Bayesian Statistics

Bas Kleijn, KdV Institute for Mathematics
Bayesian philosophy

Bayesian school of statistics differs from the Frequentist school.

Bayesians have a different perspective on data and models. In particular, no true, underlying distribution $P_0$ of the data.

Bayesians have a belief concerning the mechanism that generates the data. The data itself is used to correct this belief.

Mathematically, Belief is represented by a prior probability measure $\Pi$ on the model. The data $X_1, \ldots, X_n$ is incorporated by conditioning, resulting in a posterior $\Pi(\cdot|X_1,\ldots,X_n)$ probability measure on the model.
Motivating example (Savage, 1961)

Example 3.1 Consider the following three statistical experiments:

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 is correct every time.

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.

A drunken friend says that he can predict heads or tails of a fair coin-flip. In ten trials, he is right every time.
Frequentist analysis

We analyse the Bayesian procedure from a frequentist perspective.

**Assumption** sample $X_1, \ldots, X_n$ i.i.d. $P_0$-distributed

We shall concentrate on the **large-sample behaviour of the posterior**.

Typical questions

- **Consistency** Does the posterior concentrate in the point $P_0 \in \mathcal{P}$
- **Rate of convergence** How fast does concentration occur?
- **Limiting shape** Which shape does a concentrating posterior have?
- **Asymptotic testing** Is the Bayes factor consistent?

in the limit $n \to \infty$. 
Goal

The question

Given the model, which priors give rise to posteriors with good frequentist convergence properties?

The answer

To formulate theorems that assert asymptotic properties of the posterior, under conditions on the prior and the model.
Course schedule

Lecture 1  **Bayesian Basics**  
Bayesian formalism, estimation, coverage, testing

Lecture 2  **The Bernstein-von Mises theorem**  
Limit shape in smooth parametric models, semi-parametrics

Lecture 3  **Bayes and the Infinite**  
Consistency, Doob’s theorem, Schwartz’s theorem

Lecture 4  **More posterior consistency**  
Barron’s, Walker, Ghosh-Ghosal-van der Vaart

Lecture 5  **Remote contiguity and Bayes factors**  
Consistency with non-\(i.i.d\). data, testing of hypotheses
References


S. Ghosal, A. van der Vaart, *Foundations of Bayesian statistics*, (unpublished) (201?).

In the first lecture, the basic formalism of Bayesian statistics is introduced and its formulation as a frequentist method of inference is given. We discuss such notions as the prior and posterior, Bayesian point estimators like the posterior mean and MAP estimators, credible intervals, odds ratios and Bayes factors. All of these are compared to more common frequentist inferential tools, like the MLE, confidence sets and Neyman-Pearson tests.
Bayesian and Frequentist statistics

**sample space** \((X, \mathcal{B})\)  
**measurable space**

**i.i.d. data** \(X^n = (X_1, \ldots, X_n) \in X^n\)  
**frequentist/Bayesian model** \((\mathcal{P}, \mathcal{G})\) model subsets \(B, V \in \mathcal{G}\)

**parametrization** \(\Theta \rightarrow \mathcal{P} : \theta \mapsto P_\theta\)  
**model distributions**

**prior** \(\Pi : \mathcal{G} \rightarrow [0, 1]\)  
**probability measure**

**posterior** \(\Pi(\cdot | X^n) : \mathcal{G} \rightarrow [0, 1]\)  
**Bayes’s rule, inference**

**Frequentist**  
assume there is \(P_0\)  
\(X^n \sim P_0^n\)

**Bayes**  
assume \(P \sim \Pi\)  
\(X^n | P \sim P^n\)
Bayes’s Rule and Disintegration

**Definition 10.1** Assume that all $P \mapsto P^n(A)$ are $\mathcal{G}$-measurable. Given prior $\Pi$, a posterior is any $\Pi(\cdot|X^n = \cdot) : \mathcal{G} \times \mathcal{X}^n \to [0, 1]$ s.t.

