignment in the planted bi-section model can be recovered exactly [178], if and only if,

$$(a_n + b_n - 2\sqrt{a_nb_n} - 1)\log n + \frac{1}{2} \log \log n \rightarrow \infty.$$

(10.3)

[178] also find a sharp condition for detection in the Kesten-Stigum phase. To summarize, the sparse Erdős-Rényi phases as well as the proximity of the Erdős-Rényi sub-model in the parameter space play a role in the statistical perspective on communities in the stochastic block model.
Estimation methods used for the community detection problem include spectral clustering (see [149] and many others), maximization of the likelihood and other modularities [107, 30, 53, 5], semi-definite programming [113, 112], and penalized ML detection of communities with minimax optimal misclassification ratio [229, 97]. More generally, we refer to [1] and the very informative introduction of [97] for extensive bibliographies and a more comprehensive discussion. Bayesian methods have been popular throughout, e.g. the original work by Snijders and Nowicki (2001) [182], the work of [61, 62] and more recently, [210], based on an empirical prior choice, and [177].

Interest in the stochastic block model has generated a wealth of algorithms that estimate the class assignment. Naturally, great emphasis has been placed on computational efficiency of these algorithms. Methods that maximize the likelihood or other modularities [107, 30] do not compare well in this respect. For example, [97] argue along these lines (and favour a localized method over global (penalized) ML-estimation) [229]. Broadly speaking, MCMC methods to simulate from posteriors are comparable to the relevant maximization methods as far as computational burdens are concerned.

Given that algorithms exists with much more favourable computational properties, calculation or simulation of posteriors seems unnecessarily laborious. As long as point-estimation is the only goal there is little to argue, but estimation is only the first step in statistical inference: when a consistent estimator has been found, immediate questions regarding (limiting) accuracy and reliability arise. From a Bayesian perspective, the posterior provides estimates of accuracy and credibility without further process, but to the algorithmic frequentist relying on a point-estimator, more detailed inferential questions concerning the sampling distributions of point-estimators are often (prohibitively) hard to analyse. In the stochastic block model, questions concerning accuracy have been addressed in [229], but, to the best of the authors’ knowledge, frequentist uncertainty quantification with confidence sets for class assignment has not been addressed in the literature.

In this chapter, our goal is to explore the limits of what is possible from the statistical point of view, similar to what Mossel et al. do from the probabilistic point of view and Abbe et al. from the information theoretic point of view. In section 10.4, it is shown that posteriors detect and recover underlying class assignments under conditions on $(p_n, q_n)$ that are sharp. In section 10.5, it is demonstrated that Bayesian credible sets can be converted to asymptotically consistent confidence sets in various ways [146]. If we assume posterior consistency in the form of exact recovery (or detection), credible sets (or enlarged credible sets) are consistent confidence sets. Moreover, if credible levels grow to one quickly enough, credible sets can be interpreted as frequentist confidence sets without conditions on the parameters. When,

$$n|p_n - q_n| \to 0, \quad n\left|p_n^{1/2} - q_n^{1/2}\right| \to \infty,$$

close to the Erdős-Rényi submodel where communities are the hardest to distinguish, credible sets may be enlarged to achieve frequentist asymptotic coverage. We conclude that, in the context of the planted bi-section model (and also much wider),
10.2 The planted bi-section model

In a stochastic block model, each vertex is assigned to one of $K \geq 2$ classes through an unobserved class assignment vector $\theta'$. Each vertex belongs to a class and any edge occurs (independently of others) with a probability depending on the classes of the vertices that it connects. In the planted bi-section model, there are only two classes ($K = 2$) and, at the $n$-th iteration ($n \geq 1$), there are $2n$ vertices (labelled with indices $1 \leq i \leq 2n$), $n$ in each class, with class assignment vector $\theta' \in \Theta'_n$ (with components $\theta'_1, \ldots, \theta'_{2n} \in \{0, 1\}$), where $\Theta'_n$ is the subset of $\{0, 1\}^{2n}$ of all finite binary sequences that contain as many ones as zeroes. Denote that space in which the random graph $X^n$ takes its values by $\mathcal{X}_n$ (e.g. represented by its adjacency matrix with entries $\{X_{ij}: 1 \leq i, j \leq 2n\}$). The (n-dependent) probability of an edge occurring ($X_{ij} = 1$) between vertices $1 \leq i, j \leq 2n$ within the same class is denoted $p_n \in (0, 1)$; the probability of an edge between classes is denoted $q_n \in (0, 1)$.

$$Q_{ij}(\theta') := P_{\theta', p}(X_{ij} = 1) = \begin{cases} p_n, & \text{if } \theta'_{n,i} = \theta'_{n,j}, \\ q_n, & \text{if } \theta'_{n,i} \neq \theta'_{n,j}. \end{cases} \quad (10.4)$$

Note that if $p_n = q_n$, $X^n$ is the Erdős-Rényi graph $G(2n, p_n)$ and the class assignment $\theta_n \in \Theta'_n$ is not identifiable. Another identifiability issue that arises is that the model is invariant under interchange of class labels 0 and 1. This is expressed in the parameter spaces $\Theta'_n$ through equivalence relations: $\theta'_1 \sim_n \theta'_2$, if $\theta'_{2,n} = -\theta'_{1,n}$ (by componentwise negation). To prevent non-identifiability, we parametrize the model for $X^n$ in terms of a parameter $\theta_n$ in a quotient space $\Theta_n = \Theta'_n/\sim_n$, for every $n \geq 1$. For $\theta'_n \in \Theta'_n$ we denote the equivalence class $\{\theta'_{n,i} = -\theta'_{n,i}\}$ by $\theta_n$. Note that the set $\Theta_n$ can be identified with the set of partitions of $\{1, \ldots, 2n\}$ consisting of exactly two sets, via the identification

$$\theta_n \leftrightarrow \{ \{i: \theta'_{n,i} = 0\}, \{i: \theta'_{n,i} = 1\} \},$$

and note that this is independent of the choice of the representation.
The probability measure for the graph $X^n$ corresponding to parameter $\theta$ is denoted $P_{\theta,n}$. The likelihood is given by,
\[
p_{\theta,n}(X^n) = \prod_{i<j} Q_{i,j}(\theta)^{X_{i,j}}(1 - Q_{i,j}(\theta))^{1 - X_{i,j}}.
\]

For the sparse versions of the planted bi-section model, we also define edge probabilities that vanish with growing $n$: take $(a_n)$ and $(b_n)$ such that $a_n \log n = np_n$ and $b_n \log n = nq_n$ for the Chernoff-Hellinger phase; take $(c_n)$ and $(d_n)$ such that $c_n = np_n$ and $d_n = nq_n$ for the Kesten-Stigum phase. The fact that we do not allow loops (edges that connect vertices with themselves) leaves room for $2 \cdot 2^n(n - 1) + n^2 = 2n^2 - n = \frac{1}{2} \cdot (2n)(2n - 1)$ possible edges in the random graph $X^n$ observed at iteration $n$.

The statistical question of interest in this model is to reconstruct the unobserved class assignment vectors $\theta_n$ consistently, that is, correctly with probability growing to one as $n \to \infty$. This can be stated in a strong and in a weak version, defined below.

**Definition 10.1.** Let $\theta_{0,n} \in \Theta_n$ be given. An estimator sequence $\hat{\theta}_n : \mathcal{X}_n \to \Theta_n$ is said to recover the class assignment $\theta_{0,n}$ exactly if,
\[
P_{\theta_{0,n}}(\hat{\theta}_n(X^n) = \theta_{0,n}) \to 1,
\]
that is, if $\hat{\theta}_n$ indicates the correct partition assignment with high probability.

Based on the definition of $k$ (just before eq. (10.9) below) we also relax this consistency requirement somewhat in the form of the following definition, c.f. [178] and others.

**Definition 10.2.** Let $\theta_{0,n} \in \Theta_n$ be given. An estimator sequence $\hat{\theta}_n : \mathcal{X}_n \to \Theta_n$ is said to detect the class assignment $\theta_{0,n}$ if,
\[
\frac{1}{2n} \left| \sum_{i=1}^{2n} (-1)^{\hat{\theta}_i}(-1)^{\theta_{0,i}} \right| P_{\theta_{0,n}} \to 1,
\]
that is, if the fraction of correct assignments in $\hat{\theta}_n$ grows to one with high probability.

Below, we specialize to the Bayesian approach: we choose prior distributions $\pi_n$ for all $\Theta_n$, $(n \geq 1)$ and calculate the posterior: denoting the likelihood by $p_{\theta,n}(X^n)$, the posterior for the parameter $\theta_n$ is written as a fraction of sums, for all $A \subset \Theta_n$,
\[
\Pi(A | X^n) = \sum_{\theta_n \in A} p_{\theta,n}(X^n) \pi_n(\theta_n) \left/ \sum_{\theta_n \in \Theta_n} p_{\theta,n}(X^n) \pi_n(\theta_n) \right.,
\]
where $\pi_n : \Theta_n \to [0, 1]$ is the probability mass function prior $\Pi_n$. Here, we only consider uniform priors $(\Pi_n)$ for $\theta_n \in \Theta_n$, so for all $n \geq 1$ and $\theta_n \in \Theta_n$, $\pi(\theta_n) = \pi_n := (|\Theta_n|)^{-1}$. 
10.3 Existence of suitable tests

Given \( n \geq 1 \) and two class assignment vectors \( \theta_0, \theta_n \in \Theta_n \), we are interested in calculation of the likelihood ratio \( \frac{dP_{\theta_n}}{dP_{\theta_0}} \), because it determines testing power as well as the various forms of remote contiguity that play a role.

Choose representations \( \theta'_0 \) of \( \theta_0 \) and \( \theta'_n \) of \( \theta_n \) so that \( k'((\theta'_0, \theta'_n)) = k(\theta_0, \theta) \), where \( k \) and \( k' \) are as in section 10.4. Recall that, \( Z_n(\theta'_0) \subset \{1, \ldots, 2n\} \) is class zero and the complement \( Z'_n(\theta'_0) \) class one. For the sake of presentation (in figure 10.1 below), relabel the vertices such that \( Z(\theta'_0) = \{1, \ldots, n\} \) and \( Z'(\theta'_0) = \{n+1, \ldots, 2n\} \). In the case \( n = 4 \), figure 10.1 shows edge probabilities in the familiar block arrangement.

Recall that the likelihood under \( \theta_0 \) is given by,

\[
p_{\theta_0}(X^n) = \prod_{i<j} Q_{i,j}(\theta_0)^{X_{ij}} (1 - Q_{i,j}(\theta_0))^{1-X_{ij}}.
\]

If we assume that \( \theta'_0 \) and \( \theta'_n \) differ by \( k \) pair-exchanges among respective members of the zero- and one-classes, then a look at figure 10.1 reveals that the likelihood-ratio depends only on the edges for which exactly one of its end-points changes class. Define,

\[
A_n = \{(i, j) \in \{1, \ldots, 2n\} : i < j, \theta_{0,n,i} = \theta'_{0,n,i}, \theta'_{n,j} \neq \theta_{n,j}\},
\]

\[
B_n = \{(i, j) \in \{1, \ldots, 2n\} : i < j, \theta'_{0,n,i} \neq \theta_{0,n,i}, \theta'_{n,j} = \theta_{n,j}\}.
\]

Also define,

\[
(S_n, T_n) := \bigg(\sum_{(i,j) \in A_n} X_{ij}, \sum_{(i,j) \in B_n} X_{ij}\bigg),
\]
and note that the likelihood ratio can be written as,

$$\frac{p_{\theta,n}(X^n)}{p_{\theta_0,n}(X^n)} = \left(\frac{1 - p_n}{p_n}\right)^{S_n - T_n}$$

(10.5)

where,

$$(S_n, T_n) \sim \begin{cases} \text{Bin}(2k(n-k), p_n) \times \text{Bin}(2k(n-k), q_n), & \text{if } X^n \sim P_{\theta_0,n}, \\ \text{Bin}(2k(n-k), q_n) \times \text{Bin}(2k(n-k), p_n), & \text{if } X^n \sim P_{\theta,n}. \end{cases}$$

(10.6)

Based on that, we derive the following lemma.

**Lemma 10.1.** Let $n \geq 1$, $\theta_{0,n}, \theta_n \in \Theta_n$ be given. Assume that $\theta_{0,n}$ and $\theta_n$ differ by $k$ pair-exchanges. Then there exists a test function $\phi_n : X_n \rightarrow [0, 1]$ such that,

$$P_{\theta_{0,n}}(X^n) + P_{\theta,n}(1 - \phi_n(X^n)) \leq a_{n,k},$$

with testing power,

$$a_{n,k} = (1 - p_n - q_n + 2p_nq_n + 2\sqrt{p_n(1 - p_n)\sqrt{q_n(1 - q_n)}})^{2k(n-k)}.$$

**Proof.** The likelihood ratio test $\phi_n(X^n)$ has testing power bounded by the so-called Hellinger transform,

$$P_{\theta_{0,n}}(X^n) + P_{\theta,n}(1 - \phi_n(X^n)) \leq \inf_{0 \leq \alpha \leq 1} P_{\theta_{0,n}} \left( \frac{p_{\theta,n}(X^n)}{p_{\theta_{0,n}}(X^n)} \right)^{\alpha},$$

(see, e.g. [162] and proposition 2.6 in [146]). Using $\alpha = 1/2$ (which is the minimum), we find that,

$$P_{\theta_{0,n}} \left( \frac{p_{\theta,n}(X^n)}{p_{\theta_{0,n}}(X^n)} \right)^{1/2} = P_{\theta_{0,n}} \left( \frac{p_n}{1 - p_n} \right)^{S_n - T_n} = Pe^{\frac{1}{2} \lambda_n S_n} Pe^{-\frac{1}{2} \lambda_n T_n}$$

where $\lambda_n := \log(1 - p_n) - \log(p_n) + \log(q_n) - \log(1 - q_n)$ and $(S_n, T_n)$ are distributed binomially, as in the first part of (10.6). Using the moment-generating function of the binomial distribution, we conclude that,
\begin{align*}
P_{\theta_0, n} \left( \frac{p_{\theta_0, n}(X^n)}{p_{\theta_0, n}} \right)^{1/2} &= \left( 1 - p_n + p_n \left( \frac{1 - p_n}{p_n} \frac{q_n}{1 - q_n} \right)^{1/2} \right) \\
&\times \left( 1 - q_n + q_n \left( \frac{p_n}{1 - p_n} \frac{1 - q_n}{q_n} \right)^{1/2} \right)^{2k(n-k)} \\
&= \left( (1 - p_n) + p_n \frac{q_n^{1/2}}{\frac{1 - q_n}{q_n}} \left( \frac{1 - q_n}{1 - p_n} \right)^{1/2} \right)^{2k(n-k)} \\
&= \left( (1 - p_n)(1 - q_n) + 2(p_n q_n (1 - p_n)(1 - q_n))^{1/2} + p_n q_n \right)^{2k(n-k)}
\end{align*}

which proves the assertion.

### 10.4 Exact recovery and detection with posteriors

Consider the sequence of experiments in which we observe random graphs $X^n \in \mathcal{P}_n$, generated by the planted bi-section model of definition (10.4). We assume, for every $n \geq 1$, that the prior is the uniform distribution over $\Theta_n$: since we can choose $n$ vertices from a total of $2n$ in $\binom{2n}{n}$ ways and $\theta \sim -\theta$, $\pi_n = \left( \frac{1}{2} \binom{2n}{n} \right)^{-1}$.

Given true parameters $\theta_{0,n} \in \Theta_n \ (n \geq 1)$, choose representations $\theta_{0,n}^l \in \Theta_n'$ and define $Z_n(\theta_{0,n}^l) \subset \{1, \ldots, 2n\}$ to be class zero (the set of all those $i$ such that $\theta_{0,i}^l = 0$) and call the complement $Z_n^c(\theta_{0,n}^l)$ class one. For the questions concerning exact recovery and detection, we are interested in the sets $V_{n,k} \subset \Theta_n'$, defined to contain all those $\theta_{0,n}^l$ that differ from $\theta_{0,n}^l$ by exactly $k$ exchanges of pairs: for $\theta_{0,n}^l \in \Theta_n'$ we have $\theta_{0,n}^l \in V_{n,k}$, if the set of vertices in class zero c.f. $\theta_{0,n}^l$, $Z(\theta_{0,n}^l) = \{1 \leq i \leq 2n : \theta_{0,n,i}^l = 0\}$, from which we leave out the set of vertices in class zero c.f. $\theta_{0,n}^l$, $Z(\theta_{0,n}^l) = \{1 \leq i \leq 2n : \theta_{0,n,i}^l = 0\}$, has $k$ elements. Conversely, for any $\theta_{1,n}^l$ and $\theta_{2,n}^l$ in $\Theta_n'$, we denote the minimal number of pair-exchanges necessary to take $\theta_{1,n}^l$ into $\theta_{2,n}^l$ by $k'(\theta_{1,n}^l, \theta_{2,n}^l)$. Note that $k'(\theta_{1,n}^l, -\theta_{1,n}^l) = n - k'(\theta_{1,n}^l, \theta_{2,n}^l)$, which leads to the distance measure between two representation classes

\begin{equation}
k(\theta_{1,n}, \theta_{2,n}) = k'(\theta_{1,n}^l, \theta_{2,n}^l) \wedge k'(\theta_{1,n}^l, -\theta_{2,n}^l)
\end{equation}

and note that this is independent of choice of the representations and that this function $k$ takes values in $\{0, \ldots, \lfloor n/2 \rfloor\}$. Now define,

\begin{equation}
V_{n,k} = V_{n,k}(\theta_{0,n}) = \{ \theta_n : k(\theta_n, \theta_{0,n}) = k \} = \{ \theta_n : \theta_{n}^l \in V_{n,k} \},
\end{equation}
for \( k \in \{1, \ldots, \lfloor n/2 \rfloor \} \). Given some sequence \((k_n)\) of positive integers we then define \(V_n\) as the disjoint union,

\[
V_n = \bigcup_{k=k_n}^{\lfloor n/2 \rfloor} V_{n,k}
\]

(10.9)

Since we can choose two subsets of \( k \) elements from two sets of size \( n \) in \( \binom{n}{k}^2 \) ways, the cardinal of \( V_{n,k} \) is \( \frac{n}{k} \), when \( k < n/2 \) and \( \frac{1}{2} \binom{n}{k}^2 \) when \( n \) is even and \( k = n/2 \). In both cases the number of elements in \( V_{n,k} \) is therefore bounded by \( \binom{n}{k}^2 \).

According to lemma 2.2 in [146] (with \( B_n = \{ \theta_{0,n} \} \)), for any test sequences \( \phi_{k,n} : \mathcal{R}_n \to [0,1] \) \( (k \geq 1, n \geq 1) \), we have,

\[
P_{\theta_{0,n}} \Pi(V_n|X^n) = \sum_{k=k_n}^{\lfloor n/2 \rfloor} P_{\theta_{0,n}} \Pi(V_{n,k}|X^n)
\]

\[
\leq \sum_{k=k_n}^{\lfloor n/2 \rfloor} \left( P_{\theta_{0,n}} \phi_{k,n}(X^n) + \sum_{\theta_n \in V_{n,k}} P_{\theta_n} (1 - \phi_{k,n}(X^n)) \right)
\]

for every \( n \geq 1 \). Suppose that for any \( k \geq 1 \) there exists a sequence \((a_{n,k})_{n \geq 1} \), \( a_{n,k} \downarrow 0 \) and, for any \( \theta_n \in V_{n,k} \), a test function \( \phi_{\theta_n,n} \) that distinguishes \( \theta_0,n \) from \( \theta_n \) as follows,

\[
P_{\theta_{0,n}} \phi_{\theta_{0,n}}(X^n) + P_{\theta_{0,n}}(1 - \phi_{\theta_{0,n}}(X^n)) \leq a_{n,k}, \tag{10.10}
\]

for all \( n \geq 1 \). Then using test functions \( \phi_{\theta_n,n}(X^n) = \max\{\phi_{\theta_n,n}(X^n) : \theta_n \in V_{n,k} \} \), as well as the fact that,

\[
P_{\theta_{0,n}} \phi_{\theta_n,n}(X^n) \leq \sum_{\theta_n \in V_{n,k}} P_{\theta_{0,n}} \phi_{\theta_n,n}(X^n),
\]

we see that,

\[
P_{\theta_{0,n}} \Pi(V_n|X^n) \leq \sum_{k=k_n}^{\lfloor n/2 \rfloor} \sum_{\theta_n \in V_{n,k}} \left( P_{\theta_{0,n}} \phi_{\theta_n,n}(X^n) + P_{\theta_n} (1 - \phi_{\theta_n,n}(X^n)) \right)
\]

\[
\leq \sum_{k=k_n}^{\lfloor n/2 \rfloor} \left( \frac{n}{k} \right)^2 a_{k,n}.
\]

This inequality forms the basis for the results on exact recovery and detection in the next two subsections.
10.4.1 Posterior consistency: exact recovery

For the case of exact recovery, we are interested in the expected posterior masses of subsets of $\Theta_n$ of the form:

$$V_n = \{ \theta_n \in \Theta_n : \theta_n \neq \theta_{0,n} \} = \bigcup_{k=1}^{\lfloor n/2 \rfloor} V_{n,k}. $$

The theorem states a sufficient condition for $(p_n)$ and $(q_n)$, which is related to requirement (10.3) in the Chernoff-Hellinger phase.

**Theorem 10.1.** For some $\theta_{0,n} \in \Theta_n$, assume that $X^n \sim P_{\theta_{0,n}}$, for every $n \geq 1$. If we equip every $\Theta_n$ with its uniform prior and $(p_n)$ and $(q_n)$ are such that,

$$\left( 1 + (1 - p_n - q_n + 2p_nq_n + 2\sqrt{p_n(1 - p_n)q_n(1 - q_n)})^{n/2} \right)^{2n} \rightarrow 1, \quad (10.11)$$

as $n \rightarrow \infty$, then,

$$\Pi(\theta_n = \theta_{0,n} | X^n) \xrightarrow{p_{\theta_{0,n}}} 1, \quad (10.12)$$

as $n \rightarrow \infty$, i.e. the posterior recovers the community assignment exactly.

**Proof.** According to lemma 10.1, for every $n \geq 1, k \geq 1$ and given, $\theta_{0,n}$, there exists a test sequence satisfying (10.10) with $a_{n,k} = (1 - \mu_n)^{2k(n-k)}$ and $\mu_n = p_n + q_n - 2p_nq_n - 2(p_n(1 - p_n)q_n(1 - q_n))^{1/2} \in [0, 1]$. Therefore, with $z_n = (1 - \mu_n)^{n/2}$,

$$P_{\theta_{0,n}}\Pi(V_n | X^n) \leq \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{k} 2p_nq_n \leq \sum_{k=1}^{\lfloor n/2 \rfloor} \frac{n}{k} (1 - \mu_n)^{nk} \leq \sum_{k=1}^{\lfloor n/2 \rfloor} \frac{2n}{k} (1 - \mu_n)^{nk} \leq \sum_{l=1}^{n} \frac{2n}{l} (1 + z_n)^{2n} - 1 \quad (10.13)$$

The right-hand side goes to zero if (10.11) is satisfied.

**Example 10.1.** Consider (10.11) in the sparse Chernoff-Hellinger phase, where $np_n = a_n \log n, nq_n = b_n \log n$ with $a_n, b_n = O(1)$. In that case,

$$\left( 1 + (1 - p_n - q_n + 2p_nq_n + 2\sqrt{p_n(1 - p_n)q_n(1 - q_n)})^{n/2} \right)^{2n} \approx \left( 1 + (1 - (a_n + b_n - 2\sqrt{a_nb_n} + o(n^{-1}\log n)) \log n) / n \right)^{n/2} \approx (1 + \frac{1}{n} (a_n + b_n - 2\sqrt{a_nb_n} - 2))^{2n} = \exp\left(2e^{-\frac{1}{2}(a_n + b_n - 2\sqrt{a_nb_n} - 2)}\right)^{2n}$$

(10.13)
Accordingly, in the Chernoff-Hellinger phase (10.11) amounts to the sufficient condition,

\[(a_n + b_n - 2 \sqrt{a_nb_n} - 2) \log n \to \infty,\]  

(10.14)

which closely resembles (but is not exactly equal to) (10.3), the requirement of [178]. In fact there is a trade-off: (10.3) is slightly weaker than (10.14) but applies only if there exists a \(C > 0\) such that \(C^{-1}a_n b_n \leq C\) for large enough \(n\) [178, 229]. More particularly, one of the sequences \((a_n)\) and \((b_n)\) may fade away with growing \(n\) or equal zero outright. For example if \(b_n = 0\) and \(\liminf_n a_n > 2\), edges between classes are completely absent but, separately, the Erdös-Rényi graphs spanned by vertices in \(Z(\theta_0')\) and \(Z(\theta_0')\) respectively are connected with high probability. Similarly, if \(a_n = 0\) and \(\liminf_n b_n > 2\), the posterior succeeds in exact recovery: apparently, with \(b_n\) above 2, edges between classes are abundant enough to guarantee the existence of a path in \(X^n\) that visits all vertices at least once, with high probability.