(i) For any $G \in \mathcal{G}$, $x^n \mapsto \Pi(G|X^n = x^n)$ is $\mathcal{B}^n$-measurable

(ii) (Disintegration) For all $A \in \mathcal{B}^n$ and $G \in \mathcal{G}$

$$\int_A \Pi(G|X^n) dP^n_\Pi = \int_G P^n(A) d\Pi(P)$$

where $P^n_\Pi = \int P^n d\Pi(P)$ is the prior predictive distribution

**Remark 10.2** For frequentists $(X_1,\ldots,X_n) \sim P^n_0$, so assume $P^n_0 \ll P^n_\Pi$
Theorem 11.1 Assume $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is dominated by a $\sigma$-finite $\mu$ on $(\mathcal{Y}, \mathcal{B})$ with densities $p_\theta = dP_\theta/d\mu$. Then,

$$\Pi(\theta \in G \mid Y) = \frac{\int_G p_\theta(Y) d\Pi(\theta)}{\int_\Theta p_\theta(Y) d\Pi(\theta)},$$

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

The prior predictive has a density with respect to \( \mu \),

\[
P^\Pi(B) = \int_\Theta \int_B p_\theta(y) \, d\mu(y) \, d\Pi(\theta) = \int_B \left( \int_\Theta p_\theta(y) \, d\Pi(\theta) \right) \, d\mu(y).
\]

So the prior predictive density \( p^\Pi : \mathcal{Y} \to \mathbb{R} \) is equal to the denominator of the posterior. Note,

\[
\int_B \Pi(G|Y = y) \, dP^\Pi(y) = \int_B \left( \int_G p_\theta(Y) \, d\Pi(\theta) / \int_\Theta p_\theta(Y) \, d\Pi(\theta) \right) \, dP^\Pi(y)
\]

\[
= \int_B \int_G p_\theta(y) \, d\Pi(\theta) \, d\mu(y) = \int_G P_\theta(B) \, d\Pi(\theta),
\]

so the disintegration equality holds.
Proof

Since $P^\Pi(p^\Pi > 0) = 1$, the denominator is non-zero and the posterior is well-defined $P^\Pi$-a.s. For $y$ s.t. $p^\Pi(y) > 0$ and $(G_n)$ disjoint,

$$\prod \left( \theta \in \bigcup_{n \geq 1} G_n \mid Y = y \right) = C(y) \int_{\bigcup_n G_n} p_{\theta}(y) \, d\Pi(\theta)$$

$$= C(y) \int \sum_{n \geq 1} 1_{\{\theta \in G_n\}} p_{\theta}(y) \, d\Pi(\theta)$$

$$= \sum_{n \geq 1} C(y) \int_{G_n} p_{\theta}(y) \, d\Pi(\theta) = \sum_{n \geq 1} \Pi(\theta \in G_n \mid Y = y),$$

by monotone convergence. The posterior is well-defined and $\sigma$-additive, $P^\Pi$-a.s.
The Bayesian procedure consists of the following steps

(i) Based on the background of the data $Y$, choose a model $\mathcal{P}$, usually with parameterization $\Theta \rightarrow \mathcal{P} : \theta \mapsto P_{\theta}$.

(ii) Also choose a prior measure $\Pi$ on $\mathcal{P}$ (reflecting “belief”). Usually a measure on $\Theta$ is defined, inducing a measure on $\mathcal{P}$.

(iii) Calculate the posterior as a function of the data $Y$.

(iv) Observe a realization of the data $Y = y$, substitute in the posterior and do statistical inference.
Posterior predictive distribution

**Definition 15.1** Consider data $Y$ from $(Y, B)$, a model $P$ and prior $\Pi$. Assume that the posterior $\Pi(\cdot | Y)$ is regular. The **posterior predictive distribution** is defined,

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

for every event $B \in B$.