It is tempting to state the following, well-known [2, 178] sufficient condition for the sequences \(a_n > 0\) and \(b_n > 0\):

\[(\sqrt{a_n} - \sqrt{b_n})^2 > c, \text{for some } c > 2 \text{ and } n \text{ large enough,}\]  

(10.15)

(even though it ignores the logarithm in (10.14)).

**Corollary 10.1.** Under the conditions of (10.1), the MAP-ML-estimator \(\hat{\theta}_n\) recovers \(\theta_0, n\) exactly: \(P_{\theta_0, n}(\hat{\theta}_n(X^n) = \theta_0, n) \to 1.\)

**Proof.** Due to the uniformity of the prior, for every \(n \geq 1\), maximization of the posterior density (with respect to the counting measure) on \(\Theta_n\), is the same as maximization of the likelihood. Due to (10.12), the posterior density in the point \(\theta_0, n\) in \(\Theta_n\) converges to one in \(P_{\theta_0, n}\)-probability. Accordingly, the point of maximization is \(\theta_0, n\) with high probability.

### 10.4.2 Posterior consistency: detection

For the case of detection, the requirement of convergence is less stringent: we require only that estimated \(\theta_n\) differ from the true \(\theta_0, n\) by \(o(n)\) differences, rather than matching \(\theta_0, n\) exactly. It is well-known that edge-probabilities may be of orders smaller than \(o(\log(n)/n)\) and of order \(O(n^{-1})\). More precisely, *c.f.* [178, proposition 2.9] we have that

\[\frac{n(p_n - q_n)^2}{p_n + q_n} \to \infty,\]  

(10.16)

is a necessary and sufficient condition for detection. So where exact recovery can be achieved only in the Chernoff-Hellinger phase, detection of communities is possible in the Kersten-Stigum phase as well. In this section, we show that the posterior for the uniform prior also ‘detects’ the true class assignment (when that notion is translated appropriately into a property of the posterior, see also [177]).
We are interested in the expected posterior masses of subsets of $\Theta_n$ of the form:

$$W_n = \bigcup_{k=k_n}^{\lfloor n/2 \rfloor} V_{n,k},$$

for a (possibly) divergent sequence $k_n$ of order $o(n)$: the posterior concentrates on class assignments $\theta_n$ that differ from $\theta_{0,n}$ by no more than $k_n$ pair exchanges, so the fraction of misclassified vertices becomes negligible in the limit $n \to \infty$.

**Theorem 10.2.** For some $\theta_{0,n} \in \Theta_n$, let $X_n \sim P_{\theta_{0,n}}$ for every $n \geq 1$. If we equip all $\Theta_n$ with uniform priors and $(p_n), (q_n)$ are such that,

$$\frac{n}{k_n} \left(1 - p_n - q_n + 2p_n q_n + 2\sqrt{(p_n(1-p_n)q_n(1-q_n))} \right)^{n/2} \to 0,$$  \hspace{1cm} (10.17)

as $n \to \infty$, then,

$$\Pi(W_n|X^n) \overset{R_0}{\to} 0,$$  \hspace{1cm} (10.18)

as $n \to \infty$, i.e. the posterior detects $\theta_{0,n}$ at rate $k_n$.

**Proof.** (of theorem 10.2) According to lemma 10.1, for every $n \geq 1$, $k \geq 1$ and given, $\theta_{0,n}$, there exists a test sequence satisfying (10.10) with $a_{n,k} = (1-\mu_n)^{2k(n-k)}$. Therefore, using the inequalities $\binom{2n}{k} \leq \frac{(2n)^k}{k!}$ and $(n+m)! \geq n!m!$, the Stirling lower bound formula, and finally our assumption $n(1-\mu_n)^{n/2}/k_n \to 0$, we see that for big enough $n$,

$$P_{\theta_{0,n}} \Pi(W_n|X^n) \leq \sum_{k=k_n}^{\lfloor n/2 \rfloor} \binom{n}{k}^2 (1-\mu_n)^{2k(n-k)}$$

$$\leq \sum_{k=2k_n}^{n} \binom{2n}{k} (1-\mu_n)^{k(n-k/2)} \leq \sum_{k=2k_n}^{n} \frac{1}{k!} (2n)^k (1-\mu_n)^{kn/2}$$

$$\leq \frac{(2n(1-\mu_n)^{n/2})^{2k_n}}{(2k_n)!} e^{2n(1-\mu_n)^{n/2}}.$$

Application of Stirling’s approximation then leads to,

$$P_{\theta_{0,n}} \Pi(W_n|X^n) \leq \frac{1}{\sqrt{4\pi k_n}} \left(\frac{n(1-\mu_n)^{n/2}}{k_n} \right)^{2k_n} e^{2k_n + 2n(1-\mu_n)^{n/2}}$$

$$\leq \frac{1}{\sqrt{4\pi k_n}} \left(\frac{n(1-\mu_n)^{n/2}}{k_n} e^{1+n(1-\mu_n)^{n/2}/k_n} \right)^{2k_n}$$

$$\leq \frac{n(1-\mu_n)^{n/2}}{k_n} e^{1+n(1-\mu_n)^{n/2}/k_n}.$$
which converges to zero as $n \to \infty$.

Example 10.2. Note that as $p_n, q_n \to 0$, we may expand,

$$\sqrt{p_n} - \sqrt{q_n} = \frac{1}{2\sqrt{\frac{1}{2}(p_n + q_n)}} (p_n - q_n) + O(\|p_n - q_n\|^2).$$

which means that,

$$\mu_n = (\sqrt{p_n} - \sqrt{q_n})^2 + O(n^{-2}) = \frac{(p_n - q_n)^2}{2(p_n + q_n)} + O(n^{-2}).$$

Assuming only condition (10.2), we would arrive at the conclusion that $n\mu_n > 1 + O(n^{-1})$, which is insufficient in the proof of theorem 10.2. Note that a non-divergent choice $k_n = O(1)$ forces us back into the Chernoff-Hellinger phase where exact recovery is possible.

Corollary 10.2. Under the conditions of theorem 10.2 with $(p_n)$ and $(q_n)$ such that,

$$n(p_n + q_n - 2p_n q_n - 2\sqrt{p_n(1-p_n)q_n(1-q_n)}) \to \infty,$$

as $n \to \infty$, then, there exists a sequence $k_n = o(n)$ such that,

$$\Pi(k_n(\theta_n, \theta_{0,n}) \geq k_n \mid X^n) \xrightarrow{P} 0,$$

as $n \to \infty$, i.e. the posterior detects $\theta_{0,n}$.

Proof. Define, for every $\beta \in (0, 1)$, $k_{\beta,n} = \beta n$. We follow the proof of theorem 10.2 with $k_n = k_{\beta,n}$ and note that,

$$P_{\theta_{0,n}}\Pi(k_n(\theta_n, \theta_{0,n}) \geq k_{\beta,n} \mid X^n) \leq \frac{1}{\beta} (1 - \mu_n)^{n/2} e^{1 + \beta^{-1} (1 - \mu_n) n/2}.$$

Due to eq. (10.19),

$$(1 - \mu_n)^{n/2} = (1 - p_n - q_n + 2p_n q_n + 2\sqrt{p_n(1-p_n)q_n(1-q_n)})^{n/2} \to 0,$$

so $P_{\theta_{0,n}}\Pi(k_n(\theta_n, \theta_{0,n}) \geq k_{\beta,n} \mid X^n) \to 0$. Let $\beta_n \downarrow 0$ be given; if we let $m(n)$ go to infinity slowly enough, posterior convergence continues to hold with $\beta$ equal to $\beta_{m(n)}$, that is, for $k_n = k_{\beta_{m(n)},n}$.

According to [178, proposition 2.9] (see (10.16)), $n\mu_n \to \infty$ is also a necessary condition for the possibility of detection. So corollary 10.2 shows that uniform priors lead to posteriors that detect the truth under the weakest possible condition on the parameter sequences $(p_n)$ and $(q_n)$. This is encouraging to the Bayesian and to the frequentist who uses Bayesian methods in this model and in models like it, e.g. the stochastic block model.
10.5 Uncertainty quantification

The most immediate results on uncertainty quantification are obtained with the help of the results in the previous section: if we know that the sequences \((p_n)\) and \((q_n)\) satisfy requirements like (10.11) or (10.17), so that recovery or detection is guaranteed, then a consistent sequence of confidence sets is easily constructed from credible sets, as shown in subsection 10.5.1 and the sizes of these credible sets as well as the sizes of associated confidence sets are controlled.

If the sequences \((p_n)\) and \((q_n)\) are unknown, or if we require explicit confidence levels, confidence sets can still be constructed from credible sets under conditions requiring that credible levels grow to one quickly enough. Enlargement of credible sets may be used to mitigate this condition, whenever we are close to the Erdős-Rényi submodel, as discussed in subsection 10.5.2.

Regarding the sizes of credible sets, the most natural way to compile a minimal-order credible set \(E_n(X^n)\) in a discrete space like \(\Theta_n\), is to calculate the posterior weights \(\Pi(\{\theta_n\} | X^n)\) of all \(\theta_n \in \Theta_n\), order \(\Theta_n\) by decreasing posterior weight into a finite sequence \(\{\theta_{n,1}, \theta_{n,2}, \ldots, \theta_{n,m}\}\) and define \(E_n(X^n) = \{\theta_{n,1}, \ldots, \theta_{n,m}\}\), for the smallest \(m \geq 1\) such that \(\Pi(\{\theta_{n,1}, \ldots, \theta_{n,m}\} | X^n)\) is greater than or equal to the required credible level. To provide guarantees regarding the sizes of credible sets, one would like to show that these \(E_n(X^n)\) are of an order that is upper bounded with high probability. (Although it is not so clear what the upper bound should be, ideally.)

Here we shall follow a different path based on the smallest number \(k(\theta_n, \eta_n)\) of pair-exchanges between two representations \(\theta'_n\) and \(\eta'_n\) in \(\Theta'_n\) of \(\theta_n\) and \(\eta_n\) respectively, see (10.7). The map \(k : \Theta_n \times \Theta_n \rightarrow \{0, 1, \ldots, \lfloor n/2 \rfloor\}\) is interpreted in a role similar to that of a metric on larger parameter spaces: the diameter \(\text{diam}_n(C)\) of a subset \(C \subset \Theta_n\) is,

\[
\text{diam}_n(C) = \max \left\{ k(\theta_n, \eta_n) : \theta_n, \eta_n \in C \right\}.
\]

by definition.

10.5.1 Posterior recovery/detection and confidence sets

If the posteriors concentrate amounts of mass on \(\{\theta_{0,n}\}\) arbitrarily close to one with growing \(n\), then a sequence of credible sets of a certain, fixed level contains \(\theta_{0,n}\) for large enough \(n\). If such posterior concentration occurs with high \(P_{\theta_{0,n}}\)-probability, then the sequence of credible sets is also an asymptotically consistent sequence of confidence sets.

**Theorem 10.3.** Let \(c_n \in [0, 1]\) be given, with \(c_n > \varepsilon > 0\) for large enough \(n\). Suppose that the posterior recovers the communities exactly.
Then any sequence $(D_n)$ of $(P^\Pi_n$-almost-sure) credible sets of levels $c_n$ satisfies,

$$P_{\theta_0,n}(\theta_0,n \in D_n(X^n)) \rightarrow 1,$$

i.e. $(D_n)$ is a consistent sequence of confidence sets. Credible sets of minimal order/diameter equal $\{\theta_0\}$ with high $P_{\theta_0,n}$-probability.

**Proof.** Note that with the uniform priors $\Pi_n$, $P_{\theta_0,n} \ll P^\Pi_n$ for all $n \geq 1$, so that $P^\Pi_n$-almost-surely defined credible sets $D_n$ of credible level at least $\varepsilon$, also satisfy,

$$P_{\theta_0,n}(\Pi(D_n(X^n)|X^n) \geq \varepsilon) = 1.$$

So if, in addition,

$$P_{\theta_0,n}(\Pi(|\theta_0,n|)|X^n) > 1 - \varepsilon \rightarrow 1,$$

then $\theta_0,n \in D_n(X^n)$ with high $P_{\theta_0,n}$-probability. Since all posterior mass is concentrated at $\theta_0,n$ with high probability, the $\{\theta_0,n\}$ form a sequence of unique credible sets of minimal order (or minimal diameter $k_0 = 0$) with confidence levels greater than $\varepsilon > 0$ for large enough $n$.

In the Kesten-Stigum phase, enlargement of credible sets is sufficient to obtain confidence sets. Recall the definition of the $V_{n,k}(\theta_0)$ in (10.8) (with $\theta_0$ replaced by $\theta_0,n$). Given some fixed underlying $\theta_0,n \in \Theta_n$, we write $V_{n,k}(\theta_0,n)$ for $V_{n,k}(\theta_0,n)$. Making a certain choice for the upper bounds $k_0 \geq 1$, we arrive at,

$$B_n(\theta_0,n) = \bigcup_{k=0}^{k_0} V_{n,k}(\theta_0,n),$$

(10.22)

for every $n \geq 1$ and $\theta_0 \in \Theta_n$. Similar as for $V_{n,k}$ we write $B_n$ for $B_n(\theta_0,n)$. Given a subset $D_n$ of $\Theta_n$, the set $C_n \subset \Theta_n$ associated with $D_n$ under $B_n(\theta_0,n)$ (see definition 7.7) then is the set of $\theta_n \in \Theta_n$ whose $k$-distance from some element of $D_n$ is at most $k_n$,

$$C_n = \{\theta_n \in \Theta_n : \exists \eta_n \in D_n, k(\eta_n, \theta_n) \leq k_n\},$$

the $k_n$-enlargement of $D_n$. If we know that the sequences $(p_n)$ and $(q_n)$ satisfy requirement (10.17), posterior concentration occurs around $\{\theta_0,n\}$ in ‘balls’ of diameters $2k_n$ with growing $n$, and there exist credible sets $D_n'$ of levels greater than $1/2$ and of diameters $2k_n$ centred on $\theta_0,n$. The credible sets $D_n$ of minimal diameters of any level greater than $1/2$ must intersect $D_n$. Then the $k_n$-enlargements $C_n$ of the $D_n$ contain $\theta_{0,n}$.

**Theorem 10.4.** Let $c_n \in [0,1]$ be given, with $c_n > \varepsilon > 0$ for large enough $n$. Suppose that the posterior detects communities with rate $(k_n)$,

$$\Pi(k(\theta_n, \theta_0,n) \leq k_n | X^n) \xrightarrow{P_{\theta_0,n}} 1.$$
Let \((B_n)\) denote a sequence of \(P_n\)-almost-sure credible sets of levels \(c_n\) of minimal diameters. Then \(\text{diam}_n(B_n) \leq 2k_n\) with high \(P_{\theta_{0,n}}\)-probability and the \(k_n\)-enlargements \(C_n\) of the \(B_n\) satisfy,

\[
P_{\theta_{0,n}}(\theta_{0,n} \in C_n(X^n)) \to 1,
\]
i.e. the \(k_n\)-enlargements \((C_n)\) form a consistent sequence of confidence sets.

Proof. As in the proof of theorem 10.3, \(P_n\)-almost-surely defined credible sets \(D_n\) of credible level at least \(c_n\) also satisfy,

\[
P_{\theta_{0,n}}(\Pi(D_n(X^n)|X^n) \geq c_n) = 1.
\]

Now fix \(n \geq 1\). For every \(\theta_n \in \Theta_n\) and every \(x' \in \Theta_n\), let \(k_n(\theta_n, x')\) denote the radius of the smallest ball in \(\Theta_n\) centred on \(\theta_n\) of posterior mass (at least) \(c_n\). Let \(\hat{\theta}_n(x') \in \Theta_n\) be such that,

\[
k_n(\hat{\theta}_n(x')) = \min \{k_n(\theta_n, x') : \theta_n \in \Theta_n\},
\]
i.e. the centre point of a smallest \(k_n\)-ball in \(\Theta_n\). Convergence of the posterior implies that the balls \(B_n(\theta_{0,n})\) of radii \(k_n\) centred on \(\theta_{0,n}\) contain a fraction of the posterior mass arbitrarily close to one, so assuming that \(n\) is large enough, we may assume that \(c_n > \varepsilon > 0\) and \(\Pi(B_n(\theta_{0,n})|X^n) > 1 - \varepsilon\) with high \(P_{\theta_{0,n}}\)-probability. Conclude that,

\[
B_n(\theta_{0,n}) \cap B_n(\hat{\theta}_n(X^n)) \neq \emptyset,
\]
with high \(P_{\theta_{0,n}}\)-probability, which amounts to asymptotic coverage of \(\theta_{0,n}\) for the \(k_n\)-enlargement \(C_n(X^n)\) of \(B_n(\hat{\theta}_n(X^n))\).

Note that the case \(k_n = 0\) obtains in the Chernoff-Hellinger phase and the cases \(k_n = o(n)\) for the Kesten-Stigum phase. The theorem remains valid in the case \(k_n = \beta n\) (for some \(\beta \in (0, 1)\)), as in the proof of corollary 10.2.

### 10.5.2 Confidence sets directly from credible sets

To use theorems 10.3 or 10.4, the statistician needs to know that the sequences \((p_n)\) and \((q_n)\) satisfy (10.11) or (10.17), basically to satisfy the testing condition (10.10). Particularly, condition (10.19) is not strong enough to apply theorem 10.4. But even if that knowledge is not available and testing cannot serve as a condition, the use of credible sets as confidence sets remains valid, as long as credible levels grow to one fast enough. The following proposition also provides lower bounds for confidence levels of credible sets. (Write \(b_n = |\Theta_n|^{-1} = (\frac{1}{2}(\sqrt{n})^{-1})).

**Proposition 10.1.** Let \(\theta_{0,n}\) in \(\Theta_n\) with uniform priors \(\Pi_n, n \geq 1\), be given and let \(D_n\) be a sequence of credible sets, such that,

\[
\Pi(D_n(X^n)|X^n) \geq 1 - a_n.
\]
for some sequence \((a_n)\) with \(a_n = o(b_n)\). Then,

\[
P_{\theta_0,n}(\theta_0 \in D_n(X^n)) \geq 1 - b_n^{-1} a_n.
\]

**Proof.** If \(\theta_{0,n} \notin D_n(X^n)\) then \(\Pi((\theta_{0,n})|X^n) \leq a_n, P_n^{\Pi}\)-almost surely. Then,

\[
P_{\theta_0,n}(\theta_0 \in \Theta \setminus D_n(X^n)) = P_n^{\Pi}(\theta_0 \in \Theta \setminus D_n(X^n))
\]

\[
= b_n^{-1} \int_{\theta_{0,n}} P_{\theta,n}(\theta_0 \in \Theta \setminus D_n(X^n)) d\Pi_n(\theta)
\]

\[
\leq b_n^{-1} P_n^{\Pi}(\{\theta_0 \in \Theta_n \setminus D_n(X^n)\} \Pi((\theta_{0,n})|X^n)) \leq b_n^{-1} a_n,
\]

by Bayes’s Rule (A.4).

Theorem 7.6 leaves room for mitigation of the lower bound on credible levels if we are willing to use enlarged credible sets. There are two competing influences when enlarging: on the one hand, the prior masses \(b_n = \Pi_n(B_n(\theta_{0,n}))\) become larger, relaxing the rate at which credible levels are required to go to one. On the other hand, enlargement leads to likelihood ratios with random fluctuations that take them further away from one (see lemmas ?? and ??), thus interfering with notions like contiguity and remote contiguity (see appendix ??). Whether proposition 10.1 is useful and whether enlargement of credible sets helps, depends on the sequences \((p_n), (q_n)\).

We shall consider the ‘statistical phase’ where distinctions between within-class and between-class edges become less-and-less pronounced:

\[
p_n - q_n = o(n^{-1}),
\]

while satisfying also the condition that,

\[
p_n^{1/2}(1 - p_n)^{1/2} + q_n^{1/2}(1 - q_n)^{1/2} = o(\{p_n - q_n\}).
\]

In this regime, \(p_n, q_n \to 0\) or \(p_n, q_n \to 1\). If \(p_n, q_n \to 0\) as in the sparse phases, (10.24) amounts to,

\[
n(p_n^{1/2} - q_n^{1/2}) \to \infty,
\]

so differences between \(p_n\) and \(q_n\) may not converge to zero too fast. (Note however that extreme sparsity levels of order \(p_n, q_n \propto n^{-\gamma}\) with \(1 < \gamma < 2\) are allowed.) For the following lemma we define,

\[
\rho_n = \min \left\{ \left( \frac{1 - p_n}{p_n} \right)^{1/2}, \left( \frac{p_n}{1 - p_n} \right)^{1/2} \right\} = e^{-|\lambda_n|}.
\]

where \(\lambda_n := \log(1 - p_n) - \log(p_n) + \log(q_n) - \log(1 - q_n)\), and,

\[
\alpha_n = \int 2k(\theta_{0,n}, \theta_n) (n - k(\theta_{0,n}, \theta_n)) d\Pi_n(\theta_n|B_n) = \frac{1}{|B_n|} \sum_{k=0}^{k_n} \binom{n}{k} 2k(n-k)
\]
with the following rate for remote contiguity (for some $C > 1$):

$$d_n = \rho_n^{C\alpha_n|p_n - q_n|}. \quad (10.26)$$

Let $(k_n)$ and $\theta_{0,n} \in \Theta_n$ be given for all $n \geq 1$. It is shown in lemma ?? that, if (10.24) holds, then,

$$P_{\theta_{0,n}} < d_n^{-1} P_{\theta_{0,n}} \Pi |B_n|,$$

with $B_n = B_n(\theta_{0,n})$ like in (10.22). This argument amounts to a proof for the following theorem (immediate from theorem 7.6).

**Theorem 10.5.** Let $(k_n)$ be given and assume that (10.23) and (10.24) hold. Let $\theta_{0,n}$ in $\Theta_n$ with uniform priors $\Pi_n$ be given and let $D_n$ be a sequence of credible levels $1 - a_n$, for some sequence $(a_n)$ such that $b_n^{-1} a_n = o(d_n)$. Then the sets $C_n$, associated with $D_n$ under $B_n$ as in (10.22) satisfy,

$$P_{\theta_{0,n}} \left( \theta_0 \in C_n(X^n) \right) \to 1,$$

i.e. the $C_n$ are asymptotic confidence sets.

Consider the possible choices for $(a_n)$ if we assume $k_n = \beta n$ for some fixed $\beta \in (0, 1)$ (as in the proof of corollary 10.2), which leads to the type of exponential correction factor in the prior mass sequence $b_n$ that is required to move the restriction on the credible levels $1 - a_n$ substantially. First of all, Stirling’s approximation gives rise to the following approximate lower bound on the factor between prior mass and prior mass without enlargement:

$$\frac{\Pi_n(B_n)}{\Pi_n(\{\theta_{0,n}\})} = \sum_{k=0}^{k_n} \binom{n}{k}^2 \geq \binom{n}{k_n}^2 \geq \frac{1}{2\pi n} \frac{1}{\beta(1-\beta)} f(\beta)^n,$$

where $f : (0, 1) \to (1, 4]$ is given by,

$$f(\beta) = (1 - \beta)^{-2(1-\beta)\beta^{-2}}.$$

Approximating $\alpha_n \approx 2k_n(n - k_n)$ for large $n$ and using (10.23), we also have,

$$d_n = \rho_n^{C\alpha_n|p_n - q_n|} \approx \rho_n^{2C\beta n(1-\beta)|p_n - q_n|} = e^{-\lambda_n \alpha(n)},$$

So if we assume that $\lambda_n = O(1)$, $d_n$ is sub-exponential and does not play a role for the improvement factor.

Conclude as follows: (let $a_n = o(|\Theta_n|^{-1}) \approx o(4^{-n})$ denote the rates appropriate in proposition 10.1 and assume $\lambda_n = O(1)$) if we have credible sets $D_n(X^n)$ of credible levels $1 - a_n f(\beta)^n(1+o(1))$, then the sequence of enlarged confidence sets $(C_n(X^n))$, associated with $D_n(X^n)$ through $B_n$, with $k_n = \beta n$, covers the true value of the class assignment parameter with high probability. Credible levels that had to be of order $1 - a_n \approx 1 - o(4^{-n})$ previously, can be of approximate order $1 - o(c^{-n})$ for any $1 < c < 4$ by enlargement by $B_n$ if conditions (10.23) and (10.24) hold; the closer $c$ is to $1$, the closer $\lambda_n$ is to $1/2$. Theorem 10.5. Let $(k_n)$ be given and assume that (10.23) and (10.24) hold. Let $\theta_{0,n}$ in $\Theta_n$ with uniform priors $\Pi_n$ be given and let $D_n$ be a sequence of credible levels $1 - a_n$, for some sequence $(a_n)$ such that $b_n^{-1} a_n = o(d_n)$. Then the sets $C_n$, associated with $D_n$ under $B_n$ as in (10.22) satisfy,

$$P_{\theta_{0,n}} \left( \theta_0 \in C_n(X^n) \right) \to 1,$$

i.e. the $C_n$ are asymptotic confidence sets. Consider the possible choices for $(a_n)$ if we assume $k_n = \beta n$ for some fixed $\beta \in (0, 1)$ (as in the proof of corollary 10.2), which leads to the type of exponential correction factor in the prior mass sequence $b_n$ that is required to move the restriction on the credible levels $1 - a_n$ substantially. First of all, Stirling’s approximation gives rise to the following approximate lower bound on the factor between prior mass and prior mass without enlargement:

$$\frac{\Pi_n(B_n)}{\Pi_n(\{\theta_{0,n}\})} = \sum_{k=0}^{k_n} \binom{n}{k}^2 \geq \binom{n}{k_n}^2 \geq \frac{1}{2\pi n} \frac{1}{\beta(1-\beta)} f(\beta)^n,$$

where $f : (0, 1) \to (1, 4]$ is given by,

$$f(\beta) = (1 - \beta)^{-2(1-\beta)\beta^{-2}}.$$

Approximating $\alpha_n \approx 2k_n(n - k_n)$ for large $n$ and using (10.23), we also have,

$$d_n = \rho_n^{C\alpha_n|p_n - q_n|} \approx \rho_n^{2C\beta n(1-\beta)|p_n - q_n|} = e^{-\lambda_n \alpha(n)},$$

So if we assume that $\lambda_n = O(1)$, $d_n$ is sub-exponential and does not play a role for the improvement factor.

Conclude as follows: (let $a_n = o(|\Theta_n|^{-1}) \approx o(4^{-n})$ denote the rates appropriate in proposition 10.1 and assume $\lambda_n = O(1)$) if we have credible sets $D_n(X^n)$ of credible levels $1 - a_n f(\beta)^n(1+o(1))$, then the sequence of enlarged confidence sets $(C_n(X^n))$, associated with $D_n(X^n)$ through $B_n$, with $k_n = \beta n$, covers the true value of the class assignment parameter with high probability. Credible levels that had to be of order $1 - a_n \approx 1 - o(4^{-n})$ previously, can be of approximate order $1 - o(c^{-n})$ for any $1 < c < 4$ by enlargement by $B_n$ if conditions (10.23) and (10.24) hold; the closer $c$ is to $1$, the closer $\lambda_n$ is to $1/2$. Theorem 10.5. Let $(k_n)$ be given and assume that (10.23) and (10.24) hold. Let $\theta_{0,n}$ in $\Theta_n$ with uniform priors $\Pi_n$ be given and let $D_n$ be a sequence of credible levels $1 - a_n$, for some sequence $(a_n)$ such that $b_n^{-1} a_n = o(d_n)$. Then the sets $C_n$, associated with $D_n$ under $B_n$ as in (10.22) satisfy,

$$P_{\theta_{0,n}} \left( \theta_0 \in C_n(X^n) \right) \to 1,$$

i.e. the $C_n$ are asymptotic confidence sets. Consider the possible choices for $(a_n)$ if we assume $k_n = \beta n$ for some fixed $\beta \in (0, 1)$ (as in the proof of corollary 10.2), which leads to the type of exponential correction factor in the prior mass sequence $b_n$ that is required to move the restriction on the credible levels $1 - a_n$ substantially. First of all, Stirling’s approximation gives rise to the following approximate lower bound on the factor between prior mass and prior mass without enlargement:

$$\frac{\Pi_n(B_n)}{\Pi_n(\{\theta_{0,n}\})} = \sum_{k=0}^{k_n} \binom{n}{k}^2 \geq \binom{n}{k_n}^2 \geq \frac{1}{2\pi n} \frac{1}{\beta(1-\beta)} f(\beta)^n,$$

where $f : (0, 1) \to (1, 4]$ is given by,

$$f(\beta) = (1 - \beta)^{-2(1-\beta)\beta^{-2}}.$$
10.6 Exercises [EMPTY]
Appendix A  
Notation, definitions and conventions

Because we take the perspective of a frequentist using Bayesian methods, we are obliged to demonstrate that Bayesian definitions continue to make sense under the assumptions that the data \( X \) is distributed according to a true, underlying \( P_0 \).

Remark A.1. We assume given for every \( n \geq 1 \), a measurable (sample) space \((X_n, \mathcal{B}_n)\) and random sample \( X^n \in X_n \), with a model \( \mathcal{P}_n \) of probability distributions \( P_n : \mathcal{B}_n \to [0,1] \). It is also assumed that there exists an \( n \)-independent parameter space \( \Theta \) with a Hausdorff, completely regular topology \( T \) and associated Borel \( \sigma \)-algebra \( \mathcal{G} \), and, for every \( n \geq 1 \), a bijective model parametrization \( \Theta \to \mathcal{P}_n : \theta \mapsto P_{\theta,n} \) such that for every \( n \geq 1 \) and every \( A \in \mathcal{B}_n \), the map \( \Theta \to [0,1] : \theta \mapsto P_{\theta,n}(A) \) is measurable. Any prior \( \Pi \) on \( \Theta \) is assumed to be a Borel probability measure \( \Pi : \mathcal{G} \to [0,1] \) and can vary with the sample-size \( n \). (Note: in i.i.d. setting, the parameter space \( \Theta \) is \( P_1 \), \( \theta \) is the single-observation distribution \( P \) and \( \theta \mapsto P_{\theta,n} \) is \( P \to P^n \).) As frequentists, we assume that there exists a ‘true, underlying distribution for the data; in this case, that means that for every \( n \geq 1 \), there exists a distribution \( P_{0,n} \) from which the \( n \)-th sample \( X^n \) is drawn.

Often one assumes that the model is well-specified: that there exists a \( \theta_0 \in \Theta \) such that \( P_{0,n} = P_{\theta_0,n} \) for all \( n \geq 1 \). We think of \( \Theta \) as a topological space because we want to discuss estimation as a procedure of sequential, stochastic approximation of and convergence to such a ‘true parameter value \( \theta_0 \). In theorem 8.7 and definition 6.4 we assume, in addition, that the observations \( X^n \) are coupled, i.e. there exists a probability space \((\Omega, \mathcal{F}, P_0)\) and random variables \( X^n : \Omega \to X_n \) such that \( P_0((X^n)^{-1}(A)) = P_{\theta_0,n}(X^n \in A) \) for all \( n \geq 1 \) and \( A \in \mathcal{B}_n \).

Definition A.1. Given \( n,m \geq 1 \) and a prior probability measure \( \Pi_n : \mathcal{G} \to [0,1] \), define the \( n \)-th prior predictive distribution on \( X_m \) as follows:

\[
P_{m}^{\Pi_n}(A) = \int_\Theta P_{\theta,m}(A) d\Pi_n(\theta),
\]  \hspace{1cm} (A.1)

for all \( A \in \mathcal{B}_m \). If the prior is replaced by the posterior, the above defines the \( n \)-th posterior predictive distribution on \( X_m \).
\[ P_{m|X^n}^{\Pi_n}(A) = \int_\Theta P_{\theta,m}(A) \, d\Pi_n(X^n), \quad (A.2) \]

for all \( A \in \mathcal{B}_m \). For any \( B_n \in \mathcal{G} \) with \( \Pi_n(B_n) > 0 \), define also the \( n \)-th local prior predictive distribution on \( \mathcal{B}_m \),

\[ P_{m|B_n}^{\Pi_n}(A) = \frac{1}{\Pi_n(B_n)} \int_{B_n} P_{\theta,m}(A) \, d\Pi_n(\theta), \quad (A.3) \]

as the predictive distribution on \( X_m \) that results from the prior \( \Pi_n \) when conditioned on \( B_n \). If \( m \) is not mentioned explicitly, it is assumed equal to \( n \).

The prior predictive distribution \( P_{m|n}^{\Pi_n} \) is the marginal distribution for \( X^n \) in the Bayesian perspective that considers parameter and sample jointly \((\theta, X^n) \in \Theta \times \mathcal{X}_n\) as the random quantity of interest.

**Definition A.2.** Given \( n \geq 1 \), a (version of) the posterior is any map \( \Pi(\cdot|X^n = \cdot) : \mathcal{G} \times \mathcal{X}_n \to [0, 1] \) such that,

1. for \( B \in \mathcal{G} \), the map \( \mathcal{X}_n \to [0, 1] : x_n \mapsto \Pi(B|X^n = x_n) \) is \( \mathcal{B}_n \)-measurable,
2. for all \( A \in \mathcal{B}_n \) and \( V \in \mathcal{G} \),

\[ \int_A \Pi(V|X^n) \, dP_{m|X^n}^{\Pi_n} = \int_V P_{\theta,n}(A) \, d\Pi_n(\theta). \quad (A.4) \]

Bayes’s Rule is expressed through equality (A.4) and is sometimes referred to as a ‘disintegration’ (of the joint distribution of \((\theta, X^n)\)). If the posterior is a Markov kernel, it is a \( P_{m|X^n}^{\Pi_n} \)-almost-surely well-defined probability measure on \( (\Theta, \mathcal{G}) \). But it does not follow from the definition above that a version of the posterior actually exists as a regular conditional probability measure. Under mild extra conditions, regularity of the posterior can be guaranteed: for example, if sample space and parameter space are Polish, the posterior is regular; if the model \( P_n \) is dominated (denote the density of \( P_{\theta,n} \) by \( p_{\theta,n} \)), the fraction of integrated likelihoods,

\[ \Pi(V|X^n) = \int_V p_{\theta,n}(X^n) \, d\Pi_n(\theta) / \int_\Theta p_{\theta,n}(X^n) \, d\Pi_n(\theta), \quad (A.5) \]

for \( V \in \mathcal{G} \), \( n \geq 1 \) defines a regular version of the posterior distribution. (Note also that there is no room in definition (A.4) for \( X^n \)-dependence of the prior, so ‘empirical Bayes’ methods must be based on data \( Y^n \) independent of \( X^n \), i.e. sample-splitting.)

**Remark A.2.** As a consequence of the frequentist assumption that \( X^n \sim P_{0,n} \) for all \( n \geq 1 \), the \( P_{m|X^n}^{\Pi_n} \)-almost-sure definition (A.4) of the posterior \( \Pi(V|X^n) \) does not make sense automatically [92, 145]: null-sets of \( P_{m|X^n}^{\Pi_n} \) on which the definition of \( \Pi(\cdot|X^n) \) is ill-determined, may not be null-sets of \( P_{0,n} \). To prevent this, we impose the domination condition,

\[ P_{0,n} \ll P_{m|X^n}^{\Pi_n} \quad (A.6) \]

for every \( n \geq 1 \).
To understand the reason for (A.6) in a perhaps more familiar way, consider a dominated model and assume that for certain \( n \), (A.6) is not satisfied. Then, using (A.1), we find,

\[
P_{0,n}\left( \int p_{\theta,n}(X^n) \, d\Pi_n(\theta) = 0 \right) > 0,
\]

so the denominator in (A.5) evaluates to zero with non-zero \( P_{0,n} \)-probability. A sufficient condition for (A.6) is obtained with the help of the topologies \( \mathcal{T}_n \) (see also remark 3.6 (2) in Strasser (1985) [208]).

**Definition A.3.** For all \( n \geq 1 \), let \( F_n \) denote the class of all bounded, \( \mathcal{B}_n \)-measurable \( f : \mathcal{X}_n \to \mathbb{R} \). The topology \( \mathcal{T}_n \) is the initial topology on \( \mathcal{P}_n \) for the functions \( \{ P \mapsto Pf : f \in F_n \} \).

If we model single-observation distributions \( P \in \mathcal{P} \) for an i.i.d. sample, the topology \( \mathcal{T}_n \) on \( \mathcal{P}_n = \mathcal{P}^n \) induces a topology on \( \mathcal{P} \) (which we also denote by \( \mathcal{T}_n \)) for each \( n \geq 1 \). The union \( \mathcal{T}_n = \bigcup_n \mathcal{T}_n \) is an inverse-limit topology that allows formulation of conditions for the existence of consistent estimates that are not only sufficient, but also necessary [156], offering a precise perspective on what is estimable and what is not in i.i.d. context. The associated strong topology is that generated by total variation (or, equivalently, the Hellinger metric).

For more on these topologies, the reader is referred to Strasser (1985) [208] and to Le Cam (1986) [162]. We note explicitly the following fact, which is a direct consequence of Hoeffding’s inequality.

**Proposition A.1.** (Uniform \( \mathcal{T}_n \)-tests)
Consider a model \( \mathcal{P} \) of single-observation distributions \( P \) for i.i.d. samples \( (X_1, X_2, \ldots, X_n) \sim P^n \), \( (n \geq 1) \). Let \( m \geq 1 \), \( \varepsilon > 0 \), \( P_0 \in \mathcal{P} \) and a measurable \( f : \mathcal{X}^m \to [0, 1] \) be given. Define \( B = \{ P \in \mathcal{P} : \sup \{ |(P^n - P_0^n) f| < \varepsilon \} \} \), and \( V = \{ P \in \mathcal{P} : \sup \{ |(P^n - P_0^n) f| \geq 2\varepsilon \} \} \).

There exist a uniform test sequence \( (\phi_n) \) such that,

\[
\sup_{P \in B} P^n \phi_n \leq e^{-nD}, \quad \sup_{Q \in V} Q^n (1 - \phi_n) \leq e^{-nD},
\]

for some \( D > 0 \).

**Proof.** The proof is an application of Hoeffding’s inequality for the sum \( \sum_{i=1}^{n} f(X_i) \) and is left to the reader.

The topologies \( \mathcal{T}_n \) also play a role for condition (A.6).

**Proposition A.2.** Let \( (\Pi_n) \) be Borel priors on the Hausdorff uniform spaces \( (\mathcal{P}_n, \mathcal{T}_n) \).
For any \( n \geq 1 \), if \( R_{0,n} \) lies in the \( \mathcal{T}_n \)-support of \( \Pi_n \), then \( R_{0,n} \ll \Pi_n \).

**Proof.** Let \( n \geq 1 \) be given. For any \( A \in \mathcal{P}_n \) and any \( U' \subset \Theta \) such that \( \Pi_n(U') > 0 \),

\[
P_{0,n}(A) \leq \int P_{\theta,n}(A) \, d\Pi_n(\theta | U') + \sup_{\theta \in U'} | P_{\theta,n}(A) - R_{0,n}(A) |.
\]

Let \( A \in \mathcal{B}_n \) be a null-set of \( \Pi_n U' \); since \( \Pi_n(U') > 0 \), \( \int P_{\theta,n}(A) \, d\Pi_n(\theta | U') = 0 \). For some \( \varepsilon > 0 \), take \( U' \) equal to the \( \mathcal{T}_n \)-basis element \( \{ \theta \in \Theta : | P_{\theta,n}(A) - R_{\theta,n}(A) | < \varepsilon \} \) to conclude that \( P_{\theta,n}(A) < \varepsilon \) for all \( \varepsilon > 0 \).
In many situations, priors are Borel for the Hellinger topology, so it is useful to observe that the Hellinger support of $\Pi_n$ in $\mathcal{P}_n$ is always contained in the $\mathcal{R}_n$-support.

**Notation and conventions**

*l.h.s.* and *r.h.s.* refer to left- and right-hand sides respectively. For given probability measures $P, Q$ on a measurable space $(\Omega, \mathcal{F})$, we define the Radon-Nikodym derivative $dP/dQ : \Omega \to [0, \infty)$, $P$-almost-surely, referring only to the $Q$-dominated component of $P$, following [162]. We also define $(dP/dQ)^{-1} : \Omega \to (0, \infty]$ : $\omega \mapsto 1/(dP/dQ(\omega))$, $Q$-almost-surely. Given a $\sigma$-finite measure $\mu$ that dominates both $P$ and $Q$ (e.g. $\mu = P + Q$), denote $dP/d\mu = q$ and $dQ/d\mu = p$. Then the measurable map $p/q 1\{q > 0\} : \Omega \to [0, \infty)$ is a $\mu$-almost-everywhere version of $dP/dQ$, and $q/p 1\{q > 0\} : \Omega \to [0, \infty)$ of $(dP/dQ)^{-1}$. Define total-variational and Hellinger distances by $\|P - Q\| = \sup_A |P(A) - Q(A)|$ and $H(P, Q)^2 = 1/2 \int (p^{1/2} - q^{1/2})^2 d\mu$, respectively. Given random variables $Z_n \sim P_n$, weak convergence to a random variable $Z$ is denoted by $Z_n \xrightarrow{P\text{-w.}} Z$, convergence in probability by $Z_n \xrightarrow{P\text{-a.s.}} Z$ and almost-sure convergence (with coupling $P^\omega$) by $Z_n \xrightarrow{P^\omega\text{-a.s.}} Z$. The integral of a real-valued, integrable random variable $X$ with respect to a probability measure $P$ is denoted $PX$, while integrals over the model with respect to priors and posteriors are always written out in Leibniz’s notation. For any subset $B$ of a topological space, $\overline{B}$ denotes the closure, $\mathring{B}$ the interior and $\partial B$ the boundary. Given $\varepsilon > 0$ and a metric space $(\Theta, d)$, the covering number $N(\varepsilon, \Theta, d) \in \mathbb{N} \cup \{\infty\}$ is the minimal cardinal of a cover of $\Theta$ by $d$-balls of radius $\varepsilon$. Given real-valued random variables $X_1, \ldots, X_n$, the first order statistic is $X_{(1)} = \min_{1 \leq i \leq n} X_i$. The Hellinger diameter of a model subset $C$ is denoted $\text{diam}_H(C)$ and the Euclidean norm of a vector $\theta \in \mathbb{R}^n$ is denoted $\|\theta\|_{2,n}$. 
Appendix B
Measure theory

In this appendix we collect some important notions from measure theory. The goal is not a self-contained presentation but rather to establish the basic definitions and theorems from the theory for reference in the main text. As such, the presentation omits certain existence theorems and many of the proofs of other theorems (although references are given). The focus is strongly on finite (e.g. probability-)measures, in places at the expense of generality. Some background in elementary set-theory and analysis is required. As a comprehensive reference we note Kingman and Taylor (1966) [138], alternatives being Dudley (1989) [76] and Billingsley (1986) [31].

B.1 Sets and sigma-algebras

It is assumed that the reader is familiar with the following notions in set theory: set, subset, empty set, union, intersection, complement, set difference and disjointness. Let \( \Omega \) be a set. The powerset \( 2^\Omega \) is the collection of all subsets of \( \Omega \). A partition of \( \Omega \) is an \( \mathcal{A} \subset 2^\Omega \) such that \( \Omega = \bigcup_{A \in \mathcal{A}} A \) and \( A \cap A' = \emptyset \) for any \( A, A' \in \mathcal{A} \) such that \( A \neq A' \). Let \( (A_n) \) be a sequence of subsets of \( \Omega \). We say that \( (A_n) \) is monotone decreasing (resp. monotone increasing) if \( A_{n+1} \subset A_n \) (resp. \( A_n \subset A_{n+1} \)) for all \( n \geq 1 \). A monotone decreasing (resp. increasing) sequence \( (A_n) \) has a set-theoretic limit \( \lim A_n \) defined as \( \cap_{n \geq 1} A_n \) (resp. \( \cup_{n \geq 1} A_n \)). For any sequence of subsets \( (A_n) \), the sequence \( (\bigcup_{m \geq n} A_m)_{n \geq 1} \) (resp. \( (\bigcap_{m \geq n} A_m)_{n \geq 1} \)) is monotone decreasing (resp. increasing) and, accordingly, for any sequence \( (A_n) \) we define

\[
\limsup A_n = \cap_{n \geq 1} \bigcup_{m \geq n} A_m, \quad \liminf A_n = \bigcup_{n \geq 1} \cap_{m \geq n} A_m.
\]

The sequence \( (A_n) \) is said to converge, if \( \limsup A_n = \liminf A_n \).

**Definition B.1.** Let \( \Omega \) be a set. A collection \( \mathcal{F} \) of subsets of \( \Omega \) is called a \( \sigma \)-algebra, if \( \mathcal{F} \) has the following properties.

1. \( \emptyset \in \mathcal{F} \),
2. If $A \in \mathcal{F}$, then $\Omega \setminus A \in \mathcal{F}$.
3. If $(A_n) \subset \mathcal{F}$, then $\bigcup_{n \geq 1} A_n \in \mathcal{F}$.

**Definition B.2.** A measurable space $(\Omega, \mathcal{F})$ consists of a set $\Omega$ and a $\sigma$-algebra $\mathcal{F}$ of subsets of $\Omega$.

A subset $A$ of a measurable space $(\Omega, \mathcal{F})$ is called measurable if $A \in \mathcal{F}$. It can be shown that a $\sigma$-algebra is a monotone class which means that if $(A_n) \subset \mathcal{F}$ is a monotone sequence, then $\lim A_n \in \mathcal{F}$.

**Definition B.3.** Let $\Omega$ be a non-empty set and let $\mathcal{C}$ be a collection of subsets of $\Omega$. The $\sigma$-algebra generated by $\mathcal{C}$, denoted $\sigma(\mathcal{C})$ is the smallest $\sigma$-algebra that contains $\mathcal{C}$.

**Lemma B.1.** Let $\Omega$ be a non-empty set and let $\mathcal{C}$ be a collection of subsets of $\Omega$. Then,

$$\sigma(\mathcal{C}) = \bigcap \{ \Sigma \subset 2^\Omega : \mathcal{C} \subset \Sigma, \Sigma \text{ is a } \sigma\text{-algebra} \}$$

**Example B.1.** Let $X$ be a topological space. The Borel $\sigma$-algebra is the $\sigma$-algebra $\sigma(\mathcal{F})$ generated by the open (or closed) sets. The Borel $\sigma$-algebra on $X$ is denoted $\mathcal{B}(X)$.

### B.2 Measures

From here on, let $(\Omega, \mathcal{F})$ denote a measurable space. A set-function $\nu$ is any mapping $\mathcal{F} \to \mathbb{R}$.

**Definition B.4.** A set-function $\nu : \mathcal{F} \to \mathbb{R}$ is said to be additive if, for any $k \geq 1$ and any $\mathcal{A} = \{A_1, \ldots, A_k\} \subset \mathcal{F}$ such that $A_i \cap A_j = \emptyset$ for all $1 \leq i < j \leq k$,

$$\nu \left( \bigcup_{i=1}^k A_i \right) = \sum_{i=1}^k \nu(A_i).$$

A set-function $\nu$ is said to be countably additive (or $\sigma$-additive) if the above holds for any countable $\mathcal{A} = \{A_n : n \geq 1\} \subset \mathcal{F}$ such that $A_i \cap A_j = \emptyset$ for all $i, j \geq 1, i \neq j$.