**Lemma 15.2** The posterior predictive distribution is a **probability measure**, almost surely.

**Lemma 15.3** Endow $P$ with the topology of total variation and a Borel prior $\Pi$. Suppose, either, that $P$ is relatively compact, or, that $\Pi$ is Radon. Then $\hat{P}$ lies in the **closed convex hull of $P$**, almost surely.
Proof

Let \( \epsilon > 0 \) be given. There exist \( \{P_1, \ldots, P_N\} \subset \mathcal{P} \) such that the balls \( B_i = \{P' \in \mathcal{P} : \|P' - P_i\| < \epsilon\} \) cover \( \mathcal{P} \). Define \( C_{i+1} = B_{i+1} \setminus \bigcup_{j=1}^{i} B_j \) (\( C_1 = B_1 \)), then \( \{C_1, \ldots, C_N\} \) is a partition of \( \mathcal{P} \). Define \( \lambda_i = \Pi(C_i | Y) \) (almost surely) and note,

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

So there exist elements in \( \text{co}(\mathcal{P}) \) that are arbitrarily close to \( \hat{P} \) in total variation. Conclude that \( \hat{P} \) lies in its closure.
Posterior mean

**Definition 17.1** Let $\mathcal{P}$ be a model parameterized by a closed, convex $\Theta$, subset of $\mathbb{R}^d$. Let $\Pi$ be a Borel prior. If $\theta$ is integrable with respect to the posterior, the posterior mean is defined

$$\hat{\theta}_1(Y) = \int_{\Theta} \theta \ d\Pi(\theta \mid Y) \in \Theta,$$

almost-surely.

**Remark 17.2** Convexity of $\Theta$ is necessary for interpretation.
Maximum-a-posteriori estimator

**Definition 18.1** Let \( \mathcal{P} \) be a model parametrized by \( \Theta \) with prior \( \Pi \).
Assume that the **posterior is dominated** by \( \sigma \)-finite measure \( \mu \) on \( \Theta \),
with density \( \theta \mapsto \pi(\theta|Y) \). The **maximum-a-posteriori (MAP) estimator** \( \hat{\theta}_2 \) is defined by,

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

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

Typically, the MAP-estimator maximizes

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

which is equivalent to log-likelihood maximization with **penalty** \( \log \pi(\theta) \).
Frequentist coverage

**Definition 19.1** Assume that $Y \sim P_{\theta_0}$ for some $\theta_0 \in \Theta$. Choose a confidence level $\alpha \in (0, 1)$. Then a subset $C_{\alpha}$ of $\Theta$ is a level-$\alpha$ confidence set if,

$$P_{\theta}(\theta \in C_{\alpha}) \geq 1 - \alpha,$$

for all $\theta \in \Theta$.

An asymptotic version exists, where we require that a sequence $(C_{\alpha,n})$ satisfies,

$$\liminf_{n \to \infty} P_{\theta}^{n}(\theta \in C_{\alpha,n}) \geq 1 - \alpha,$$

for all $\theta \in \Theta$.

Typically confidence sets are based on an estimator $\hat{\theta}$, or rather, on the distribution $\hat{\theta}$ has (the so-called **sampling distribution**).
Credible sets

**Definition 20.1** Let $\Theta$ parameterizing a model $P$ for data $Y$, with prior $\Pi$. Choose a credible level $\alpha \in (0, 1)$. Then a subset $D_\alpha \in \mathcal{G}$ of $\Theta$ is a level-$\alpha$ credible set if,

$$\Pi\left( \theta \in D_\alpha \mid Y \right) \geq 1 - \alpha,$$

almost-surely.

An asymptotic version exists, where we require that a sequence $(D_{\alpha,n})$ satisfies,

$$\liminf_{n \to \infty} \Pi\left( \theta \in D_{\alpha,n} \mid Y_n \right) \geq 1 - \alpha,$$

almost-surely.

Typically, credible sets in parametric models are level sets of the posterior density, the so-called **HPD-credible sets**.
Randomized testing

**Definition 21.1** Let \( \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \) be a model for data \( Y \). Assume given two hypotheses \( H_0 \) and \( H_1 \) for \( \theta \),

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

where \( \{\Theta_0, \Theta_1\} \) are a partition of \( \Theta \). A test function \( \phi \) is a map \( \phi : \mathcal{Y} \to [0, 1] \) used as a randomized test: given a realisation \( Y = y \) we reject \( H_0 \) with probability \( \phi(y) \).

The Neyman-Pearson lemma proves optimality of

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

for \( H_0 : P = P_{\theta_0} \) versus \( H_1 : P = P_{\theta_1} \).
Odds ratios and Bayes factors

**Definition 22.1** Let $\Theta$ parameterize a model $P$ for data $Y$ with prior $\Pi$. Let $\{\Theta_0, \Theta_1\}$ be a partition of $\Theta$ such that $\Pi(\Theta_0) > 0$ and $\Pi(\Theta_1) > 0$. The *prior odds ratio* and *posterior odds ratio* are defined by $\Pi(\Theta_0)/\Pi(\Theta_1)$ and $\Pi(\Theta_0|Y)/\Pi(\Theta_1|Y)$. The *Bayes factor* for $\Theta_0$ versus $\Theta_1$ is defined,

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

**Subjectivist** Accept $H_0$ if the posterior odds are greater than 1  
**Objectivist** Accept $H_0$ if the Bayes factor is greater than 1
Test sequences and asymptotics

Consider the case of data that forms a sequence \((Y_n)\), modelled with \(\mathcal{P}_n = \{P_{\theta,n} : \theta \in \Theta\}\) and hypotheses \(H_0 : \theta \in B\) and \(H_1 : \theta \in V\) for subsets \(B, V \subset \Theta\) s.t. \(B \cap V = \emptyset\).

A typical example: \(Y_n = (X_1, \ldots, X_n)\) i.i.d., with \(P_{\theta,n} = P_{\theta}^n\)

A test sequence \((\phi_n)\) is (asymptotically) consistent if,

\[
P_{\theta,n}\phi_n \to 0 \quad \text{and} \quad Q_{n,\theta'}(1 - \phi_n) \to 0,
\]
for all \(\theta \in B, \theta' \in V\). \((\phi_n)\) is uniformly (asymptotically) consistent if,

\[
\sup_{\theta \in B} P_{\theta,n}\phi_n \to 0 \quad \text{and} \quad \sup_{\theta' \in V} Q_{n,\theta'}(1 - \phi_n) \to 0,
\]
Bayesian tests

A test sequence \((\phi_n)\) is \(\Pi\)-a.s. (asymptotically) consistent if,

\[
P_{\theta,n} \phi_n \to 0 \text{ and } Q_{n,\theta'}(1 - \phi_n) \to 0,
\]
for all \(\Pi\)-almost-all \(\theta \in B, \theta' \in V\).

**Theorem 24.1** (BK, unpublished) Let \((\mathcal{P}, \mathcal{G}, \Pi)\) be given. For any \(B, V \in \mathcal{G}, B \cap V = \emptyset\), the following are equivalent,

(i) There exists a \(\Pi\)-a.s. consistent test sequence for \(B\) versus \(V\);

(ii) There exists a test sequence \((\phi_n)\) s.t.

\[
\int_B P_{\theta,n} \phi_n \, d\Pi(\theta) + \int_V Q_{n,\theta'}(1 - \phi_n) \, d\Pi(\theta) \to 0
\]

(iii) The posterior satisfies \(\Pi(V|X_n) \xrightarrow{P\text{-a.s.}} 0\) and \(\Pi(B|X_n) \xrightarrow{Q\text{-a.s.}} 0\), for \(\Pi\)-almost-all \(P \in B, Q \in V\).
Optimal tests and the minimax theorem

We say that \((\phi_n)\) is minimax optimal if,

\[
\sup_{\theta \in \Theta_0} P^n_\theta \phi_n + \sup_{\theta \in \Theta_1} P^n_\theta (1 - \phi_n) = \inf_{\psi} \left( \sup_{\theta \in \Theta_0} P^n_\theta \psi + \sup_{\theta \in \Theta_1} P^n_\theta (1 - \psi) \right),
\]