**Definition B.5.** Given a measurable space $(\Omega, \mathcal{F})$, a set-function $\mu : \mathcal{F} \to \mathbb{R}$ is a signed measure if $\mu$ is countably additive and $\mu$ is a measure if $\mu$ is countably additive and $\mu \geq 0$. A measure with a Borel $\sigma$-algebra for a domain is called a Borel measure. If $\mu$ is a measure, $(\Omega, \mathcal{F}, \mu)$ is called a measure space. If $\mu$ is finite (resp. $\sigma$-finite), $(\Omega, \mathcal{F}, \mu)$ is called a finite (resp. $\sigma$-finite) measure space.

**Definition B.6.** Let $(\mathcal{V}, \mathcal{B})$ be a measurable space. Given a set-function $\nu : \mathcal{B} \to \mathbb{R}$, the total-variation norm of $\nu$ is defined:

$$\|\nu\| = \sup_{\mathcal{A}} \sum_{A \in \mathcal{A}} |\nu(A)|,$$

(B.1)
where the supremum is taken over all countable, measurable partitions \( \mathcal{A} \). Alternatively, we may decompose \( \nu = \nu_+ - \nu_- \) uniquely into two positive measures \( \nu_+ , \nu_- \) (the so-called Hahn-Jordan decomposition), define the total variation measure \( |\nu| = \nu_+ + \nu_- \) and note that \( \|\nu\| = |\nu|(\mathcal{Y}) \). A set-function \( \nu \) is said to be finite if its total variation is finite. A \( \sigma \)-additive set-function \( \nu : \mathcal{F} \to \mathbb{R} \) is said to be \( \sigma \)-finite if there exists a measurable countable partition \( \{A_n\} \) of \( \Omega \) such that \( |\nu|(A_n) < \infty \) for all \( n \geq 1 \). A positive measure \( \nu \) such that \( \|\nu\| = \nu(\mathcal{Y}) = 1 \) is a probability measure. Then \( (\mathcal{Y} , \mathcal{B} , \nu) \) is called a probability space.

A null-set of a measure \( \mu \) on \( (\mathcal{Y} , \mathcal{B}) \) is an \( A \in \mathcal{B} \) such that \( \mu(A) = 0 \). If a property holds for all points in \( \mathcal{Y} \), except in a null-set \( A \subset \mathcal{Y} \) of a measure \( \mu \), we say that the property holds \( (\mu\text{-})\text{almost-everywhere} \) (notation: \( \mu\text{-a.e.} \)) or, if \( \mu \) is a probability measure, \( (\mu\text{-})\text{almost-surely} \) (notation: \( \mu\text{-a.s.} \)). For any two positive measures \( \mu \) and \( \nu \) on \( (\mathcal{Y} , \mathcal{B}) \), we say that \( \mu \text{ dominates} \nu \) (notation: \( \nu \ll \mu \)), if \( \mu(A) = 0 \) implies \( \nu(A) = 0 \) for all \( A \in \mathcal{B} \). We say that \( \mu \) and \( \nu \) are (mutually) singular (notation: \( \mu \perp \nu \)), if there exists a measurable partition \( \{A,B\} \) of \( \mathcal{Y} \) such that \( \mu(C) = 0 \) for any measurable \( C \subset B \) and \( \nu(D) = 0 \) for any measurable \( D \subset A \).

**Proposition B.1.** Let \((\Omega , \mathcal{F})\) be a measurable space. The collection of all finite signed measures on \( \mathcal{F} \) forms a linear space \( \mathcal{M}(\Omega , \mathcal{F}) \) which is a Banach space for the total variation norm.

As a result of \( \sigma \)-additivity, measures display a form of continuity expressed by the following theorem.

**Theorem B.1.** Let \((\Omega , \mathcal{F})\) be a measurable space with measure \( \mu : \mathcal{F} \to [0,\infty] \). Then,

(i) for any monotone decreasing sequence \((F_n)_{n \geq 1}\) in \( \mathcal{F} \) such that \( \mu(F_n) < \infty \) for some \( n \),

\[
\lim_{n \to \infty} \mu(F_n) = \mu\left(\bigcap_{n=1}^{\infty} F_n\right),
\]

(B.2)

(ii) for any monotone increasing sequence \((G_n)_{n \geq 1}\) in \( \mathcal{F} \),

\[
\lim_{n \to \infty} \mu(G_n) = \mu\left(\bigcup_{n=1}^{\infty} G_n\right),
\]

(B.3)

Theorem B.1 is sometimes referred to as the continuity theorem for measures, because if we view \( \bigcap_n F_n \) as the monotone limit \( \lim_n F_n \), (B.2) can be read as \( \lim_n \mu(F_n) = \mu(\lim_n F_n) \), expressing continuity from below. Similarly, (B.3) expresses continuity from above. Note that theorem B.1 does not guarantee continuity for arbitrary sequences in \( \mathcal{F} \). It should also be noted that theorem B.1 is presented here in simplified form: the full theorem states that continuity from below is equivalent to \( \sigma \)-additivity of \( \mu \) (for a more comprehensive formulation and a proof of theorem B.1, see [138], theorem 3.2).

**Example B.2.** Let \( \Omega \) be a discrete set and let \( \mathcal{F} \) be the powerset \( 2^\Omega \) of \( \Omega \), i.e. \( \mathcal{F} \) is the collection of all subsets of \( \Omega \). The counting measure \( n : \mathcal{F} \to [0,\infty] \) on \((\Omega , \mathcal{F})\)
is defined simply to count the number \( n(F) \) of points in \( F \subseteq \Omega \). If \( \Omega \) contains a finite number of points, \( n \) is a finite measure; if \( \Omega \) contains a countably infinite number of points, \( n \) is \( \sigma \)-finite. The counting measure is \( \sigma \)-additive.

**Example B.3.** We consider \( \mathbb{R} \) with any \( \sigma \)-algebra \( \mathcal{F} \), let \( x \in \mathbb{R} \) be given and define the measure \( \delta_x : \mathcal{F} \to [0,1] \) by

\[
\delta_x(A) = 1\{x \in A\},
\]

for any \( A \in \mathcal{F} \). The probability measure \( \delta_x \) is called the Dirac measure (or delta measure, or atomic measure) degenerate at \( x \) and it concentrates all its mass in the point \( x \). Clearly, \( \delta_x \) is finite and \( \sigma \)-additive. Convex combinations of Dirac measures, i.e. measures of the form

\[
P = \sum_{j=1}^{m} \alpha_j \delta_{\alpha_j},
\]

for some \( m \geq 1 \) with \( \alpha_1, \ldots, \alpha_m \) such that \( \alpha_j \geq 0 \) and \( \sum_{j=1}^{m} \alpha_j = 1 \), can be used as a statistical model for an observation \( X \) that take values in a discrete (but unknown) subset \( \{x_1, \ldots, x_m\} \) of \( \mathbb{R} \). The resulting model is not dominated. For later reference, we introduce the set of all discrete measures \( D(\mathbb{R}, \mathcal{B}) = \{P \in M(\mathbb{R}, \mathcal{B}) : P = \sum_{\alpha_j} \alpha_j \delta_{\alpha_j} \} \) for sequences \( (x_j) \subseteq \mathbb{R} \) and \( (\alpha_j) \subseteq [0,1] \) such that \( \sum_{j=1}^{m} \alpha_j = 1 \).

Often, one has a sequence of events \( (A_n) \) and one is interested in the probability of a limiting event \( A \), for example the event that \( A_n \) occurs infinitely often. The following lemmas pertain to this situation.

**Lemma B.2.** (First Borel-Cantelli lemma)
Let \( (\Omega, \mathcal{F}, P) \) be a probability space with a sequence \( (A_n) \subseteq \mathcal{F} \) and denote \( A = \limsup A_n \). If \( \sum_{n \geq 1} P(A_n) < \infty \), then \( P(A) = 0 \).

In the above lemma, the sequence \( (A_n) \) is general. To draw the converse conclusion, the sequence needs to exist of independent events: \( A, B \in \mathcal{F} \) are said to be independent under \( P \) if \( P(A \cap B) = P(A)P(B) \).

**Lemma B.3.** (Second Borel-Cantelli lemma)
Let \( (\Omega, \mathcal{F}, P) \) be a probability space and let \( (A_n) \subseteq \mathcal{F} \) be independent and denote \( A = \limsup A_n \). If

\[
\sum_{n \geq 1} P(A_n) = \infty,
\]

then \( P(A) = 1 \).

Together, the Borel-Cantelli lemmas assert that for a sequence of independent events \( (A_n) \), \( P(A) \) equals zero or one, according as \( \sum_{n \geq 1} P(A_n) \) converges or diverges. As such, this corollary is known as a zero-one law.

To conclude this section, we consider a property of random vectors called exchangeability.

**Definition B.7.** A random vector \( (X_1, \ldots, X_n) \in \mathbb{R}^n \) with distribution \( P_n \) is said to be exchangeable, if, for any permutation \( \pi \) of \( \{1, \ldots, n\} \), the random vector \( (X_{\pi(1)}, \ldots, X_{\pi(n)}) \) also has distribution \( P_n \).
This property is a generalization of \(i.i.d\).-ness: note that if \((X_1, \ldots, X_n) \sim P^n_0\) then \((X_1, \ldots, X_n)\) is exchangeable. The converse does not hold but exchangeable distributions can be characterized in terms of \(i.i.d\). distributions, as the following result demonstrates.

**Theorem B.2. (De Finetti’s theorem)** The random vector \((X_1, \ldots, X_n) \in \mathbb{R}^n\) distributed according to a probability measure \(P_n\) is exchangeable if and only if there exists a (unique) probability measure \(\Pi\) on the collection \(\mathcal{M}_1^+ (\mathbb{R})\) of all Borel probability on \(\mathbb{R}\) such that,

\[
P_n(A_1 \times \ldots \times A_n) = \int_{\mathcal{B}(\mathbb{R})} \prod_{i=1}^n P(A_i) d\Pi(P),
\]

for all \(A_1, \ldots, A_n \in \mathcal{B}(\mathbb{R})\).

### B.3 Measurability, random variables and integration

In this section we consider random variables and their expectation values. Throughout this section, let \((\Omega, \mathcal{F}, P)\) denote a probability space.

**Definition B.8.** Given a map \(X : A \to B\) and a subset \(C \subset B\), the pre-image of \(C\) under \(X\), is defined as,

\[
X^{-1}(C) = \{a \in A : X(a) \in C\} \subset A.
\]

Given two measurable spaces \((\Omega, \mathcal{F})\) and \((\mathcal{X}, \mathcal{B})\), a map \(X : \Omega \to \mathcal{X}\) is called measurable if, for all \(B \in \mathcal{B}, X^{-1}(B) \in \mathcal{F}\). These subsets form a sub-\(\sigma\)-algebra \(\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}\}\) called the \(\sigma\)-algebra generated by \(X\).

Essentially, measurability makes it possible to speak of “the probability that \(X\) lies in \(B\)”: \(P(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\})\),

is well-defined only if \(X^{-1}(B)\) belongs to the domain of \(P\). Specializing to real-valued measurable maps, it follows from elementary manipulation of set-limits that suprema of sequences of measurable maps are again measurable. This statement can be framed in the following central theorem in measure theory.

**Theorem B.3. (Monotone class theorem)** For every \(n \geq 1\), let \(f_n : \Omega \to \mathbb{R}\) be measurable and assume that \(f_{n+1}(\omega) \geq f_n(\omega)\) for all \(n \geq 1\) and \(\omega \in \Omega\). Then \(f(\omega) = \lim_{n \to \infty} f_n(\omega)\) defines a measurable map \(f : \Omega \to \mathbb{R}\).

This means that the set of all measurable \(f : \Omega \to \mathbb{R}\) forms what is know as a **monotone class**, an partially ordered set that is closed for limits over monotone sequences. Although measurability is preserved under linear combinations, the space of all measurable \(f : \Omega \to \mathbb{R}\) is not a linear space because if, for some \(\omega \in \Omega\),...
$f(\omega) = \infty$ and $g(\omega) = -\infty$, then $(f + g)(\omega) = \infty - \infty$ is ill-defined. No such problems arise when we restrict to the set of all measurable $f \geq 0$, which form a cone. Restriction to measurable $f : \Omega \to \mathbb{R}$, on the other hand, invalidates the monotone class theorem.

**Definition B.9.** Let $(\Omega, \mathcal{F}, P)$ be a probability space. A *random variable* is a measurable map $X : \Omega \to \mathbb{R}$ with the property that $P(|X| = \infty) = 0$. Therefore, every random variable can be represented by a real-valued $X' : \Omega \to \mathbb{R}$, up to null-sets of $P$, i.e. $P(X = X') = 1$.

Note that random variables do not form a monotone class (take $f_n = n$), but they do form a linear space. To define expectations (integrals with respect to $P$), we extend by monotone limit starting from the following definition.

**Definition B.10.** A measurable map $f : \Omega \to \mathbb{R}$ is called *simple* if there exists a $k \geq 1$, a $k$-set partition $A_1, \ldots, A_k$ of $\Omega$ and $a_1, \ldots, a_k \in \mathbb{R}$ such that,

$$f(\omega) = \sum_{i=1}^{k} a_i 1_{A_i}(\omega).$$

The *integral* of a simple $f$ with respect to $P$ is defined as,

$$\int f \, dP = \sum_{i=1}^{k} a_i P(A_i).$$

A straightforward construction shows that for every measurable $f \geq 0$, there exists an increasing sequence $(f_n)$ of non-negative, simple functions such that $f_n(\omega) \uparrow f(\omega)$ for all $\omega \in \Omega$. By the monotony of $(f_n)$, this defines an integral for every non-negative, measurable $f$,

$$\int f \, dP = \lim_{n \to \infty} \int f_n \, dP,$$

(after one demonstrates that the l.h.s. does not depend on the particular $(f_n)$ we choose to approximate $f$). Extension to real-valued measurable functions that take on negative values as well is done by treating negative $f_-$ and non-negative $f_+$ parts of $f$ separately. Extension to $\mathbb{R}^d$ with $d > 1$ proceeds component-wise. The most important result in integration theory is the following elementary theorem.

**Theorem B.4. (Monotone convergence)** Let $(f_n)$ be a monotone sequence of measurable maps $\Omega \to \mathbb{R}$. Then $\lim_n \int f_n \, dP = \int (\lim_n f_n) \, dP$.

Before we can state Fatou’s lemma and the dominated convergence theorem, we define integrability of measurable maps.

**Definition B.11.** Let $(\Omega, \mathcal{F}, P)$ be a probability space. A real-valued measurable function $f : \Omega \to \mathbb{R}$ is said to be *integrable* with respect to $P$ if

$$\int_{\Omega} f \, dP < \infty.$$
It follows immediately from the definition that an integrable \( f \) is a random variable. Note that any sequence of measurable \( f_n \) is dominated by the sequence \( (\sup_{m \geq n} f_m) \). By the monotone class theorem the suprema are measurable and the resulting sequence of maps is monotone decreasing.

**Lemma B.4. (Fatou’s lemma)** Let \( f_n : \Omega \to \mathbb{R} \) be a sequence of measurable maps such that \( f_n \leq g \), \( P \)-almost-surely for all \( n \geq 1 \), for some \( P \)-integrable \( g : \Omega \to \mathbb{R} \). Then,

\[
\limsup_{n \to \infty} \int f_n \, dP \leq \int \left( \limsup_{n \to \infty} f_n \right) \, dP.
\]

An obvious extension provides an inequality for the limes inferior. When combined, the \( \limsup \) and \( \liminf \) versions of Fatou’s lemma imply the following result, known as Lebesgue’s (dominated convergence) theorem.

**Theorem B.5. (Dominated convergence)** Let \( f_n : \Omega \to \mathbb{R} \) be a sequence of measurable maps such that \( \lim_{n} f_n : \Omega \to \mathbb{R} \) exists and \( |f_n| \leq g \), \( P \)-almost-surely for all \( n \geq 1 \), for some \( P \)-integrable \( g : \Omega \to \mathbb{R} \). Then,

\[
\lim_{n \to \infty} \int f_n \, dP = \int (\lim_{n \to \infty} f_n) \, dP.
\]

For any two probability spaces \((\Omega_1, \mathcal{F}_1, P_1)\) and \((\Omega_2, \mathcal{F}_2, P_2)\), the set \( \Omega_1 \times \Omega_2 \) can be endowed with the \( \sigma \)-algebra generated by products of the form \( A_1 \times A_2 \) where \( A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2 \), which is called the product \( \sigma \)-algebra, denoted \( \mathcal{F} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2) \) and a product measure \( P = P_1 \times P_2 \), to arrive at a probability space \((\Omega, \mathcal{F}, P)\), for which the following elementary theorem on the interchangability of integrals applies.

**Theorem B.6. (Fubini’s theorem)** Let \((\Omega_1, \mathcal{F}_1, P_1)\) and \((\Omega_2, \mathcal{F}_2, P_2)\) be probability spaces and denote their product by \((\Omega, \mathcal{F}, P)\). For any non-negative, \( \mathcal{F} \)-measurable \( f : \Omega \to \mathbb{R} \) and any \( \omega_1 \in \Omega_1 \), \( f(\omega_1, \cdot) : \Omega_2 \to \mathbb{R} \) is \( \mathcal{F}_2 \)-measurable. Furthermore, for any \( A_1 \in \mathcal{F}_1 \) and \( A_2 \in \mathcal{F}_2 \),

\[
\int_{A_1 \times A_2} f(\omega_1, \omega_2) \, dP(\omega) = \int_{A_1} \left( \int_{A_2} f(\omega_1, \omega_2) \, dP_2(\omega_2) \right) dP_1(\omega_1) = \int_{A_2} \left( \int_{A_1} f(\omega_1, \omega_2) \, dP_1(\omega_1) \right) dP_2(\omega_2).
\]

Another central result from integration theory forms the foundation for the probability density we associate with many distributions.

**Theorem B.7. (Radon-Nikodym theorem)** Let \((\Omega, \mathcal{F})\) be a measurable space and let \( \mu, \nu : \mathcal{F} \to [0, \infty] \) be two \( \sigma \)-finite measures on \((\Omega, \mathcal{F})\). There exists a unique decomposition

\[
\mu = \mu_\parallel + \mu_\perp,
\]

such that \( \nu_\parallel \ll \nu \) and \( \mu_\perp \) and \( \nu \) are mutually singular. Furthermore, there exists a finite-valued, \( \mathcal{F} \)-measurable function \( f : \Omega \to \mathbb{R} \) such that for all \( F \in \mathcal{F} \),
\[
\mu_\nu(F) = \int_F f \, d\nu. \tag{B.5}
\]

The function \( f \) is \( \nu \)-almost-everywhere unique.

The function \( f : \Omega \to \mathbb{R} \) in the above theorem is called the Radon-Nikodym derivative of \( \mu \) with respect to \( \nu \). If \( \mu \) is a probability distribution, then \( f \) is called the (probability) density for \( \mu \) with respect to \( \nu \). The Radon-Nikodym derivative is sometimes denoted \( d\mu/d\nu \). The assertion that \( f \) is “\( \nu \)-almost-everywhere unique” means that if there exists a measurable function \( g : \Omega \to \mathbb{R} \) such that (B.5) holds with \( g \) replacing \( f \), then \( f = g \), (\( \nu \)-a.e.), i.e. \( f \) and \( g \) may differ only on a set of \( \nu \)-measure equal to zero. Through a construction involving increasing sequences of simple functions, we see that the Radon-Nikodym theorem has the following implication.

**Corollary B.1.** Assume that the conditions for the Radon-Nikodym theorem are satisfied. Let \( X : \Omega \to [0, \infty] \) be measurable and \( \mu \)-integrable. Then the product \( X f \) is \( \nu \)-integrable and

\[
\int X \, d\mu = \int X f \, d\nu.
\]

**Remark B.1.** Integrability is not a necessary condition here, but the statement of the corollary becomes rather less transparent if generalized.

### B.4 Existence of stochastic processes

A stochastic process has the following broad definition.

**Definition B.12.** Let \((\Omega, \mathcal{F}, P)\) be a probability space, let \( T \) be an arbitrary set. A collection of \( \mathcal{F} \)-measurable random variables \((X_t : \Omega \to \mathbb{R} : t \in T)\) is called a stochastic process indexed by \( T \).

When using this definition, one often starts from a collection of (possibly dependent) random quantities without the certainty that there exists a coupling, that is, a probability space like \((\Omega, \mathcal{F}, P)\) above, on which all random quantities can be represented simultaneously as random variables.

Kolmogorov’s existence theorem provides an explicit construction of \((\Omega, \mathcal{F}, P)\). Clearly, if the \( X_t \) take their values in a measurable space \((\mathcal{X}, \mathcal{B})\), the obvious choice for \( \Omega \) is the product \( \mathcal{B}^T \) in which the process takes its values. The question remains how to characterize \( P \) and its domain \( \mathcal{F} \). Kolmogorov assumes that for any finite subset \( S = \{t_1, \ldots, t_k\} \subseteq T \), the distribution \( P_{t_1 \ldots t_k} \) of the \( k \)-dimensional stochastic vector \((X_{t_1}, \ldots, X_{t_k})\) is given.

**Example B.4.** Choose \( T = [0, 1] \) and define random quantities \( f(t) \) for each \( t \in T \). For any finite subset \( S \subseteq T \), we may specify the distribution \( P_S \) of the random vector,

\[
(f(s) : s \in S) \sim P_S.
\]
For example, we may consider multivariate normal distributions $P$ with the property that, for any $s, t \in S$, the expectation $P(f(s) - f(t)) = 0$ and the variance of $f(s) - f(t)$ is proportional to $|s - t|$ (which gives rise to Brownian motion). If a coupling exists, the resulting stochastic process $(f(t) : t \in [0, 1])$ describes random functions $f : [0, 1] \to \mathbb{R}$.

Since the distributions $P_{t_1 \ldots t_k}$ are as yet unrelated and given for all finite subsets of $T$, consistency requirements are implicit if they are to serve as marginals to the probability distribution $P$; if two finite subsets $S_1, S_2 \subset T$ satisfy $S_1 \subset S_2$, then the distribution of $\{X_t : t \in S_1\}$ should be marginal to that of $\{X_t : t \in S_2\}$. Similarly, permutation of the components of the stochastic vector in the above display should be reflected in the respective distributions as well. The requirements for consistency are formulated in two requirements called Kolmogorov’s consistency conditions:

(K1) Let $k \geq 1$ and $\{t_1, \ldots, t_{k+1}\} \subset T$ be given. For any $C \in \sigma(\mathcal{B}^k)$,
$$P_{t_1 \ldots t_k}(C) = P_{t_1 \ldots t_{k+1}}(C \times \mathcal{F}).$$

(K2) Let $k \geq 1$, $\{t_1, \ldots, t_k\} \subset T$ and a permutation $\pi$ of $k$ elements be given. For any $A_1, \ldots, A_k \in \mathcal{B}$,
$$P_{\pi(t_1) \ldots \pi(t_k)}(A_1 \times \ldots \times A_k) = P_{t_1 \ldots t_k}(A_{\pi^{-1}(1)} \times \ldots \times A_{\pi^{-1}(k)}).$$

**Theorem B.8.** (Kolmogorov’s existence theorem)

Let a collection of random quantities $\{X_t : t \in T\}$ be given. Suppose that for any $k \geq 1$ and all $t_1, \ldots, t_k \in T$, the finite-dimensional marginal distributions
$$\{X_t_1, \ldots, X_t_k\} \sim P_{t_1 \ldots t_k}, \quad (B.6)$$
are defined and satisfy conditions (K1) and (K2). Then there exists a probability space $(\Omega, \mathcal{F}, P)$ and a stochastic process $\{X_t : \Omega \to \mathcal{X} : t \in T\}$ such that all distributions of the form (B.6) are marginal to $P$.