**Theorem 25.1** Assume that \(\Phi\) and \(\Theta\) are convex, that \(\phi \mapsto R(\theta, \phi)\) is convex for every \(\theta\) and that the map \(\theta \mapsto R(\theta, \phi)\) is concave for every \(\phi\). Furthermore, suppose that \(\Phi\) is compact and \(\phi \mapsto R(\theta, \phi)\) is continuous for all \(\theta\). Then there exists a minimax optimal test \(\phi^*\) and,

\[
\sup_{\theta \in \Theta} R(\theta, \phi^*) = \inf_{\phi \in \Phi} \sup_{\theta \in \Theta} R(\theta, \phi) = \sup_{\theta \in \Theta} \inf_{\phi \in \Phi} R(\theta, \phi).
\]
Examples of uniform test sequences

In the following, fix \( n \geq 1 \) and consider \( i.i.d. \) data \( Y = (X_1, \ldots, X_n) \). The model \( \mathcal{P} \) contains probability measures \( P \) s.t. \( Y \sim P^n \).

**Lemma 26.1** (Minimax Hellinger tests) Let \( B, V \subset \mathcal{P} \) be convex with \( H(B, V) > 0 \). There exist a uniform test sequence \( (\phi_n) \) s.t.

\[
\sup_{P \in B} P^n \phi_n \leq e^{-\frac{1}{2}n H^2(B, V)}, \quad \sup_{Q \in V} Q^n (1 - \phi_n) \leq e^{-\frac{1}{2}n H^2(B, V)}.
\]
Proof

The minimax risk $\pi(B, V)$ for testing $B$ versus $Q$ is

$$\pi(B, V) = \inf_{\phi} \sup_{(P, Q) \in B \times V} (P\phi + Q(1 - \phi))$$

Apply the minimax theorem,

$$\inf_{\phi} \sup_{(P, Q)} (P\phi + Q(1 - \phi)) = \sup_{P, Q} \inf_{\phi} (P\phi + Q(1 - \phi))$$

On the r.h.s. $\phi$ can be chosen $(P, Q)$-dependently; minimal for $\phi = 1\{p < q\}$ (remember the Neyman-Pearson test)

$$\pi(B, V) = \sup_{P, Q} (P(p < q) + Q(p \geq q))$$
Proof

Note that:

\[ P(p < q) + Q(p \geq q) = \int_{p<q} p \, d\mu + \int_{p\geq q} q \, d\mu \]

\[ \leq \int_{p<q} p^{1/2}q^{1/2} \, d\mu + \int_{p\geq q} p^{1/2}q^{1/2} \, d\mu = 1 - \frac{1}{2}H^2(P, Q) \leq e^{-\frac{1}{2}H^2(P, Q)}. \]

This relates minimax testing power to the Hellinger distance between \( P \) and \( Q \). For product measures, \( n \)-th power.

\[ \pi(P^n, Q^n) \leq e^{-\frac{1}{2}nH^2(P, Q)}. \]
In the following, fix \( n \geq 1 \) and consider \emph{i.i.d.} data \( Y = (X_1, \ldots, X_n) \). The model \( \mathcal{P} \) contains probability measures \( P \) s.t. \( Y \sim P^n \).

**Lemma 29.1 (Weak tests)** Let \( \epsilon > 0 \), \( P_0 \in \mathcal{P} \) and a measurable \( f : \mathcal{X}^n \to [0, 1] \) be given. Define,

\[
B = \left\{ P \in \mathcal{P} : |(P^n - P^n_0)f| < \epsilon \right\}, \quad V = \left\{ P \in \mathcal{P} : |(P^n - P^n_0)f| \geq 2\epsilon \right\}.
\]

There exist a \( D > 0 \) and \emph{uniformly consistent test sequence} \( (\phi_n) \) s.t.

\[
\sup_{P \in B} P^n \phi_n \leq e^{-nD}, \quad \sup_{Q \in V} Q^n (1 - \phi_n) \leq e^{-nD}.
\]

Proof relies on Hoeffding's inequality.
Lecture II

The Bernstein-Von Mises theorem

The second lecture is devoted to regular estimation problems and the Bernstein-von Mises theorem, both parametrically and semi-parametrically. We discuss regularity, local asymptotic normality, efficiency and the consequences and applications of the parametric Bernstein-von Mises theorem. We then turn to semiparametrics, considering consistency under perturbation, integral IAN and the semi-parametric Bernstein-von Mises theorem. Semi-parametric bias is mentioned as a major obstacle.
Example Parametric regression

Questions

Observe \( i.i.d. \ Y_1, \ldots, Y_n, \ Y_i = \theta + e_i \) (or \( Y_i = \theta X_i + e_i, \) etcetera) with a normally distributed error (of known variance). The density for the observation is,

\[
p_{\theta_0}(x) = \phi(x - \theta_0),
\]

where \( \phi \) is the density for the relevant normal distribution. Note the Fisher information for location is non-singular.