Kolmogorov’s approach to the definition and characterization of stochastic processes in terms of finite-dimensional marginals is of great practical value: it allows one to restrict attention to finite-dimensional marginal distributions when characterising the process. The drawback of the construction becomes apparent only upon closer inspection of the $\sigma$-algebra $\mathcal{F}$: it is the $\sigma$-algebra generated by the so-called cylinder sets, which implies that measurability of events restricting an uncountable number of $X_t$’s simultaneously can not be guaranteed. This is the ever-recurring trade-off between generality and strength of a mathematical result: Kolmogorov’s existence theorem always works but it does not give rise to a comfortably ‘large’ domain for the resulting $P : \mathcal{F} \to [0, 1]$. The discussion continues in appendix D, where the Bochner-Komogorov existence theorem is discussed.
In this section, we consider conditioning of probability measures. In first instance, we consider straightforward conditioning on events and illustrate Bayes’s rule, but we also cover conditioning on $\sigma$-algebras and random variables, to arrive at the posterior distribution and Bayes’s rule for densities.

**Definition B.13.** Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $B \in \mathcal{F}$ be such that $P(B) > 0$. For any $A \in \mathcal{F}$, the conditional probability of the event $A$ given event $B$ is defined:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$  \hfill (B.7)

Conditional probability given $B$ describes a set-function on $\mathcal{F}$ and one easily checks that this set-function is a probability measure that assigns probability one to $B$. The conditional probability measure $P(\cdot | B) : \mathcal{F} \to [0, 1]$ can be viewed as the restriction of $P$ to $\mathcal{F}$-measurable subsets of $B$, normalized to be a probability measure. Definition B.13 gives rise to a relation between $P(A|B)$ and $P(B|A)$ (in case both $P(A) > 0$ and $P(B) > 0$, of course), which is called Bayes’s Rule.

**Proposition B.2.** (Bayes’s Rule)

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $A, B \in \mathcal{F}$ be such that $P(A) > 0$, $P(B) > 0$. Then

$$P(A|B) P(B) = P(B|A) P(A).$$

However, only being able to condition on events $B$ of non-zero probability is too restrictive. Moreover, $B$ above is a definite event; it is desirable also to be able to discuss probabilities conditional on events that have not been measured yet, i.e. to condition on a whole $\sigma$-algebra of events like $B$ above.

**Definition B.14.** Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{C}$ be a sub-$\sigma$-algebra of $\mathcal{F}$ and let $X$ be a real-valued $P$-integrable random variable. The conditional expectation of $X$ given $\mathcal{C}$ is any $\mathcal{C}$-measurable random variable $E[X|\mathcal{C}] : \Omega \to \mathbb{R}$ such that,

$$\int_C X \, dP = \int_C E[X|\mathcal{C}] \, dP,$$

for all $C \in \mathcal{C}$.

If we consider some $B \in \mathcal{F}$ with $P(B) > 0$ and the $\sigma$-algebra $\sigma_B = \{\emptyset, B, \Omega \setminus B, \Omega\}$, and we consider definition B.14 for $X = 1_A$, we recover,

$$E[1_A|\sigma_B] = P(A|B) 1_B + P(A|\Omega \setminus B) 1_{\Omega \setminus B}.$$

The condition that $X$ be $P$-integrable is sufficient for existence and uniqueness of $E[X|\mathcal{C}]$ $P$-almost-surely, the proof being an application of the Radon-Nikodym theorem (see theorem 10.1.1 in Dudley (1989)). So conditional expectations are not unique but if we have two different random variables $e_1$ and $e_2$ satisfying the defining conditions for $E[X|\mathcal{C}]$, then $e_1 = e_2$, $P$-almost-surely. Often, the $\sigma$-algebra $\mathcal{C}$ is
the \( \sigma \)-algebra \( \sigma(Z) \) generated by another random variable \( Z \). In that case we denote the conditional expectation by \( E[X|Z] \) and realizations are denoted \( E[X|Z=z] \).

**Definition B.15.** Let \( (\Omega, \mathcal{F}, P) \) be a probability space and let \( \mathcal{C} \) be a sub-\( \sigma \)-algebra of \( \mathcal{F} \). Furthermore, let \( (\mathcal{Y}, \mathcal{B}) \) be a measurable space and let \( Y : \Omega \to \mathcal{Y} \) be a random variable. For each \( A \in \mathcal{F} \) the conditional distribution of \( Y \) given \( \mathcal{C} \) is defined as follows:

\[
P_{Y|\mathcal{C}}(A, \omega) = E[1\{Y \in A\}|\mathcal{C}](\omega),
\]

\( P \)-almost-surely.

Although seemingly innocuous, the fact that conditional expectations are defined only \( P \)-almost-surely poses a rather subtle problem: for every \( A \in \mathcal{B} \) there exists an \( A \)-dependent null-set on which \( P_{Y|\mathcal{C}}(A, \cdot) \) is not defined. This is not a problem if we are interested only in \( A \) (or in a countable number of sets). But usually, we wish to view \( P_{Y|\mathcal{C}} \) as a probability measure, that is to say, it must be well-defined as a map on the \( \sigma \)-algebra \( \mathcal{B} \) almost-surely. Since most \( \sigma \)-algebras are uncountable, there is no guarantee that the corresponding union of exceptional null-sets has measure zero as well. This means that definition B.15 only defines \( P_{Y|\mathcal{C}}(A, \omega) \) per individual \( A \), and not as a map \( A \mapsto P_{Y|\mathcal{C}}(A, \omega) \) for \( P \)-almost-all \( \omega \in \Omega \). The extra property that the conditional distribution is well-defined \( P \)-almost-surely as a map is called regularity of the conditional distribution.

**Definition B.16.** If \( \pi : \mathcal{B} \times \Omega \to [0, 1] \) is such that,

1. for every \( B \in \mathcal{B} \), \( \omega \mapsto \pi(B, \omega) \) is \( \mathcal{C} \)-measurable,
2. there is an \( E \in \mathcal{C} \) with \( P(E) = 0 \) such that for all \( \omega \in \Omega \setminus E \), \( B \mapsto \pi(B, \omega) \) is a probability measure,
3. for all \( C \in \mathcal{C} \),

\[
\int_C \pi(B, \omega) dP(\omega) = P(B \cap C),
\]

then \( \pi \) is said to be a regular conditional distribution.

The existence of a regular conditional probability cannot be guaranteed without further conditions on the underlying probability space.

**Definition B.17.** A topological space \((S, \mathcal{T})\) is said to be a Polish space if \( \mathcal{T} \) is metrizable, complete and separable. Any topological space \( T \) that is the continuous image of a Polish space is called a Souslin space.

Polish spaces appear in many subjects in measure theory: the existence of a countable, dense subset in a metric setting allows constructions based countable covers by metric balls. In this manner Polish spaces allow countable formulations for properties that would involve uncountable collections of subsets otherwise, in correspondence with countability restrictions arising from measure theory. Such a construction occurs in a theorem that guarantees the existence of regular conditional distributions.

**Theorem B.9.** Let \( (\Omega, \mathcal{F}, P) \) be a probability space and let \( \mathcal{Y} \) be a Polish space. If \( Y : \Omega \to \mathcal{Y} \) is a Borel measurable random variable taking values in \( \mathcal{Y} \) and \( \mathcal{C} \) is
any sub-$\sigma$-algebra of $\mathcal{F}$, there exists a (P-almost-surely unique) regular conditional distribution $\Pi_{Y|C} : \mathcal{B} \times \Omega \to [0,1]$.

**Proof.** Proof For a proof of this theorem, the reader is referred to Dudley (1989) [76], theorem 10.2.2).

In Bayesian context we can be more specific regarding the sub-$\sigma$-algebra $C$ since $\Omega = \mathcal{X} \times \Theta$. To condition on $\theta$ we choose $C = \{ \mathcal{X} \times G : G \in \mathcal{F} \}$. Due to this special choice for $C$, $\mathcal{C}$-measurability implies that $\Pi_{Y|\mathcal{F}}(\cdot, (y, \theta))$ is a $\mathcal{C}$-measurable function of $\theta$ which we denote $\Pi_{Y|\theta} : \mathcal{B} \times \Theta \to [0,1]$. Similarly, since the data $Y$ usually takes its values in $\mathbb{R}$ or $\mathbb{R}^n$ (and certainly in a Polish space), the existence of a regular version of the posterior $\Pi_{\theta|Y} : \mathcal{G} \times \mathcal{Y} \to [0,1]$ is guaranteed.
Appendix C
Topologies on spaces of measures

C.1 Convergence in spaces of probability measures

Let $\mathcal{M}(\mathbb{R}, \mathcal{B})$ denote the space of all probability measures on $\mathbb{R}$ with Borel $\sigma$-algebra $\mathcal{B}$.

**Definition C.1. (topology of weak convergence)**
Let $(Q_n)$ and $Q$ in $\mathcal{M}(\mathbb{R}, \mathcal{B})$ be given. Denote the set of points in $\mathbb{R}$ where $Q \rightarrow [0,1]: t \mapsto Q(-\infty,t]$ is continuous by $C$. We say that $Q_n$ converges weakly to $Q$ if, for all $t \in C$, $Q_n(-\infty,t] \rightarrow Q(-\infty,t]$.

Weak convergence has several equivalent definitions.

**Lemma C.1. Portmanteau lemma**
Let $(Q_n)$ and $Q$ in $\mathcal{M}(\mathbb{R}, \mathcal{B})$ be given. The following are equivalent:

(i) $Q_n$ converges weakly to $Q$.
(ii) For every bounded, continuous $f: \mathbb{R} \rightarrow \mathbb{R}$, $Q_nf \rightarrow Qf$.
(iii) For every bounded, Lipschitz $g: \mathbb{R} \rightarrow \mathbb{R}$, $Qng \rightarrow Qg$.
(iv) For all non-negative, continuous $h: \mathbb{R} \rightarrow \mathbb{R}$, $\liminf_{n \rightarrow \infty} Qnf \geq Qf$.
(v) For every open set $F \subset \mathbb{R}$, $\liminf_{n \rightarrow \infty} Qn(F) \geq Q(F)$.
(vi) For every closed set $G \subset \mathbb{R}$, $\limsup_{n \rightarrow \infty} Qn(G) \leq Q(G)$.
(vii) For every Borel set $B$ such that $Q(\partial B) = 0$, $Qn(B) \rightarrow Q(B)$.

In (vii) above, $\partial B$ denotes the boundary of $B$, which is defined as the closure of $B$ minus the interior of $B$.

**Lemma C.2.** When endowed with the topology of weak convergence, the space $\mathcal{M}(\mathbb{R}, \mathcal{B})$ is Polish, i.e. complete, separable and metric.

**Definition C.2. (topology of pointwise convergence)**
Let $(Q_n)$ and $Q$ in $\mathcal{M}(\mathbb{R}, \mathcal{B})$ be given. We say that $Q_n$ converges pointwise to $Q$ if, for all $B \in \mathcal{B}$, $Q_n(B) \rightarrow Q(B)$. 
Definition C.3. (topology of total variation)
Let \((Q_n)\) and \(Q\) in \(\mathcal{M}(\mathbb{R},\mathcal{B})\) be given. We say that \(Q_n\) converges in total variation to \(Q\) if,
\[
\|Q_n(B) - Q(B)\| = \sup_{B \in \mathcal{B}} |Q_n(B) - Q(B)| \to 0.
\]
In exercise 4.6, it is shown that this distance can also be calculated as the \(L_1\)-difference between densities for \(Q_n\) and \(Q\).

Lemma C.3. (Scheffé’s lemma)
Let \((\Omega,\mathcal{F},\mu)\) be a measure space. Given a sequence \((f_n)\) of integrable functions and a measurable function \(f\) such that \(f_n(\omega) \to f(\omega)\) for \(\mu\)-almost-all \(\omega\), then
\[
\int |f_n - f| \, d\mu \to 0 \text{ if and only if } \int |f_n| \, d\mu \to \int |f| \, d\mu.
\]

Corollary C.1. Let \((\Omega,\mathcal{F},\mu)\) be a measure space. If a sequence of probability densities \((p_n)\) converges to a probability density \(p\) \(\mu\)-almost-everywhere, then
\[
\|P_n - P\| \to 0.
\]

C.2 Inverse limits and locally convex spaces

Bourbaki introduces the so-called inverse limit space (known also as projective limit space [198]) as a construction that can be interpreted at many levels of detail. In a set-theoretic way ([41], Ch. III, §7, No. 1; [41], Ch. III, §7, No. 2; [41], Ch. R, §6, No. 2), we describe a directed set \(I\) and a system of sets \((X_\alpha)_{\alpha \in I}\) with maps \(f_{\alpha \beta} : X_\beta \to X_\alpha\) such that,
(i) for all \(\alpha \leq \beta \leq \gamma\), we have \(f_{\alpha \gamma} = f_{\alpha \beta} \circ f_{\beta \gamma}\),
(ii)for all \(\alpha \in I\), \(f_{\alpha \alpha} = i_\alpha\), the identity mapping on \(X_\alpha\).

Conceptually, the maps \(f_{\alpha \beta}\) should be thought of as a system of projections. The inverse limit of the sets \((X_\alpha)\) is then defined as the set \(X\) of all \(x\) in the set-product \(\prod_{\alpha \in I} X_\alpha\) that satisfy,
\[
pr_\alpha(x) = f_{\alpha \beta}(pr_\beta(x)),
\]
for all \(\alpha \leq \beta\). The restriction of \(pr_\alpha\) to \(X\) is denoted \(f_\alpha : X \to X_\alpha\) and called the canonical mapping of \(X\) into \(X_\alpha\); these mappings form a so-called coherent family, i.e. they satisfy \(f_\alpha = f_{\alpha \beta} \circ f_\beta\) for all \(\alpha \leq \beta\). (An immediate point of caution concerns the still-open possibility that \(X = \emptyset\).) For an inverse system of topological spaces \((\mathcal{X}_\alpha, \mathcal{A}_\alpha)_{\alpha \in I}\) more can be said (see [42], Ch. I, §4, No. 4).

Definition C.4. In the above setup, assume that the \(X_\alpha\) are topological spaces and that the maps \(f_{\alpha \beta}\) are continuous for all \(\alpha \leq \beta\). The inverse limit \(X\) of the topological spaces \(X_\alpha\) is the set-theoretic inverse limit \(X\), endowed with the coarsest topology that makes all \(f_\alpha : X \to X_\alpha\) continuous.

Example C.1. Let \(I\) be a directed set and denote by \(\mathcal{I}\) the collection of all finite \(S \subseteq I\). Given a family of topological spaces \((X_\alpha)_{\alpha \in I}\), the product space \(X = \prod_{\alpha \in I} X_\alpha\) can be defined equivalently as the inverse limit of the finite topological products,
\[ \mathcal{X}_S = \prod_{\alpha \in S} \mathcal{X}_\alpha \]

where \( \mathcal{X} \) is the directed by inclusion. The maps \( f_{ST} : \mathcal{X}_T \to \mathcal{X}_S \) for \( S \subset T \) are projections between finite product spaces. The canonical projections \( f_S : X \to \mathcal{X}_S \) are the usual projections \( \pi_S \). The corresponding inverse limit topology is therefore the coarsest topology on the set-theoretic product \( \mathcal{X} \) that makes all projections continuous, the usual definition of the topological product.

To characterize convergence and continuity in inverse limit spaces, we have the following specification of [42], Ch. I, § 2, No. 3, Prop. 4.

**Proposition C.1.** Let the topological space \( (X, \mathcal{I}) \) be the inverse limit of the inverse system of topological spaces \( ((X_\alpha, \mathcal{I}_\alpha))_{\alpha \in I} \). Then the collection of all finite intersections of sets of the form \( f_\alpha^{-1}(U) \) (\( \alpha \in I \) and \( U \in \mathcal{I}_\alpha \)), forms a basis for \( \mathcal{I} \). Furthermore, given a topological space \( \mathcal{Y} \), a map \( h : \mathcal{Y} \to X \) is continuous, if and only if \( f_\alpha \circ h : \mathcal{Y} \to \mathcal{X}_\alpha \) for all \( \alpha \in I \).

One may wonder which topological properties survive inverse limits. For example, an inverse limit is compact and non-empty whenever the spaces \( \mathcal{X}_\alpha \) are compact and non-empty ([42], Ch. I, § 9, No. 6, prop. 8); a so-called inverse limit of uniform spaces requires the \( f_{\alpha\beta} \) and \( f_\alpha \) to be uniformly continuous and leads to a uniformity on the limit space \( \mathcal{X} \) (the coarsest that makes all \( f_\alpha \) uniformly continuous [42], Ch. II, § 2, No. 7, prop. 8). In case the uniform spaces \( \mathcal{X}_\alpha \) are complete, the inverse limit \( \mathcal{X} \) is also complete ([42], Ch. II, § 3, No. 5, Prop. 10 and Cor.). Note the following criterion for the Cauchy property (which is a specific version of [42], Ch. II, § 3, No. 1, Prop. 4).

**Proposition C.2.** Let \((X, \mathcal{Y})\) be the inverse limit of uniform spaces \( ((\mathcal{X}_\alpha, \mathcal{U}_\alpha))_{\alpha \in I} \). Then the collection of all finite intersections of subsets of the form \( (f_\alpha^{-1}, f_\alpha^{-1}) \) \((V) \), where \( \alpha \in I \) and \( V \) an entourage from \( \mathcal{U}_\alpha \) forms a fundamental system of entourages for \( \mathcal{Y} \). Furthermore, given a uniform space \( \mathcal{Y} \), a map \( h : \mathcal{Y} \to X \) is uniformly continuous, if and only if \( f_\alpha \circ h : \mathcal{Y} \to \mathcal{X}_\alpha \) is uniformly continuous for all \( \alpha \in I \). Moreover, a filter base \( \mathcal{C} \) on \( \mathcal{X} \) is Cauchy, if and only if \( f_\alpha(\mathcal{C}) \) is Cauchy for all \( \alpha \in I \).

**Example C.2.** Statistical models for the distributions of i.i.d. samples \( X = (X_1, X_2, \ldots) \) carry a natural uniformity that arises as an inverse limit. Let \((\mathcal{X}_1, \mathcal{B}_1)\) denote the measurable space in which each of the sample points \( X_i \) (\( i \geq 1 \)) takes its values, so \( X \) lies in \( \mathcal{X} = \bigotimes_{i=1}^{\infty} \mathcal{X}_i \). Like before, the sample space \( \mathcal{X}_1^n \) for the first \( n \) sample points is denoted \( \mathcal{X}_n \) with product \( \sigma \)-algebra \( \mathcal{B}_n \). Let \( \mathcal{P} \) denote a collection of probability measures on \((\mathcal{X}_1, \mathcal{B}_1)\). Consider the \( \sigma \)-algebra \( \mathcal{B} \) of subsets of \( \mathcal{X} \) generated by the cylinder sets and note that the model \( \mathcal{P} \) is mapped one-to-one to a collection of infinite product measures \( \mathcal{P}_n = \{ P^n : \mathcal{P} \to [0, 1] : P \in \mathcal{P} \} \).

**Definition C.5.** For each \( n \geq 1 \), consider the linear space \( \mathcal{X}_n \) of all bounded \( \mathcal{B}_n \)-Borel-measurable \( f : \mathcal{X}^n \to \mathbb{R} \) and consider the fundamental system of entourages obtained by choosing \( k \geq 1 \), and \( f_1, \ldots, f_k \in \mathcal{X}_n \) to define,
\[ W_{n,f_1,\ldots,f_k} = \left\{ (P, Q) \in \mathcal{P}^\infty \times \mathcal{P}^\infty : \left| (P - Q)f_i \right| < 1, 1 \leq i \leq k \right\}. \]

For every \( n \geq 1 \), these subsets of \( \mathcal{P}^\infty \times \mathcal{P}^\infty \) form a fundamental system of entourages for a uniformity \( \mathcal{U}_n \) on \( \mathcal{P}^\infty \) (and by extension with slight abuse of notation, also on \( \mathcal{P} \), which is in bijective correspondence with the diagonal in \( \mathcal{P}^\infty \) and can so inherit the subspace uniformity). Associated topologies (on \( \mathcal{P}^\infty \) and \( \mathcal{P} \)) are denoted \( \mathcal{T}_n \), for all \( n \geq 1 \). Identifying \( n \geq 1 \) as the index \( \alpha \), we consider the inverse limit of the inverse system of uniform spaces \( ((\mathcal{P}^\infty, \mathcal{U}_n), f_{nm})_{n \geq 1} \) and denote the resulting uniform space by \( (\mathcal{P}^\infty, \mathcal{U}_\infty) \). Again with slight abuse of notation, we denote by \( \mathcal{U}_\infty \) the inverse limit uniformity on \( \mathcal{P} \) from the \( \mathcal{U}_n \). The associated topologies on \( \mathcal{P}^\infty \) and \( \mathcal{P} \) are both denoted \( \mathcal{T}_\infty \).

Note that \( \mathcal{P}^\infty \) may be replaced by its linear span \( \mathcal{L} \), with straightforward extension of the definition of the entourages \( W_{n,f_1,\ldots,f_k} \), without changing the conclusions above. In that case, also define the linear space \( \mathcal{F} \) that consists of all maps \( f : X^\infty \to \mathbb{R} \) that are in the union of the images of the spaces \( \mathcal{F}_n \) under the canonical embeddings of \( \mathcal{F}_n \) in the space of all maps \( X^\infty \to \mathbb{R} \). To define \( \mathcal{F} \) topologically, we view \( \{ \mathcal{F}_n : n \geq 1 \} \) as a system of locally convex spaces (by means of the collection semi-norms \( p_\mu : \mathcal{F}_n \to \mathbb{R} : f \mapsto |\mu f| \) for \( \mu \in \mathcal{L} \)). The space \( \mathcal{F} \) is the direct (or inductive limit) of the system \( (\mathcal{F}_n, f_{nm}) \), with canonical injections for all \( n \leq m \), \( f_{nm} : \mathcal{F}_n \to \mathcal{F}_m \) that are trivially continuous (see [46], Ch. II, §4, No. 5, Example II).

The locally convex spaces \( \mathcal{L} \) and \( \mathcal{F} \) are then placed in dual correspondence, via the bilinear form \( B(\mu, f) = \mu f \). Particularly, the topology \( \mathcal{T}_\infty \) on \( \mathcal{L} \) associated with \( \mathcal{U}_\infty \) coincides with the weak topology \( \sigma(\mathcal{L}, \mathcal{F}) \); the topology on the direct limit \( \mathcal{F} \) is \( \sigma(\mathcal{F}, \mathcal{L}) \).

Alternatively, we consider the collection \( C_b(\mathcal{F}) \) of all bounded continuous maps \( f : \mathcal{F} \to \mathbb{R} \) to define a fundamental system of entourages.

**Definition C.6.** Define, for every \( f \in C(\mathcal{F}^n) \), the entourage,

\[ W^C_{n,f} = \{(P, Q) \in \mathcal{P} \times \mathcal{P} : |P^n f - Q^n f| < 1\}. \]

The collection of finite intersections \( \bigcap_{i=1}^l W_{n,f_i} \) (for some \( f_1, \ldots, f_n \in C(\mathcal{F}^n) \)) forms a basis for a uniformity \( \mathcal{U}^C_n \). Define \( \mathcal{U}^C = \bigcup_{n \geq 1} \mathcal{U}^C_n \). Denote the corresponding uniform topologies on \( \mathcal{P} \) by \( \mathcal{T}^C_n \) and \( \mathcal{T}^C_\infty \). Clearly, \( \mathcal{T}^C_1 \subset \mathcal{T}^C_2 \subset \cdots \subset \mathcal{T}^C_\infty \).

The topology \( \mathcal{T}^C_\infty \) is called the tight topology in [45], chapter XI. The topologies \( \mathcal{T}_1 \) and \( \mathcal{T}^C_\infty \) are very different.

**Example C.3.** Suppose that \( \mathcal{F} = [0,1] \) with the Borel \( \sigma \)-algebra and we consider the (rather deterministic) collection of all atomic measures \( \mathcal{P} = \{ \delta_x : x \in \mathcal{F} \} \) as our model. We identify \( \mathcal{F} \) and \( \mathcal{P} \) through the bijection, \( \delta : \mathcal{F} \to \mathcal{P} : x \mapsto \delta_x \). Note that for every \( x \in \mathcal{F} \), there exists a (measurable but discontinuous) \( f \) such that \( f(x) = 1 \) and \( f(y) = 0 \) for all \( y \in [0,1], y \neq x \). Conclude that \( \mathcal{U}_1 \) is the discrete uniformity on \( \mathcal{P} \), and hence, so is \( \mathcal{U}_\infty \). That means that any function \( g \) with \( \mathcal{P} \) as its domain is \( \mathcal{U}_\infty \)-uniformly-continuous. According to the Le Cam-Schwartz theorem, this fact
renders any disjoint model subsets $B, V$ (uniformly) testable, which is appropriate in deterministic setting. Note that the map $x \mapsto \delta_x$ is not continuous (and hence, not a parametrization c.f. the definition at the beginning of section 8.2 unless we equip $[0, 1]$ also with the discrete topology).

By contrast, the map $x \mapsto \delta_x$ is continuous if we equip $\mathcal{P}$ with the $T_\infty$ topology. Since $[0, 1]$ is compact, $\delta \cdot$ is actually a homeomorphism in this case (and that fact remains true for any Polish $\mathcal{X}$, see Proposition 10 of [45], Ch. IX, § 5, No. 4. for the general statement). That means that $\mathcal{P}$ with the $T_\infty$ topology is Polish (another fact that remains true for any Polish $X$, see Proposition 13 of [44], Ch. III, § 1, No. 9.) That means that $\mathcal{P}$ if we restrict attention to the $T_\infty$-topology, a connection that is lost when we refine to $T_1$. For example, the range $E$ of the map $x \mapsto \delta_x$ is the set of extremal points in $M_1^+(\mathcal{X}, B)$, and the convex hull of $E$ is $T_\infty$-dense in $M_1^+(\mathcal{X}, B)$ (but not $T_\infty$-dense unless $\mathcal{X}$ is countable). This topological difference between $T_\infty$ and $T_1$ is essential and explains the “inaccessibility” [156, 162] of the latter in a relative fashion.

(Relative) compactness with respect to $T_1$ is the domain of the Dunford-Pettis theorem [71]. For the next theorem only, assume that $\mathcal{P}$ is norm-bounded collection of finite, positive measures, dominated by a probability measure $Q$ and represented as a family $\mathcal{P}_Q = \{dP/dQ : P \in \mathcal{P}\}$ in $L^1(Q)$. The continuous dual of $L^1(Q)$ is $L^\infty(Q)$ and the model $\mathcal{P}$ with the $T_\infty$-topology is homeomorphic with $\mathcal{P}_Q$ as a subspace of $L^1(Q)$ with the weak topology.

**Theorem C.1.** (Dunford-Pettis) Assume $\mathcal{P}_Q$ is a norm-bounded subset of $L^1(Q)$; $\mathcal{P}_Q$ is relatively weakly compact, if and only if, for every $\epsilon > 0$ there is an $M > 0$ such that,

$$\sup_{P \in \mathcal{P}} \int [dP/dQ > M] \frac{dP}{dQ} dQ < \epsilon,$$

that is, $\mathcal{P}_Q$ is uniformly $Q$-integrable.

It is shown in the proof of lemma 3 of section 17.5 of Le Cam (1986) [162] (in the somewhat broader context of theorem 6 of appendix 8 in [162]) that weak convergence of a net $f_a \to f$ in $L^1(Q)$ implies weak convergence of product densities $f^n_a \to f^n$ weakly in $L^1(Q^n)$, as a result of the Dunford-Pettis theorem (see also lemma 3.8 in [208]). Consequently, a net in $\mathcal{P}$ that has a $T_1$-convergent subnet, also has a $T_\infty$-convergent subnet, so $T_1$-compactness implies $T_\infty$-compactness for all $n \geq 1$, which implies $T_\infty$-compactness.

**Proposition C.3.** Let $\mathcal{P}$ be a model for i.i.d. data $X^n$; $\mathcal{P}$ is $T_1$-compact, if and only if, $\mathcal{P}$ is $T_\infty$-compact.

(Relative) compactness of the model $\mathcal{P}$ for $T_C$ is the realm of Prokhorov’s theorem. Here it is assumed that $\mathcal{X}$ is a Hausdorff topological space with Borel $\sigma$-algebra. In [43], Ch. IX, § 5, No. 5, the following is referred to as Prokhorov’s property.

**Definition C.7.** Let $H$ be a subset of $M(\mathcal{X})$; $H$ is said to be uniformly tight if,
(i) sup\{∥µ∥ : µ ∈ H\} < ∞,
(ii) for every ε > 0, there is a compact K in X such that,
\[ \sup\{µ(X \setminus K) : µ ∈ H\} ≤ ε. \]

For probability models P the uniform bound in norm is always satisfied and only the second condition plays a role when one verifies relative compactness for TC.

**Theorem C.2.** Assume that X is completely regular. A subset H of M(X) that is uniformly tight, is H relatively compact for TC.

**Proof.** For a proof, see [43], Ch. IX, §5, No. 5, theorem 1.

In locally compact or Polish spaces, uniform tightness is equivalent to TC-relative compactness.

**Theorem C.3.** Assume that X is locally compact or Polish. A subset H of M(X) that is H relatively compact for TC, is uniformly tight.

**Proof.** For a proof, see [43], Ch. IX, §5, No. 5, theorem 2.

Note that also regarding matters of compactness, the T1 and TC topologies are different in that the TC compactness criterion refers to a topological feature of the sample space (the compact subset K of X), while the T1 compactness criterion does not and is formulated as a property that derives from X as a measurable space (uniform integrability). The associated strong topologies also maintain a distinction of the type.

**Proposition C.4.** The strong topologies associated with T1 and T∞ are equal to the total-variational topology. The strong topology associated with TC is TC itself.

To conclude with an example, we consider a sequence \((P_n)\) of probability measures that converges in TC but not T1. The example also shows how TC-compact sets can be non-compact for T1.

**Example C.4.** Consider X = [0, 1] with the Borel σ-algebra with distributions Pn defined by their Lebesgue measures pn for all n ≥ 1, pn(x) = nI{0 ≤ x ≤ 1/n}. For any continuous g : [0, 1] → R,
\[ \inf_{0 ≤ x ≤ 1/n} g(x) ≤ \int_0^1 g(x) dP_n(x) ≤ \sup_{0 ≤ x ≤ 1/n} g(x), \]
and both bounds go to g(0) as n → ∞, so \(P_n → δ_0\) in TC. However, the collection \\{P_n : n ≥ 1\} does not satisfy the condition of theorem C.1, so \(P_n\) does not converge for T1.
C.3 Contiguity

First, let us recall the definition of contiguity [157] (see [162] for alternatives, e.g. in terms of limiting domination in a sequence of binary experiments).

**Definition C.8.** Given measurable spaces \((X_n, B_n), n \geq 1\) with two sequences \((P_n)\) and \((Q_n)\) of probability measures, we say that \(Q_n\) is **contiguous with respect to** \(P_n\), notation \(Q_n \triangleleft P_n\), if,

\[
P_n \phi_n(X^n) = o(1) \quad \Rightarrow \quad Q_n \phi_n(X^n) = o(1),
\]

for every sequence of \(B_n\)-measurable \(\phi_n : X_n \rightarrow [0, 1]\).

The value of the notion of contiguity does not just reside with the usefulness of the property itself, but also with the multitude of accessible characterizations listed in Le Cam’s famous First Lemma (see, e.g., Hajek and Sidak (1967) [114]). (One of the formulations requires that we define the so-called Hellinger transform \(\psi(P, Q; \alpha) = \int p^\alpha q^{1-\alpha} d\mu\), where \(p\) and \(q\) denote densities for \(P\) and \(Q\) with respect to a \(\sigma\)-finite measure that dominates both \(P\) and \(Q\).)

**Lemma C.4. (Le Cam’s First Lemma)**

Given measurable spaces \((X_n, B_n) : n \geq 1\) with two sequences \((P_n)\) and \((Q_n)\) of probability measures, the following are equivalent:

(i) \(Q_n \triangleleft P_n\),

(ii) for any measurable \(T_n : X_n \rightarrow \mathbb{R}\), if \(T_n \xrightarrow{P_n} 0\), then \(T_n \xrightarrow{Q_n} 0\),

(iii) given \(\varepsilon > 0\), there is a \(b > 0\) such that \(Q_n(dQ_n/dP_n > b) < \varepsilon\), for large enough \(n\),

(iv) given \(\varepsilon > 0\), there is a \(c > 0\) such that \(\|Q_n - Q_n \wedge cP_n\| < \varepsilon\), for large enough \(n\),

(v) if \(dP_n/dQ_n \xrightarrow{Q_n}-w. f\) along a subsequence, then \(P(f > 0) = 1\),

(vi) if \(dQ_n/dP_n \xrightarrow{P_n}-w. g\) along a subsequence, then \(Eg = 1\),

(vii) Hellinger transforms satisfy, \(\liminf_n \limsup_{\alpha \uparrow 1} \psi(P_n, Q_n; \alpha) = 1\).

A proof of this form of the First Lemma can be found in [162], section 6.3. Note the relation to testing: for two sequences \((P_n)\), \((Q_n)\) that are mutually contiguous \((P_n \triangleleft Q_n\) and \(Q_n \triangleleft P_n\), there exists no test sequence that separates \((P_n)\) from \((Q_n)\) asymptotically. Loosely said, \((P_n)\) and \((Q_n)\) are indistinguishable statistically regardless of the amount of data available. Much more can be said about contiguity (to begin with, see, Roussas (1972) [196] and Greenwood and Shiryaev (1985) [111]), for instance in relation to Le Cam’s convergence of experiments, but also, specific relations that exist in the locally asymptotically normal case (e.g. Le Cam’s Third lemma [114], which relates the laws of a statistic under \(P_n\) and \(Q_n\) in such context).
Appendix D
Inverse limit measures

Existence of the Dirichlet process prior of section 6.4, defined in terms of an inverse system of marginals like (6.7), was left as an open issue following lemma 6.1, because its proof appears bona-fide but does not prove the assertion (see exercise 6.1). Here we consider the question of existence for the family of inverse limit measures, which contains the Dirichlet family, the Pólya-urn family and other well-known examples. In this appendix, we formulate a necessary and sufficient condition for the existence of measures and distributions defined by inverse systems of marginals on completely regular spaces, with special attention to the case of Polish sample spaces.

D.1 The Bochner-Kolmogorov existence theorem

There are many ways to define random probability measures in the full model: the most direct methods involve, for example, defining (finite or countable) sums of randomly weighted (fixed or random) pointmasses or other kernels, or defining random laws for moments or quantiles. Most appealing, however, is a construction called an inverse limit measure: for a refining collection $\mathcal{A}$ of finite partitions $\alpha$, we specify, for every $\alpha = \{A_1, \ldots, A_N\}$ the finite-dimensional distribution,

$$(P(A_1), \ldots, P(A_N)) \sim \Pi_\alpha.$$ 

Here, additivity is reflected when $\alpha, \beta \in \mathcal{A}$ and $\beta$ refines $\alpha$: the distribution of the sum of constituent probabilities must match the distribution of the probability of the whole (like those of (6.7), which satisfy (6.8), see also (D.4) below). Such coherency of the definitions for the marginals $\Pi_\alpha$ is a first requirement, but it is not sufficient (see, e.g., example D.1), there is a further condition. It is also instructive to keep the converse in mind from the beginning: for Bayesians, random probability measures are central and every random probability measure projects to a coherent inverse system of random ‘histograms’.

319
The literature on inverse limit measures displays a wide variety of sufficient conditions. First explorations of the subject date back to Bochner and Kolmogorov (see Bochner (1955) [38]). The well-known Bochner-Kolmogorov theorem formulates sufficient conditions for existence of a random distribution function on $\mathbb{R}$; compact Hausdorff spaces are considered in early work by Choksi (1958) [54]. The work of Kingman (1967) [139] explores completely random measures, which arise when elements of a partition are assigned independent measures. Ferguson (1973,1974) [89, 90] defines the Dirichlet process as a normalized completely random measure, and gives the flawed existence proof of lemma 6.1. Alternative methods to construct the Dirichlet, Pólya-urn and an array of other concrete families of priors on the full model (e.g. stick-breaking, Chinese restaurant and Indian buffet processes) have been studied extensively and imply their own, specific proofs of existence (for overviews, see the very complete discussions in [188, 106]). Ghosh and Ramamoorthi (2003) [103] construct priors with Bochner-Kolmogorov-type conditions on an inverse system for distribution functions on $\mathbb{R}$ and prove existence. Most recently, in theorem 3.1 of Ghosal and van der Vaart (2018) [106], it is required that the prior predictive distribution,

$$A \mapsto \int_{\mathcal{A}} P(A) \, d\Pi(P)$$

defines a Borel measure.

(Strictly speaking this conditions is tautological. However, one realizes that, in examples, the prior predictive distribution is often available, defined at least on a generating ring by the marginal distributions.) Note that the more generally accepted [105, 183] condition that the random $P$ satisfy,

$$\Pi(P \text{ is a probability distribution}) = \Pi(0 \leq P \leq 1 \text{ is } \sigma\text{-additive}) = 1, \quad (D.1)$$

also refers to $\Pi$ before proving it exists, unless we have an exactly defined notion of what (D.1) means in terms of the marginals (which is far from immediate). What is really needed are accessible necessary and sufficient conditions, imposed on the inverse system of marginals rather than its presumed limit.

The conditions we derive here are based on an equivalence in [45] (see theorem D.3 below) that focusses on the existence of approximating compacta, much in the way of Prokhorov’s condition for uniform tightness (c.f. definition C.7). This leads to topological characterization of those inverse systems $(\Pi_\alpha : \alpha \in \mathcal{A})$ that consistently define random measures $\Pi$. That statement depends on the model topology through the domain for the (Borel) measure $\Pi$, and characterizations vary. For the full model with topology $\mathcal{T}_I$, which is the case studied in Rao (1971) [192], Orbanz (2011) [184] and Beznea and Cimpean (2014) [26], the application of theorem D.3 is straightforward but leads to a rather inaccessible necessary and sufficient condition; for the full model with topology $\mathcal{T}_C$, theorem D.3 is applied through a zero-dimensional version of the samplespace, and leads to an accessible and intuitive necessary and sufficient condition (effectively doubling definition C.7 when applied in Polish samplespaces, the most common case).
D.2 Topological characterization of existence

We rely on the theory of integration of N. Bourbaki [44, 45], which revolves around
the observation that on Hausdorff completely regular spaces, bounded regular mea-
sure are identified uniquely with elements of the dual of the normed space of
bounded, continuous functions. (See [45], Ch. IX, § 5, No. 2, proposition 5.) (For
a Hausdorff completely regular space \( T \), \( C^b(T) \) denotes the linear space of all
bounded, continuous \( f : T \to \mathbb{R} \), equipped with the uniform norm.)

**Theorem D.1.** (Riesz-Markov-Kakutani) Let \( T \) be a Hausdorff completely regular
space and \( I \) a continuous linear form on the normed space \( C^b(T) \). In order that
there exist a bounded regular measure \( \mu \) such that \( I(f) = \int f \, d\mu \), for all \( f \in C^b(T) \),
it is necessary and sufficient that,

\[ (R) \text{ for every } \epsilon > 0, \text{ there is a compact } K \text{ in } T, \text{ such that } ||g|| \leq 1, \ g|_K = 0 \text{ imply } |I(g)| \leq \epsilon. \]

Regarding the partitions of the samplespace \( \mathcal{X} \), we require \( \mathcal{X} \) to be a Haus-
dorff completely regular spaces (or more specific, in particular, Polish spaces). We
consider two types of partitions: measure-theoretically the most natural choice, is
to consider \( \mathcal{A} \) consisting of all finite partitions of \( \mathcal{X} \) into measurable subsets, or a
directed subset thereof. When viewed from the topological point of view, the natural
choice for \( \mathcal{A} \) is much smaller: let \( \mathcal{X} \) be a Hausdorff completely regular space with
basis \( \mathcal{U} \).

**Definition D.1.** We say that a partition \( \alpha \) is generated by the basis \( \mathcal{U} \), if \( \alpha \) can
be constructed through a finite number of intersections of \( \mathcal{X} \) with \( U \) and \( \mathcal{X} \setminus U \),
\( U \in \mathcal{U} \). We say that \( \mathcal{A} \) is generated by the basis \( \mathcal{U} \), if \( \mathcal{A} \) contains all partitions
generated by \( \mathcal{U} \).

(For illustrations as well as more specific requirements for smaller \( \mathcal{A} \), refer to ex-
amples D.3 and D.4.)

Once a choice for \( \mathcal{A} \) is made, we consider the projection maps, for all \( \alpha \in \mathcal{A} \)
(with \( N \) denoting the order of \( \alpha \)),

\[ \phi_\alpha : \mathcal{X} \to \mathcal{X}_\alpha : x \mapsto (1_{A_1}(x), \ldots, 1_{A_N}(x)), \]

where \( \mathcal{X}_\alpha \) is the (finite) set of unit vectors in \( \mathbb{R}^N \). Implied is a projection of bounded,
regular measures,

\[ \hat{\phi}_\alpha : M(\mathcal{X}) \to M(\mathcal{X}_\alpha) : \mu \mapsto \mu_\alpha = (\mu(A_1), \ldots, \mu(A_N)), \]

If we think of the projections \( \mu_\alpha \) as random probability measures on \( \mathcal{X}_\alpha \), the asso-
ciated system of marginal probability measures \( \Pi_\alpha \) on the spaces \( M(\mathcal{X}_\alpha) \) must be
coherent. Our most accessible version of the Bochner-Kolmogorov theorem for the
existence of inverse limit measures is then formulated as follows. Below, the space
\( M^1(\mathcal{X}) \) is equipped with Prokhorov’s weak topology.
Theorem D.2. For a Polish samplespace $\mathcal{X}$ with $\mathcal{A}$ generated by a basis $\mathcal{U}$, a coherent system of marginal distributions $\Pi_{\alpha}$ on the spaces $M^1(\mathcal{X}_{\alpha})$, gives rise to a bounded, regular distribution $\Pi$ on $M^1(\mathcal{X})$, if and only if, for every $\epsilon, \delta > 0$, there exists a compact $K$ in $\mathcal{X}$ such that, for all $\alpha \in \mathcal{A}$,

$$\Pi_{\alpha}(\{\mu_{\alpha} \in M(\mathcal{X}_{\alpha}) : \mu_{\alpha}(\mathcal{X}_{\alpha} \setminus K_{\alpha}) > \delta\}) < \epsilon,$$

(D.2)

where $K_{\alpha} = \cup\{A_i \in \alpha : K \cap A_i \neq \emptyset\}$. The measure $\Pi$ is then unique and,

$$\Pi(L) = \inf\{\Pi_{\alpha}(\hat{\phi}_{s,\alpha}(L)) : \alpha \in \mathcal{A}\},$$

for every compact set $L$ in $M(\mathcal{X})$.

Proof. See corollary D.1, applied in the case that the $\Pi_{\alpha}$ are probability measures on $M^1(\mathcal{X}_{\alpha})$. The latter specification implies that, with $R = 1$, any $\epsilon > 0$ and any $\beta \in \mathcal{A}$, $\Pi_{\beta}(\mu_{\beta}(\mathcal{X}_{\beta}) \leq R) = 1 > 1 - \epsilon$. Condition (D.2) then shows that property $(P''')$ is satisfied.

We generalize to the case of a completely regular samplespace with a random finite measure, specified as a coherent system of inverse measures on smaller collections $\mathcal{A}$: the double Prokhorov property (D.2) changes slightly, a necessary and sufficient condition is provided in theorem D.5. We also provide a necessary and sufficient condition in case the space $M(\mathcal{X})$ is equipped with the topology $\mathcal{T}_1$ in proposition D.1, but argue that it is impractical and easily grows into tautologies like those formulated above.

To illustrate the intention of condition (D.2), consider the following example.

Example D.1. Here we refer to the sequence $(\alpha_n)$ of partitions for $\mathbb{R}$ of example D.4, each of which contains an element of the form,

$$A_{0,n} = (-\infty, q_n] \cup [q_n, \infty),$$

with $q_n \to \infty$. We give a coherent definition for the marginals $\Pi_n = \Pi_{\alpha_n}, n \geq 1$, that does not satisfy condition (D.2). For all $n \geq 1$, consider degenerate marginals $\Pi_n$ which assign $\mu_n (= \mu_{\alpha_n})$-mass one to the set $A_{0,n}$, with $\Pi_n$-probability one:

$$\Pi_n(\mu_n(\{(1,0,\ldots,0)\}) = 1) = 1.$$

These $(\Pi_n)$ form a consistent inverse system. For any compact $K$ in $\mathbb{R}$, $K \cap A_{0,n} = \emptyset$ for large enough $n$, which invalidates (D.2) for $\epsilon < 1$. Explained more intuitively, the problem occurs because the marginals $\mu_n$ on which the $\Pi_n$ fixate, shift mass towards $\pm \infty$ without limitation, which would mean that compact sets $K$ receive mass zero from any presumed limit measure. There is no such limit, because any measure on $\mathbb{R}$ is regular. Not all counterexamples are so extreme: it is enough to have a consistent system $(\Pi_n)$, such that, along a subsequence $(\alpha_{n(m)})$ and for some $\delta > 0$,

$$\lim_{m \to \infty} \Pi_{n(m)}(\mu_{n(m)}(\{(1,0,\ldots,0)\}) > \delta) > 0,$$
D.3 Inverse limits of positive measures

In this section, we assume that $\mathcal{X}$ is a a Hausdorff completely regular space and denote by $\mathscr{A}$ a collection of finite partitions of $\mathcal{X}$.

**Definition D.2.** A collection $\mathscr{A}$ is said to be *separating* if, for every $x, y \in \mathcal{X}, x \neq y$, there exist $\alpha \in \mathscr{A}$ and $A \in \alpha$ such that $x \in A$ and $y \notin A$.

For any $\alpha \in \mathscr{A}$, let $N(\alpha) \in \mathbb{N}$ denote its order, let $I(\alpha)$ denote the set $\{1, \ldots, N(\alpha)\}$ and write $\alpha = \{A_1, \ldots, A_{N(\alpha)}\}$ for non-empty $A_i (1 \leq i \leq N(\alpha))$. For every $\alpha \in \mathscr{A}$, let $\mathcal{X}_\alpha$ denote the discrete space of unit vectors in $\mathbb{R}^{N(\alpha)}$, $\mathcal{X}_\alpha = \{e_1, \ldots, e_{N(\alpha)}\} \subset \mathbb{R}^{N(\alpha)}$ (where $e_i$ denotes the $i$-th unit vector in $\mathbb{R}^{N(\alpha)}$) and,

$$\varphi_\alpha : \mathcal{X} \rightarrow \mathcal{X}_\alpha : x \mapsto (1_{A_1}(x), \ldots, 1_{A_{N(\alpha)}}(x)).$$

**Definition D.3.** Given a space $\mathcal{X}$ and a collection of spaces $\mathcal{X}_\alpha$, $\alpha \in \mathscr{A}$, a collection of functions $\varphi_\alpha : \mathcal{X} \rightarrow \mathcal{X}_\alpha$ is said to be *separating*, if for all $x, y \in \mathcal{X}, x \neq y$, there exist an $\alpha \in \mathscr{A}$ and such that $\varphi_\alpha(x) \neq \varphi_\alpha(y)$.