What should we expect of the posterior for \( \theta \) in this model?

If we generalize to include non-parametric modelling freedom, what can be said about the (marginal) posterior for \( \theta \)?
Convergence of the posterior distribution with growing sample size $n = 0, 1, 4, \ldots, 400$. Note: concentration at correct $\theta_0$, at parameteric rate $\sqrt{n}$ and variance is the inverse Fisher information.)
Local Asymptotic Normality \textbf{LAN}

\textbf{Definition 33.1 (Le Cam (1960))}

There is a $\dot{\ell}_{\theta_0} \in L_2(P_{\theta_0})$ with $P_{\theta_0} \dot{\ell}_{\theta_0} = 0$ s.t. for any $(h_n) = O(1)$,

$$\prod_{i=1}^{n} \frac{p_{\theta_0+n^{-1/2}h_n}}{p_{\theta_0}}(X_i) = \exp(h_n \Delta'_{n,\theta_0} - \frac{1}{2} h_n I_{\theta_0} h_n + o_{P_{\theta_0}}(1)),$$

where $\Delta'_{n,\theta_0}$ is given by,

$$\Delta'_{n,\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) \xrightarrow{P_{\theta_0}-w.} N(0, I_{\theta_0}),$$

and $I_{\theta_0} = P_{\theta_0} \dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0}^T$ is the Fisher information.
Differentiability in quadratic mean (DQM)

**Definition 34.1 (Le Cam (1960))**
A model \( \mathcal{P} \) is differentiable in quadratic mean at \( \theta_0 \) with score \( \dot{\ell}_{\theta_0} \) if

\[
\int \left( p^{1/2}_\theta - p^{1/2}_{\theta_0} - \frac{1}{2}(\theta - \theta_0) \dot{\ell}_{\theta_0} p^{1/2}_{\theta_0} \right)^2 d\mu = o(\|\theta - \theta_0\|^2).
\]

Then \( P_0 \dot{\ell}_{\theta_0} = 0 \), \( \dot{\ell}_{\theta_0} \in L_2(P_{\theta_0}) \) and \( I_{\theta_0} = P_0 \dot{\ell}_{\theta_0} \dot{\ell}_{\theta_0} \) is the Fisher information.

**Lemma 34.2 (Le Cam (1960))**
The model \( \mathcal{P} \) is DQM at \( \theta_0 \) iff \( \mathcal{P} \) is LAN at \( \theta_0 \).

**Remark 34.3** Sufficient is differentiability of \( \theta \mapsto p_{\theta}(y) \) for every \( y \).
Estimator regularity and the convolution theorem

**Definition 35.1** An estimator sequence $\hat{\theta}_n$ for a parameter $\theta_0$ is said to be regular, if for every $h_n = O(1)$, with $\theta_n = \theta_0 + n^{-1/2}h_n$

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \overset{P_{\theta_n}\text{-w.}}{\longrightarrow} L_{\theta_0}$$

for some $(h_n)$-independent limit distribution $L_{\theta_0}$.