We assume that $\mathscr{A}$ is a *directed set* with respect to the natural pre-order: for $\alpha, \beta \in \mathscr{A}$, $\alpha \leq \beta$ whenever $\beta$ refines $\alpha$. Let $\alpha, \beta \in \mathscr{A}$ be such that $\alpha \leq \beta$. Denote $\alpha = \{A_1, \ldots, A_{N(\alpha)}\}$ and $\beta = \{B_1, \ldots, B_{N(\beta)}\}$ and for every $1 \leq i \leq N(\alpha)$, let $J_{\alpha\beta}(i) \subset \{1, \ldots, N(\beta)\}$ be such that $A_i = \bigcup_{j \in J_{\alpha\beta}(i)} B_j$. Equivalently, this can be expressed through a mapping $i_{\alpha\beta} : I(\beta) \rightarrow I(\alpha)$ such that $i_{\alpha\beta}(j) = i$ whenever $j \in J_{\alpha\beta}(i)$. Based on that, define $\varphi_{\alpha\beta} : \mathcal{X}_\beta \rightarrow \mathcal{X}_\alpha$ to be the map that takes $e_j$ into $e_i$ whenever $j \in J_{\alpha\beta}(i)$. Note that $\varphi_{\alpha\beta}(e_j) = e_{i_{\alpha\beta}(j)}$ for all $1 \leq j \leq N(\beta)$. The maps $\varphi_{\alpha\beta}$ are coherent (in the sense that for $\alpha \leq \beta \leq \gamma$, $\varphi_{\alpha\gamma} \circ \varphi_{\beta\gamma} = \varphi_{\beta\gamma}$) and (since the spaces $\mathcal{X}_\alpha$ are discrete) continuous. The projection maps $\varphi'_\alpha$ form a coherent system, i.e. for all $\alpha, \beta \in \mathscr{A}$, if $\alpha \leq \beta$

$$\varphi'_\alpha \circ \varphi_{\alpha\beta} = \varphi'_\beta.$$

Next note that any $g : \mathcal{X}_\alpha \rightarrow \mathbb{R}$ is continuous: such $g$ lie in the (finite-dimensional) Banach space $C(\mathcal{X}_\alpha)$ of all continuous, real-valued maps with the uniform norm, with continuous dual we denote by $M(\mathcal{X}_\alpha)$, the finite-dimensional space of all finite, signed measures on $\mathcal{X}_\alpha$, which is a Banach space with the total-variational norm. For $\mu \in M(\mathcal{X}_\alpha)$ and $g \in C(\mathcal{X}_\alpha)$, $\int g \, d\mu$ is denoted in bi-linear form $\langle \mu, g \rangle_\alpha$.

Let $\alpha, \beta \in \mathscr{A}$ such that $\alpha \leq \beta$ be given. For any $g \in C(\mathcal{X}_\alpha)$, the map $g \circ \varphi_{\alpha\beta} : \mathcal{X}_\beta \rightarrow \mathbb{R}$ is an element of $C(\mathcal{X}_\beta)$. Because $\varphi_{\alpha\beta}$ is surjective, the induced map $\varphi^*_{\alpha\beta} : C(\mathcal{X}_\alpha) \rightarrow C(\mathcal{X}_\beta)$ is a bounded linear operator with norm equal to one. The transpose map $\varphi_{\alpha\beta} : M(\mathcal{X}_\beta) \rightarrow M(\mathcal{X}_\alpha)$ is defined by,
\[(\varphi_{\alpha\beta}(\mu), g)_{\alpha} = (\mu, \varphi_{\alpha\beta}(g))_{\beta} = (\mu, g \circ \varphi_{\alpha\beta})_{\beta}\]

for all \(\mu \in M(\mathcal{A})\) and \(g \in C(\mathcal{A})\). The linear map \(\varphi_{\alpha\beta}\) is bounded with norm less than or equal to one. Note that if we express \(\mu \in M(\mathcal{A})\) as a vector \((\mu_1, \ldots, \mu_{N(\alpha)})\) in \(\mathbb{R}^{N(\alpha)}\),

\[
(\mu, g \circ \varphi_{\alpha\beta})_{\beta} = \sum_{j \in I(\beta)} \mu_j g(\varphi_{\alpha\beta}(e_j)) = \sum_{j \in I(\beta)} \mu_j \left(\sum_{i \in I(\alpha)} \mu_i \right) g(e_i),
\]

from which one determines that,

\[
\varphi_{\alpha\beta}(\mu)_i = \sum_{j \in I(\beta)} \mu_j,
\]

(D.3)

for all \(1 \leq i \leq N(\alpha)\).

**Definition D.4.** Any collection of Borel measures \(\mu_\alpha \in M(\mathcal{A})\) on a coherent system of topological spaces \((\mathcal{A}_\alpha, \varphi_{\alpha\beta})\), such that for all \(\alpha, \beta \in \mathcal{A}\) with \(\alpha \leq \beta\),

\[
\mu_\alpha = \varphi_{\alpha\beta} \circ \mu_\beta,
\]

is called a (coherent) inverse system of measures \((\mu_\alpha, \varphi_{\alpha\beta})\) on the inverse system of spaces \((\mathcal{A}_\alpha, \varphi_{\alpha\beta})\).

The relations (D.4) express coherency, in the same role as the Kolmogorov consistency relations that play a role in the proof of theorem 6.3. But here, we consider (D.4) from the dual point of view: for all \(\alpha, \beta, \gamma \in \mathcal{A}\) such that \(\alpha \leq \beta \leq \gamma\), \(\varphi_{\alpha\beta} \circ \varphi_{\beta\gamma} = \varphi_{\alpha\gamma}\) and \(\varphi_{\alpha\alpha}\) is the identity on \(M(\mathcal{A}_\alpha)\), so \((M(\mathcal{A}_\alpha), \varphi_{\alpha\beta})\) is an inverse system of (non-empty, compact) topological spaces, with (non-empty, compact) inverse limit \(N\). And, again, all \(\mu \in M(\mathcal{A})\) map to points in \(N\), but not all points in \(N\) correspond to bounded regular measures on \(\mathcal{A}\). To relate \(N\) and \(M(\mathcal{A})\) directly, consider for \(\alpha \in \mathcal{A}\), with \(\alpha = (A_1, \ldots, A_{N(\alpha)})\), the mapping \(\Phi_{\alpha}: M(\mathcal{A}) \rightarrow M(\mathcal{A}_\alpha)\),

\[
\Phi_{\alpha}(\mu) = (\mu(A_1), \ldots, \mu(A_{N(\alpha)})),
\]

(D.5)

which form a coherent system. We are now in a position to formulate the following theorem (see [45], Ch. IX, § 4, No. 2).

**Theorem D.3.** (Bourbaki-Prokhorov) Let \((\mathcal{A}_\alpha, \varphi_{\alpha\beta})\) be an inverse system of topological spaces, indexed by \(\mathcal{A}\), \(T\) a topological space and \(\varphi_\alpha: T \rightarrow \mathcal{A}\) a coherent and separating family of continuous mappings. Finally, let \((\mu_\alpha, \varphi_{\alpha\beta})\) be an inverse system of positive measures on \((\mathcal{A}_\alpha, \varphi_{\alpha\beta})\). For there to exist a bounded regular measure \(\mu\) on \(T\) such that \(\Phi_{\alpha}(\mu) = \mu_\alpha\) for all \(\alpha \in \mathcal{A}\), it is necessary and sufficient that the following condition is satisfied:
(P) for every \( \varepsilon > 0 \), there is a compact \( K \subset T \) such that \( \mu_\alpha(\mathcal{F}_\alpha \setminus \phi_\alpha(K)) \leq \varepsilon \) for all \( \alpha \in \mathcal{A} \).

When this is so, the measure \( \mu \) is uniquely determined and
\[
\mu(L) = \inf \{ \mu_\alpha(\phi_\alpha(L)) : \alpha \in \mathcal{A} \},
\]
for every compact set \( L \) in \( T \).

The most direct application of this theorem to the inverse systems at hand is through the measure-theoretic approach.

Example D.2. Consider the case where \( T = M(\mathcal{F}) \) with topology \( \mathcal{T}_1 \), that is, \( \mu \to \nu \) if \( \mu f \to \nu f \) for every bounded, measurable \( f : \mathcal{F} \to \mathbb{R} \). Let \( \mathcal{A} \) denote the directed set of all finite measurable partitions of \( \mathcal{X} \). Then the maps \( \phi_\alpha \) that project \( M(\mathcal{F}) \) onto the \( M(\mathcal{X}_\alpha) \) are continuous (and separating) and they form a coherent system. If we require specified marginals \( \Pi_\alpha \) to form a coherent system of probability measures, we can say the following.

**Proposition D.1.** There exists a bounded regular measure \( \Pi \) on \( M(\mathcal{F}) \) with marginals \( \Pi_\alpha \) for all \( \alpha \), if and only if,
\[
\text{for every } \varepsilon > 0, \text{ there is a } \mathcal{T}_1\text{-compact } K \text{ in } M(\mathcal{F}) \text{ such that } \mu_\alpha(\mathcal{F}_\alpha \setminus \phi_\alpha(K)) \leq \varepsilon \text{ for all } \alpha \in \mathcal{A}.
\]

But now the problem becomes apparent: this type of weak compactness is the domain of the Dunford-Pettis theorem, which is formulated in the context of dominated models only, or must be extended to the context of the \( L^p \)- and \( M \)-spaces that feature centrally in Le Cam’s perspective [162]. The former option appears difficult to formulate from the point of view of an inverse system of priors, unless one considers reasonable the assumption that the measures ultimately described in the model are all dominated by a single probability measure \( Q \) and even then, that road leads to the type of tautology we have seen earlier: the requirement becomes hard to formulate but would look something like:
\[
\text{for every } \varepsilon, \delta > 0, \text{ there is an } M > 0 \text{ such that,}
\]
\[
\Pi_\alpha \left( \phi_\alpha(\mu \in M(\mathcal{F}) : \int_{(d\mu/dQ) > M} \frac{d\mu}{dQ} dQ < \delta) \right) \leq \varepsilon,
\]
for all \( \alpha \in \mathcal{A} \).

which is difficult to verify from the point of view that only an inverse system of marginal measures is available. Somewhat more generally, we may assume that there exists a probability measure \( Q \) and bounded linear operators \( T_\alpha : L^1(Q) \to \mathcal{P} \) with compact \( K_\alpha \subset L^1(Q) \) for all \( \alpha \in \mathcal{A} \) such that,
\[
\Pi_\alpha(\mathcal{F}_\alpha \setminus T_\alpha(K_\alpha)) \geq \delta), \rightarrow 0,
\]
but things are not improving, the above amounts basically to the same unverifiable condition from the point of view that only an inverse system of marginal measures.
is available. Not only do we see that the inclusion of a large $A$ (like that of all finite, measurable partitions) implies that the compactness condition is to be verified for a large set of projections, compacta do not organise with the projections in a harmonious way, as expressed by the explicit appearance of the rather difficult map $\hat{\phi}_\mu$ above.

In what follows, we apply theorem D.3 with Prokhorov’s weak topology and a zero-dimensional version $\hat{X}$ of the sample-space $T = M(\mathcal{X})$, which leads to a more controllable condition.

D.4 Partitions and inverse limit sample-spaces

Let $\mathcal{X}$ be a topological sample-space; ordinarily, $\mathcal{X}$ is spaces like $\mathbb{R}$ or $\mathbb{R}^d$, but functional data [190] takes its values in spaces of functions, curves, manifolds or distributions (in principle, and in some sampled form in practice), which we assume to be Polish (or, with more abstract conditions and assertions, Hausdorff completely regular). Dirichlet and other species-sampling priors [140] have been formulated for functional data as well, and the matter of Bochner-Kolmogorov existence has been raised (see Petrone, Guindani and Gelfand (2009) [186]).

Here, we start from the assumption that $\mathcal{X}$ is a Hausdorff completely regular space with basis $\mathcal{U}$. We define $\mathcal{A}$ to be a collection of partitions generated by the basis $\mathcal{U}$.

**Definition D.5.** A collection of partitions $\mathcal{A}$ of a Hausdorff completely regular space $\mathcal{X}$ is said to resolve $\mathcal{X}$ if, for every open $U \subset \mathcal{X}$, there exists a $\beta \in \mathcal{A}$ and an element $A \in \beta$, such that $A \subset U$.

**Example D.3.** In the special case of a Polish (or rather: separable, metrizable) sample-space $\mathcal{X}$, there exists a countable basis $\mathcal{U}$. Consider a sequence of partitions $(\alpha_n)$ constructed as follows. Let the sequence $(U_n)$ enumerate $\mathcal{U}$, set $\alpha_1 = \mathcal{X}$ and define recursively,

$$\alpha_{n+1} = \{A \cup U_n, A \setminus U_n : A \in \alpha_n, U_n \in \mathcal{U}, A \in \alpha_n\},$$

for all $n \geq 1$. Note that for all $x, y \in \mathcal{X}, x \neq y$, there is an $n \geq 1$ with $x$ and $y$ in distinct subsets of $\alpha_n$ (because $\mathcal{X}$ is Hausdorff) and for every open $U \subset \mathcal{X}$, there exist an $n \geq 1$ and an element $A \in \alpha_n$, such that $A \subset U$ (because $\mathcal{U}$ is a basis) above.

More generally (but still in the special case of a separable, metrizable sample-space $\mathcal{X}$), consider sequences $(\alpha_n)$ of finite partitions of $\mathcal{X}$, based on a countable basis $\mathcal{U}$ for $\mathcal{X}$ with the following properties: for every $n \geq 1$,

(i) $\alpha_n$ results from finite intersections of $\mathcal{X}$ with $U$ and $\mathcal{X} \setminus U$, $(U \in \mathcal{U})$,
(ii) $\alpha_{n+1}$ refines $\alpha_n$,

and, the sequence $(\alpha_n)$ has tails such that,

(iii) for every open $U \subset \mathcal{X}$, there is an $n \geq 1$ and an $A \in \alpha_n$, such that $A \subset U$. 
Example D.4. Consider $\mathcal{X} = \mathbb{R}$ with (countable, dense) subset $\mathbb{Q}$ of centre points, a (countable, dense) set $R = \{ r \in \mathbb{Q} : r > 0 \}$ of radii, and the (countable, dense) collection $\mathcal{U}$ of open intervals $\{ (q - r, q + r) : q \in \mathbb{Q}, r \in R \}$ (which is a basis for the norm topology on $\mathbb{R}$). Obviously, it is possible to apply the construction of the previous example, leading to finite partitions that consist of (open, closed and half-open) intervals only.

A different, more organised choice for the partitions $(\alpha_n)$ is described as follows: choose two sequences of integers $(m_n), (k_n)$ with $m_n \geq 0$ and $k_n \geq 1$, define $q_1 = 1$, $\delta_1 = 1$ to start the recursion,

$$q_{n+1} - q_n = m_n \delta_n, \quad \delta_n = k_n \delta_{n+1},$$

and consider the partition $\alpha_n$ with elements,

$$A_{0,n} = (-\infty, -q_n] \cup [q_n, \infty), \quad A_{1,n} = (-q_n, -q_n + \delta_n),$$

$$A_{1,n} = [-q_n + (l-1)\delta_n, -q_n + l\delta_n),$$

with $l = 1, \ldots, N_n$, and $N_1 = 3$, $N_{n+1} = (N_n - 1) + 2m_n k_n + 1$. The resulting $\alpha_n$ are partitions of $\mathbb{R}$ generated by $\mathcal{U}$, which resolve $\mathbb{R}$ iff $\limsup_n m_n > 1$ and $\limsup_n k_n > 1$. If $\lim_n m_n = 0$ but $\limsup_n k_n > 2$, the interval $(-q, q)$ (with $q = \lim_n q_n = \sum_n m_n < \infty$) is resolved by the ‘subset partitions’ $\alpha'_n = \{ A_{1,n} : 1 \leq l \leq N_n \}$ that cover $(-q, q)$ for large enough $n$. (If $\lim_n k_n = 1$, the resulting sequence of partitions does not resolve $\mathbb{R}$ (nor $(-q, q)$).) This construction relates directly to branching processes and stick-breaking processes, and associates directly with the Cantor set. Examples: $m_n = 0$ and $k_n = 2$ for all $n \geq 1$ leads to the dyadic partitions of $(0,1)$ that underpin the Pólya-urn process, while the choice $m_n = 1$ and $k_n \geq 2$ gives partitions $\alpha_n$ of $\mathbb{R}$ that refine geometrically on the intervals $[-n, n]$.

Through the $\varphi_{\alpha\beta}$, the spaces $\mathcal{X}_\alpha$ are organized into an inverse system of topological spaces, $(\mathcal{X}_\alpha, \varphi_{\alpha\beta})$. Since each of the $\mathcal{X}_\alpha$ is non-empty and compact, the inverse limit $Y$ is non-empty and compact. For any $\alpha \in \mathcal{A}$, denote by $\varphi_\alpha$ the continuous projection that takes $Y$ into $\mathcal{X}_\alpha$. The maps $(\varphi_\alpha : \alpha \in \mathcal{A})$ form a coherent family, but it is not the only one: for any $\alpha = (A_1, \ldots, A_{N(\alpha)})$, $\mathcal{X}$ is mapped surjectively to $\mathcal{X}_\alpha$ by $\varphi_\alpha$ and for all $\alpha, \beta \in \mathcal{A}$ with $\alpha \leq \beta$, $\varphi_{\alpha\beta} \circ \varphi'_\beta = \varphi'_\alpha$. For every $x \in \mathcal{X}$, $(\varphi'_{\alpha}(x) : \alpha \in \mathcal{A}) \in \prod_{\alpha \in \mathcal{A}} \varphi'_\alpha(\mathcal{X}) = \prod_{\alpha \in \mathcal{A}} \mathcal{X}_\alpha$, is a point in the inverse limit space $Y$. This defines a mapping $\varphi' : \mathcal{X} \to Y$ that is neither injective nor surjective in general.

Proposition D.2. Let $\mathcal{X}$ be a Hausdorff space with a directed set of finite partitions $\mathcal{A}$ that resolves $\mathcal{X}$ and let $(\mathcal{X}_\alpha, \varphi_{\alpha\beta})$ denote the associated inverse system. The inverse limit $Y$ is a zero-dimensional compact Hausdorff space that contains a subspace $\tilde{\mathcal{X}}$ with a continuous bijection $i : \tilde{\mathcal{X}} \to \mathcal{X}$ as well as a coherent family of continuous surjective maps $\varphi_\alpha : \tilde{\mathcal{X}} \to \mathcal{X}_\alpha$ such that for all $\alpha \in \mathcal{A}$,

$$\begin{align*}
\tilde{\mathcal{X}} &\xrightarrow{i} \mathcal{X} \\
\varphi_{\alpha} &\downarrow \mathcal{X}_\alpha
\end{align*} \quad \tag{D.6}$$
is a commutative diagram.

Proof. The inverse system \((X_\alpha, \varphi_{\alpha\beta})\) consists of compact, non-empty spaces \(X_\alpha\), so the inverse limit \(Y\) is non-empty and compact. Since \(\mathcal{A}\) is a directed set, a basis for the topology on \(Y\) is given by the collection of all sets \(\varphi_{\alpha}^{-1}(A)\), where \(\alpha \in \mathcal{A}\) and \(A\) any subset of \(X_\alpha\). For any two distinct \(x, y \in Y\), there exists an \(\alpha\) such that \(\varphi_{\alpha}(x) \neq \varphi_{\alpha}(y)\). Conclude that, since \(X_\alpha\) is Hausdorff, \(Y\) is Hausdorff. Since any \(A \subset X_\alpha\) is clopen in \(X_\alpha\) and \(\varphi_{\alpha}: Y \to X_\alpha\) is continuous, \(\varphi_{\alpha}^{-1}(A)\) is clopen in \(Y\) for any \(\alpha\) and \(A\). With a clopen neighbourhood basis for any \(y \in Y\), \(Y\) is a zero-dimensional (see [134], ch. 7, sec. 7, and [43], Ch. IX, §6, No. 4, ) space. Note that if \(x, y \in X\), \(x \neq y\), the Hausdorff property of \(X\) and the assumption that \(\mathcal{A}\) is a directed set of partitions that resolve \(X\), imply that there exists an \(\alpha \in \mathcal{A}\) with distinct \(A_\alpha, A_\beta \in \alpha\) such that \(x \in A_\alpha\) and \(y \in A_\beta\). Therefore, the map \(\varphi': X' \to Y\) is injective, and has an inverse defined on \(\hat{X} = \varphi'(X')\) which we denote by \(i: \hat{X} \to X\). Note that the maps \(\varphi_{\alpha}: \hat{X} \to X_\alpha\), defined as the restrictions to \(\hat{X}\) of the maps \(\varphi_{\alpha}: Y \to X_\alpha\), are continuous. Because \(i^{-1}(x) = \varphi'(x) = (\varphi_{\alpha}'(x) : \alpha \in \mathcal{A}) \in \hat{X} \subset \prod_{\alpha} X_\alpha\) and \(\varphi_{\alpha}\) is the restriction of \(\varphi_{\alpha} : \prod_{\alpha} X_\alpha \to X_\alpha\) to \(\hat{X}\), the relation \(\varphi_{\alpha} \circ i^{-1} = \varphi_{\alpha}'\) is immediate. Composition with \(i\) shows that diagram (D.6) is commutative. Surjectivity of \(i\) and \(\varphi_{\alpha}'\) implies surjectivity of \(\varphi_{\alpha}\). As for the continuity of \(i\), let \((x_j : j \in J)\) denote a net in \(\hat{X}\), converging to \(y \in \hat{X}\). Denote \(z = i(y) \in X\) and \(U\) an open neighbourhood of \(z\). By assumption, there exists a \(\beta \in \mathcal{A}\) and an element \(A \in \beta\), such that \(A \subset U\). Since \(x_j \to y\), \(x_j\alpha \to y_\alpha\) in \(X_\alpha\) for all \(\alpha \in \mathcal{A}\), so in particular, \(x_j\beta \to y_\beta\). Because \(X_\beta\) is discrete, there exists a \(J \subset J\) such that \(x_j\beta = y_\beta = (0, \ldots, 0, 1, 0, \ldots, 0)\) for all \(j \geq J\) (with the one corresponding to the element \(A \in \beta\)). Therefore, \(i(x_j) \in U\) for all \(j \geq J\), which proves that \(i\) is continuous.

Note that \(\hat{X}\) is Hausdorff completely regular (see [43], Ch. IX, §1, No. 5, proposition 3). If we assume \(\hat{X}\) is not itself zero-dimensional, it is not possible for the space \(\hat{X}\) to be closed as a subset of \(Y\). (For, if we assume that \(\hat{X}\) is closed, \(\hat{X}\) is compact and Hausdorff, so \(i\) has a continuous inverse. However, \(\hat{X}\) is zero-dimensional while \(X\) is not.) Therefore it is possible that there exist subsets that are compact in \(X\) and not compact in \(\hat{X}\). (To appreciate the prevalence of this phenomenon, see example D.5.)

The essential difference between the maps \(\varphi_{\alpha}\) and \(\varphi_{\alpha}'\) in diagram (D.6) is that the former are continuous (see, in particular, the conditions of theorem D.3). The space \(\hat{X}\) has a topology that is a minimal, zero-dimensional refinement of \(X\) that enables continuity of the coherent system. It’s illustrative to consider a space \(\hat{X}\) which is itself zero-dimensional. In that case, it is possible to let \(\mathcal{A}\) consist of finite partitions of clopen sets \(A \subset X\). Consequently, the maps \(\varphi_{\alpha}'\) form a coherent system of continuous maps, which means they can play the role of the maps \(\varphi_{\alpha}\) in subsequent steps.
D.5 Open and clopen topological bases

Let \( X \) be a Hausdorff completely regular space with topology \( T \) and basis \( \mathcal{U} \) (with the special example of a Polish space with countable basis \( \mathcal{U} \) also in mind). First we refine \( T \) on the set \( X \), defining a topological space \( \hat{X} \): with the same underlying set \( X \), through a sub-basis,

\[
\mathcal{J} = \{ U, X \setminus U : U \in \mathcal{U} \},
\]

for the refined topology \( \hat{T} \) in which each basis element \( U \in \mathcal{U} \) is clopen.

**Proposition D.3.** The space \( \hat{X} \) is zero-dimensional and the identity map \( i : \hat{X} \rightarrow X \) is continuous. If \( X \) is Polish, so is \( \hat{X} \).

**Proof.** The sub-basis \( \mathcal{J} \) gives rise to a basis consisting of clopen sets, so \( \hat{X} \) is zero-dimensional. Furthermore the identity \( i \) is continuous because \( T \) refines \( T \). Assuming \( X \) is Polish, the countable product space \( X^\mathbb{N} = \prod_{n \geq 1} X \) is Polish (see [134], prop. 3.3) and has a diagonal \( \Delta \) that is closed and homeomorphic to \( X \). Enumerate \( \mathcal{U} = \{ U_n : n \geq 1 \} \) and define \( X_n \) to be the topological sum of \( U \) and \( X \setminus U \), which is Polish since both \( U \) and \( X \setminus U \) are Polish. Since the underlying set for \( X_n \) is the set-theoretic union of \( U \) and its complement, there is a canonical set-theoretic identification \( i_n : X_n \rightarrow X \), which is continuous, and we view \( X_n \) as \( X \) with \( U_n \) made clopen. The product space \( \prod_{n \geq 1} X_n \) is Polish and the map \( j : \prod_{n \geq 1} X_n \rightarrow X^\mathbb{N} \) is continuous. Then \( j^{-1}(\Delta) \) is homeomorphic to \( \hat{X} \) and (closed in Polish \( X^\mathbb{N} \), therefore) Polish.

Alternatively the zero-dimensional space \( \hat{X} \) can be constructed using the Stone space of all ultra-filters of the Boolean algebra \( \mathcal{B}(\mathcal{U}) \) generated by \( \mathcal{U} \). Both representations can play the role of the space \( \hat{X} \) in the context of theorem D.3, rendering the verification of compactness in \( \hat{X} \) more feasible.

**Proposition D.4.** Let \( X \) be a Hausdorff completely regular space. The Borel sets on \( X \) and \( \hat{X} \) are equal and any bounded set function \( \mu \) is a Borel measure on \( X \) if and only if \( \mu \) is a Borel measure on \( \hat{X} \).

**Proof.** Note that the Borel \( \sigma \)-algebra on \( \hat{X} \) generated by the basis \( \mathcal{U} \) is the same \( \sigma \)-algebra as the \( \sigma \)-algebra generated by \( \mathcal{U} \) and complements, which form a basis for \( \hat{X} \): conclude that \( \mathcal{J} \) and \( \mathcal{U} \) have the same Borel sets.

As was noted, regularity of the Borel measure \( \mu \) on \( \hat{X} \) implies regularity on \( X \), but the converse may fail because, in general, \( \mathcal{U} \) has compact sets that are not compact in \( \hat{X} \): consider a typical compact subset of \( \mathbb{R} \).

**Example D.5.** Take \( \mathcal{U} = \mathbb{R} \) in its usual topology and consider the interval \( [0, 1] \). In the zero-dimensional \( \hat{X} \), the interval \( [0, 1] \) has an infinite open cover that cannot be reduced to finite: namely, take,

\[
A_0 = (-\infty, 0] \cup [1, \infty), \quad A_k = \left( \frac{1}{k}, 1 - \frac{1}{k} \right),
\]
where \( k \geq 1 \). All of these sets \( A_k, k \geq 0 \) belong to the basis of \( \mathcal{K} \), so they are open, and any finite subset does not cover \([0,1]\).

Zero-dimensional compact subsets of Polish spaces are characterized as follows.

**Theorem D.4.** (Brouwer) Any zero-dimensional, compact subset \( K \) of a Polish space \( X \) is the disjoint union of a countable open subspace \( V \) and a subspace \( C \) homeomorphic to the Cantor set. \( V \) consists of isolated points; its complement \( C \) is perfect, compact and nowhere dense in \( X \).

**Proof.** See Kechris (1994) [134], section 7.

In the application of theorem D.3 with \( T = \mathcal{K} \), property \( (P) \) requires \( K \) to be compact in \( \mathcal{K} \). Zero-dimensionality implies that \( K = V \cup C \) and \( C \) is also nowhere dense as a subset of \( \mathcal{K} \), explaining example D.5. Indeed, any zero-dimensional compact \( K \) in \( \mathcal{K} \) is also compact in \( \mathcal{K} \). Implied is the following curious specification of the Radon property for Polish spaces: for every positive, bounded Borel measure \( \mu \) on a Polish space \( \mathcal{K} \) and every \( \varepsilon > 0 \), there exists a zero-dimensional compact \( K \) such that \( \mu(\mathcal{K} \setminus K) < \varepsilon \).

By proposition D.4, any positive, bounded Borel measure \( \mu \) on \( \mathcal{K} \) is also positive, bounded Borel measure \( \mu \) on \( \mathcal{K} \), and with the help of proposition D.3 and the Radon property of Polish spaces, we conclude that bounded regular measures on \( \mathcal{K} \) and \( \mathcal{K} \) are in one-to-one correspondence. To rephrase: any continuous \( f : \mathcal{K} \to \mathbb{R} \) is also continuous when viewed as \( f : \mathcal{K} \to \mathbb{R} \), so there exists a continuous, linear, injective map \( j : C^b(\mathcal{K}) \to C^b(\mathcal{K}) \) of norm one, and dual to that, a linear \( j^* : M(\mathcal{K}) \to M(\mathcal{K}) \) of norm one.

**Proposition D.5.** If \( \mathcal{K} \) is Polish, \( M(\mathcal{K}) \) and \( M(\mathcal{K}) \) are isomorphic as locally convex spaces, each with Prokhorov’s weak topology.

**Proof.** If \( \mu \neq \nu \in M(\mathcal{K}) \), then there exists an \( f \in C^b(\mathcal{K}) \) such that \( \mu f \neq \nu f \), which remains the same if we view \( \mu, \nu \) as Borel measures on \( \mathcal{K} \) and \( f \) as a Borel measurable map on \( \mathcal{K} \to \mathbb{R} \). Conclude that \( j \) is injective. The map \( j \) is a continuous bijection with continuous inverse, if we equip both \( M(\mathcal{K}) \) and \( M(\mathcal{K}) \) with Prokhorov’s weak topologies.

The proposition below notes that the projection maps for finite, clopen partitions \( \alpha \) are continuous (as required in theorem D.3), if we equip \( M(\mathcal{K}) \) with Prokhorov’s weak topology or stronger.

**Proposition D.6.** Suppose \( \mathcal{K} \) is Hausdorff completely regular and endow \( M(\mathcal{K}) \) with Prokhorov’s weak topology. For given \( \alpha \in \mathcal{A} \), assume that all \( A \in \alpha \) are clopen in \( \mathcal{K} \). Then the map \( \phi_{\alpha} : M(\mathcal{K}) \to M(\mathcal{K}_\alpha) \) is continuous.

**Proof.** Fix \( \alpha \in \mathcal{A} \). From (D.5), we see that if we equip \( M(\mathcal{K}) \) with a topology that makes \( \mu \mapsto \mu(A) \) continuous for all \( A \in \alpha \), the map \( \phi_{\alpha} : M(\mathcal{K}) \to M(\mathcal{K}_\alpha) \) is continuous. For any clopen \( A \in \alpha \), denote,

\[
C_A = \{ f \in C(\mathcal{K}), 0 \leq f \leq 1_A \}, \quad C_{\mathcal{K} \setminus A} = \{ f \in C(\mathcal{K}), 0 \leq f \leq 1_{\mathcal{K} \setminus A} \}.
\]
Both $1_A$ and $1_{\mathcal{A}\setminus A}$ are lower-semicontinuous and satisfy, for all $x \in \hat{\mathcal{X}}$,

$$1_A(x) = \sup\{f(x) : f \in C_A\}, \quad 1_{\mathcal{A}\setminus A}(x) = \sup\{f(x) : f \in C_{\mathcal{A}\setminus A}\},$$
due to complete regularity. We have for any $\mu \in M(\hat{\mathcal{X}})$,

$$\sup\{\mu f : f \in C_A\} \leq \mu(A) \leq \inf\{\mu(1-f) : f \in C_{\mathcal{A}\setminus A}\}.$$

Let $\mu \in M(\hat{\mathcal{X}})$ and some clopen $A$ be given. For some $\epsilon > 0$, let $f \in C_A$, $g \in C_{\mathcal{A}\setminus A}$ be such that,

$$\mu g - \epsilon \leq \mu(A) \leq \mu f + \epsilon,$$

and let $U$ denote the following weak neighbourhood of $\mu$,

$$U = \{\nu \in M(\hat{\mathcal{X}}) : |\nu f - \mu f| < \epsilon, |\nu g - \mu g| < \epsilon\}.$$

Then $\nu \in U$ implies $|\mu(A) - \nu(A)| < 2\epsilon$.

### D.6 Inverse limit priors

The goal, here, is to use theorem D.3 again to prove the existence of inverse limit priors $\Pi$. Like before, the spaces $M(\mathcal{X}_a)$ are all compact and we consider the spaces $C(M(\mathcal{X}_a))$ of continuous functions $M(\mathcal{X}_a) \to \mathbb{R}$ and the continuous maps $\varphi_{a\beta} : M(\mathcal{X}_{a\beta}) \to M(\mathcal{X}_a)$ of (D.3), which induce $\varphi_{a\beta}^* : C(M(\mathcal{X}_a)) \to C(M(\mathcal{X}_{a\beta}))$ through $\varphi_{a\beta}^*(f) = f \circ \varphi_{a\beta}$, with transpose $\varphi_{a\beta}^* : M(\mathcal{X}_{a\beta}) \to M(\mathcal{X}_a)$, for all $\alpha, \beta \in \mathcal{A}$, $\alpha \leq \beta$. Like before, $(M(\mathcal{X}_a), \varphi_{a\beta})$ form an inverse system with inverse limit $\mathcal{N}$, projections $\varphi_{a\alpha} : M(\mathcal{X}_a) \to M(\mathcal{X}_a)$ and injective embedding $\hat{M} \subset \mathcal{N}$ of $M(\mathcal{X}_a)$ (with restrictions $\hat{\varphi}_{a\alpha} : \hat{M} \to M(\mathcal{X}_a)$).

If we assume continuity of the projections, $M(\mathcal{X})$ (identified with $\hat{M}$) may play the role of $T$ in theorem D.3: take an inverse system of measure $(\Pi_\alpha, \varphi_{a\alpha})$ on the inverse system $(M(\mathcal{X}_a), \varphi_{a\beta})$ and $T = M(\hat{\mathcal{X}})$, for a space $\mathcal{X}$ that is completely regular. The bounded, regular measure $\Pi$ (with which existence is proved) lies in $M(\mathcal{X})$, where $\mathcal{X}$ is the space obtained in proposition D.3.

**Theorem D.5.** Let $\mathcal{X}$ be Hausdorff completely regular with basis $\mathcal{A}$; let $\mathcal{A}$ consist of partitions generated by $\mathcal{A}$ and resolve $\mathcal{X}$. Denote the corresponding inverse system by $(\mathcal{X}_\alpha, \varphi_{a\beta})$ and by $(\Pi_\alpha, \varphi_{a\beta})$ an inverse system of positive measures on $(M(\mathcal{X}_a), \varphi_{a\beta})$. Endow $M(\hat{\mathcal{X}})$ with Prokhorov’s weak topology. For there to exist a bounded regular measure $\Pi$ on $M(\hat{\mathcal{X}})$ such that $\hat{\varphi}_{a\alpha}(\Pi) = \Pi_\alpha$ for all $\alpha \in \mathcal{A}$, it is necessary and sufficient that the following condition be satisfied:

(P) For every $\epsilon > 0$, there is a compact $H \subset M(\hat{\mathcal{X}})$ such that,

$$\Pi_\alpha(M(\mathcal{X}_a) \setminus \hat{\varphi}_{a\alpha}(H)) < \epsilon,$$
for all $\alpha \in \mathcal{A}$.

When this is so, the measure $\Pi$ is uniquely determined and

$$\Pi(L) = \inf \left\{ \Pi_\alpha(\hat{\phi}_\alpha(L)) : \alpha \in \mathcal{A} \right\},$$

for every compact set $L$ in $M(\hat{\mathcal{X}})$.

**Proof.** For every $A \in \alpha$ in every $\alpha \in \mathcal{A}$, $i^{-1}(A) \subset \hat{\mathcal{X}}$ is a clopen element of the subbasis that defines the topology for $\hat{\mathcal{X}}$. So for all $\alpha \in \mathcal{A}$, $\hat{\phi}_\alpha : M(\hat{\mathcal{X}}) \to M(\mathcal{X}_\alpha)$ is continuous with respect to Prokhorov’s weak topology, by lemma D.6. Moreover, $(\hat{\phi}_\alpha, \phi_{\alpha \beta})$ is a coherent and separating family. The assertion now follows directly from theorem D.3.

This leads to the following double version of Prokhorov’s condition for the most common types of sample spaces.

**Corollary D.1.** Let $\mathcal{X}$ be Polish in theorem D.5. For there to exist a bounded measure $\Pi$ on $M(\mathcal{X})$ such that $\hat{\phi}_\alpha(\Pi) = \Pi_\alpha$ for all $\alpha \in \mathcal{A}$, it is necessary and sufficient that the following condition be satisfied:

\((P')\) For every $\varepsilon > 0$, there exist $\beta \in \mathcal{A}$, $R > 0$ such that $\Pi_\beta(\mu_\beta(\mathcal{X}_\beta) \leq R) > 1 - \varepsilon$ and, for every $\delta > 0$, there is a compact $K \subset \mathcal{X}$ such that,

$$\Pi_\alpha(\{ \mu_\alpha(\mathcal{X}_\alpha \setminus \hat{\phi}_\alpha(K)) > \delta \}) < \varepsilon,$$

for all $\alpha \in \mathcal{A}$.

When this is so, the measure $\Pi$ is uniquely determined and

$$\Pi(L) = \inf \left\{ \Pi_\alpha(\hat{\phi}_\alpha(L)) : \alpha \in \mathcal{A} \right\},$$

for every compact set $L$ in $M(\mathcal{X})$.

**Proof.** Since $\mathcal{X}$ is Polish, $M(\mathcal{X})$ is identified with $M(\hat{\mathcal{X}})$ according to proposition D.5, and c.f. theorems C.2 and C.3, $H \subset M(\mathcal{X})$ is compact iff $H$ is closed,

$$\sup\{ \| \mu \| : \mu \in H \} < \infty,$$

and for every $\delta > 0$, there exists a compact $K$ in $\mathcal{X}$ such that for all $\mu \in H$,

$$|\mu(\mathcal{X} \setminus K)| \leq \delta.$$

The set $H_1$,

$$H_1 = \{ \mu \in M(\mathcal{X}) : \hat{\phi}_\alpha(\mu)(\mathcal{X}_\alpha \setminus \hat{\phi}_\alpha(K)) \leq \delta \},$$

is closed in $M(\mathcal{X})$, because of proposition D.5 and the fact that $\mathcal{X} \setminus \hat{\phi}_\alpha(K)$ is clopen, c.f. lemma D.6. Because $\mathcal{X}_\beta$ is clopen, the set $H_2 = \{ \mu \in M(\mathcal{X}) : \mu_\beta(\mathcal{X}_\beta) \leq L \}$ is closed. Since $\mu_\alpha(\mathcal{X}_\alpha) = \mu_\beta(\mathcal{X}_\beta)$ for all $\alpha$, c.f. (D.4), the intersection $H = H_1 \cap H_2$ is compact in $M(\mathcal{X})$ and satisfies property $(P')$. 


It is noted that property \((P'')\) is also \textit{sufficient} in the situation where \(\mathcal{X}\) is completely regular and we require that,

\((P'')\) for every \(\varepsilon, \delta > 0\), there is a compact \(K \subset \hat{\mathcal{X}}\) such that,

\[
\Pi_\alpha \left( \{ \mu_\alpha \in M(\hat{\mathcal{X}}_\alpha) : \mu_\alpha(\mathcal{X}_\alpha \setminus \hat{\phi}_\alpha(K)) > \delta \} \right) < \varepsilon,
\]

for all \(\alpha \in A\).

Note also that corollary D.1 can be read with the following consequence.

**Corollary D.2.** Let \(\mathcal{Y}\) be Polish in theorem D.5 with \(\mathcal{A}\) the partitions generated by a countable basis \(\mathcal{U}\). For any bounded measure \(\Pi\) on \(M(\mathcal{X})\) and any \(\varepsilon, \delta > 0\), there is a compact \(K \subset \hat{\mathcal{X}}\) such that,

\[
\Pi_\alpha \left( \{ \mu_\alpha \in M(\mathcal{Y}_\alpha) : \mu_\alpha(\mathcal{Y}_\alpha \setminus \hat{\phi}_\alpha(K)) > \delta \} \right) < \varepsilon,
\]

for all \(\alpha \in A\).

With reference to Brouwer’s theorem D.4, the point of this complication is that any random probability measure on a Polish space places arbitrarily mass arbitrarily close to one on fixed zero-dimensional compact sets. The suggestion is that there exists a relation with the \textit{discreteness} \(\Pi\)-almost-all realizations of the Dirichlet random probability measure \(\text{c.f. lemma 6.4}\), which is perhaps somewhat surprising without the zero-dimensional perspective on existence of inverse limit measures.
References

References

40. L. BOLTZMANN, Vorlesungen über Gastheorie, (2 Volumes), Leipzig (1895, 1898).
71. J. Diestel, Uniform integrability: an introduction, Lectures presented at the School of Measure theory and Real Analysis, Grado, Italy (1991).
References


130. S. Kakutani, Concrete representation of abstract (L)-spaces and the mean ergodic theorem, Annals of Mathematics 42 (1941), 523–537.


References

161. L. Le Cam, *An inequality concerning Bayes estimates*, University of California, Berkeley (197X), unpublished.
184. P. ORBANZ, Conjugate projective limits, arxiv:1012.0363 [math.ST]
Index

$C^b(T)$, 321
$L^1(\mu)$, 150
$L^\infty(\mu)$, 150
$R$-better, 56
$\sigma$-algebra, 299
Borel, 300
$p$-value, 44
., 90
t.i.d., 3

action, 55
additivity, 319
admissible, 56
almost-everywhere, 301
almost-surely, 301
alternative, see alternative hypothesis
approximation
asymptotic, 103
asymptotically linearity, 110
attenuation bias, 244

barycentre, 177
basis
  topological, 321, 326
Bayes factor, 51
Bayes’s billiard, 22
Bayes’s Rule, 19
belief, 14
best-regular
  ML estimator, 110
bi-polar theorem, 165
bias, 88, 106
bias correction, 88
bootstrap, 74

central limit theorem, 46
classification, 61
classifier, 61
clopen, 329
clopen set, 328
clustering methods, 86
clusters, 22
coherent family, 312, 327
coherent system, 319
complement, 299
completion, 32
condition
  Prokhorov, 320
conditional distribution, 18, 309
  regular, 172, 309
conditional expectation, 308
conditional independence, 27
conditional probability, 308
confidence level, see level, confidence
  asymptotic, 113
confidence set, 39
  asymptotic, 40, 103, 105
  asymptotically consistent, 40
Wald-type, 106, 125
confidence sets
  efficient, 103
conjugate family, 91, 156
consistency, 15, 87, 103, 145
consistency conditions, 307
consistent, 146
contiguity, 158
  remote, 144
continuity theorem, 301
convergence in total variation, 312
convergence of experiments, see limits of experiments
convex cover, 167
convex hull
Index

closed, 32
convolution, 21
counting measure, 5, 7, 301
coupling, 152, 176, 306
covering number, 246
credible interval, 42
credible region, 42
credible set, 41, 103, 113
asymptotic, 42
HPD–, 42
critical set, 44
cross-validation, 84
cylinderset, 156
data, 3
categorical, 3
functional, 3
interval, 3
nominal, 3
ordinal, 3
ranked, 3
ratio, 3
data distribution
Bayesian, 19
frequentist, 4, 27
data-tracking, 172
decision, 55
decision principle
minimax, 56
decision rule, 55
Bayes, 59
minimax, 56
randomised, 57
decision space, 55
decision theory, 55
density, see probability density 306
dependent data, 173
differentiable
in quadratic mean, 108
direct limit, 314
directed set, 323
Dirichlet distribution, 96
Dirichlet family, 97
Dirichlet process prior, 152
disintegration, 19
disjointness, 299
distribution
posterior, see posterior distribution 12
posterior predictive, 12
random, 194
sampling, 30
unimodal, 31
domination, 301
DQM, see differentiable in quadratic mean 108
duality, 165
efficiency, 90, 103, 104
empirical Bayes, 83
empty set, 299
entropy
Lindley, 79
Shannon, 79
error distribution, 244
estimate, 8
estimator, see point-estimator 8
M–, 38
formal Bayes, 35
James-Stein, 90
MAP, 37, 112
maximum-a-posteriori, 37
maximum-likelihood, 10, 15
minimax, 58
non-parametric ML, 157
penalized maximum-likelihood, 37
shrinkage, 106
small-ball, 36
unbiased, 100
estimator maximum likelihood, 124
exchangability, 302
exchangeability, 29
expectation
empirical, 9
expected loss, 35
exponential family, 92
canonical representation, 93
of full rank, 93
feature vector, 61
finite additivity, 152
formal Bayes estimators, 150
functional data analysis, 326
Hölder space, 150
hyperparameter, 81
hyperprior, 73, 81
hypothesis, 43
composite, 43
simple, 43
identifiability, 5
under P, 124
inadmissible, 56
inconsistency, 88
independence, 302
inductive limit, 314
inference, 54
infinite divisibility, 97
information criterion
nuisance, 9
of interest, 9
parameter space, 5
partition, 299
  generated by basis, 321
point-estimator, 8
pointwise convergence, 311
Polish space, 309, 322
Portmanteau lemma, 311
posterior, 19
posterior consistency, 148
  almost-surely, 148
  at a point, 148
Bayesian, 148, 150
posterior convergence
  rate of, 161
posterior distribution, 12
posterior mean, 33
posterior median, 34
posterior merging
  strong, 182
  weak, 182
posterior mode, 37
power function, 44
power sequence, 48
power-set, 7
powerset, 299, 301
pre-compact, 165
pre-image, 303
prediction, 12
predictive distribution
  posterior, 19, 31
  prior, 19
preferred
  Bayes, 59
  minimax, 56
prior, 12, 18
  conjugate, 91
  Dirichlet process, 21
  improper, 75
  informative, 70
inverse limit, 196
Jeffreys, 77
  non-informative, 74
objective, 74
reference, 79
subjective, 70
prior distribution
  conditional, 72
prior mass
  lower bound, 158, 168
  upper bound, 178
probability density, 305, 306
probability density function, 4
probability measure
  random, 319
probability space, 301
process
  branching, 327
  stick-breaking, 327
profile likelihood, 185
property
  \((K1)\), 307
  \((K2)\), 307
  \((P''')\), 333
  \((P'')\), 332
  \((P')\), 331
  \((P)\), 325
  \((R)\), 321
  Cauchy, 313
double Prokhorov, 322
Prokhorov, 315
Radon, 330
Radon-Nikodym derivative, 306
random variable, 304
randomization, 46
randomized test, 46
rate
  uniform testing, 50
rate of convergence, 9, 15, 147
reference prior, 79
regression error, 243
regression function, 244
regularity, 309
regularization, 86
representation theorem
  Riesz-Markov-Kakutani, 321
residual, 26
risk
  Bayes, 59
  minimax, 56
risk function
  Bayesian, 35, 59
sample splitting, 84
sample-average, 9
samplespace, 3, 55
sampling distribution, 40
Scheffé’s lemma, 312
second countable, 25
separating functions, 323
separating partitions, 323
separation
  uniform, 50
set, 299
set difference, 299
set-function, 300
σ-additive, 300
σ-finite, 301
additive, 300
countably additive, 300
finite, 301
shrinkage estimation, 90
sieve, 163
sieve prior, 267, 271
sigma-algebra, 3, 15
generated, 300, 303
significance level, see level, significance, 44
simple function, 304
simplex, 7
singularity, 301
Sobolev space, 150
Souslin space, 309
state, 55
state space, 55
statistic, 8, 39
statistical decision theory, see decision theory, 55, 55
stochastic process, 152, 306
studentization, 41
sub-basis
topological, 329
sub-sample, 84
subset, 299
superefficiency, 90, 106
support, 25
symmetric testing, 44, 48, 51
test
asymptotic, 46
asymptotically more powerful, 49
likelihood ratio, 49
minimax optimal, 50
uniformly asymptotically most powerful, 49
uniformly most powerful, 45
test function, 46, 144
test statistic, 43, 44
test-statistic, 44
testing power
uniform, 50
theorem
Arzelà-Ascoli, 247
Bernstein-von Mises, 42
Brouwer, 330
central limit, 9
De Finetti’s, 303
dominated convergence, 305
Fubini’s, 305
Glivenko-Cantelli, 9
minimax, 56, 166
monotone class, 303
monotone convergence, 304
Radon-Nikodym, 305
topology
norm (on R), 327
zero-dimensional, 328
type-I error, 45
type-II error, 45
unbiased inference, 70
uncertainty quantification, 103
uniformly tight, 315
union, 299
Urysohn’s metrization theorem, 149
version
posterior, 20
weak convergence, 311
weak topology
Prokhorov’s, 182
weakly bounded, 165
Weierstrass theorem, 268
zero-one law, 302