**Theorem 35.2** (Hájek, 1970)
Assume that the model is LAN at $\theta_0$ with non-singular Fisher information $I_{\theta_0}$. Suppose $\hat{\theta}_n$ is a regular estimator for $\theta_0$ with limit $L_{\theta_0}$. Then there exists a probability kernel $M_{\theta_0}$ such that $L_\theta = N(0, I_{\theta_0}^{-1}) \ast M_{\theta_0}$. 
Regular estimation and efficiency

**Definition 36.1** Given an asymptotic estimation problem with i.i.d.-$P_0$ data and non-singular Fisher information $I_0$, an influence function $\Delta_n$ is,

$$\Delta_n = I_0^{-1} \Delta' = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_0^{-1} \dot{\ell}_{\theta_0}(X_i) \xrightarrow{P_0-w.} N(0, I_0^{-1})$$

**Theorem 36.2** *(Fisher, Cramér, Rao, Le Cam, Hájek)*

An estimator $\hat{\theta}_n$ is **efficient** if and only if it is asymptotically linear:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \Delta_{n,\theta_0} + o_{P_0}(1),$$

for some influence function $\Delta_{n,\theta_0} \xrightarrow{P_{\theta_0-w.}} N(0, I_{\theta_0}^{-1}).$

**Remark 36.3** asymptotic bias equals zero because $P_0 \dot{\ell}_{\theta_0} = 0.$
Efficiency of the maximum likelihood estimator

For all \( n \geq 1 \), let \( X_1, \ldots, X_n \) denote i.i.d. data with marginal \( P_0 \).

**Theorem 37.1** (see van der Vaart (1998))

Assume that \( \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \) with \( \Theta \) open in \( \mathbb{R}^k \) and that there exists a \( \theta_0 \in \Theta \) s.t. \( P_0 = P_{\theta_0} \). Furthermore, assume that \( \mathcal{P} \) is LAN at \( \theta_0 \) and that \( I_{\theta_0} \) is non-singular. Also assume there exists an \( L^2(P_{\theta_0}) \)-function \( \dot{\ell} \) s.t. for any \( \theta, \theta' \) in a neighbourhood of \( \theta_0 \) and all \( x \),

\[
| \log p_\theta(x) - \log p_{\theta'}(x) | \leq \dot{\ell}(x) \| \theta - \theta' \|,
\]

If the ML estimate \( \hat{\theta}_n \) is consistent, it is efficient,

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{P_{\theta_0}-w.} N(0, I_{\theta_0}^{-1}).
\]
Parametric Bernstein-von Mises theorem

**Theorem 38.1** (Le Cam (1953), $h = \sqrt{n}(\theta - \theta_0)$)

Let $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ with thick prior $\Pi_\Theta$ be LAN at $\theta_0$ with non-singular $I_{\theta_0}$. Assume that for every sequence of radii $M_n \to \infty$,

$$\Pi\left( \|h\| \leq M_n \mid X_1, \ldots, X_n \right) \xrightarrow{P_0} 1$$

Then the posterior converges to normality as follows

$$\sup_B \left| \Pi\left( \theta \in B \mid X_1, \ldots, X_n \right) - N_{\Delta_n,\theta_0, I_{\theta_0}^{-1}(B)} \right| \xrightarrow{P_0} 0$$

**Remark 38.2** With $\widehat{\theta}_n$ any efficient estimator,

$$\sup_B \left| \Pi\left( \theta \in B \mid X_1, \ldots, X_n \right) - N_{\widehat{\theta}_n,(nI_{\theta_0})^{-1}(B)} \right| \xrightarrow{P_0} 0$$

**Remark 38.3** (BK and van der Vaart, 2012) There's a version for the misspecified situation ($P_0 \not\in \mathcal{P}$).
Consequences and applications

i. Bayesian point estimators are efficient

ii. Confidence intervals based on the sampling distribution of an efficient estimator and credible sets coincide asymptotically

Model selection with the Bayesian Information Criterion (BIC). Consider parameter spaces \( \Theta_k \subset \mathbb{R}^k \), \( (k \geq 1) \) with models \( \mathcal{P}_k \) for i.i.d. data \( X_1, \ldots, X_n \). Define,

\[
\text{BIC}(\theta, k) = -2 \log L_n(X_1, \ldots, X_n; \theta_1, \ldots, \theta_k) + k \log(n)
\]

Minimization of \( \text{BIC}(\theta_1, \ldots, \theta_k; k) \) with respect to \( \theta \) and \( k \) is penalized ML estimate that selects a value of \( k \). Closely related to AIC, RIC, MDL and other model selection methods.
Efficiency of formal Bayes estimators

**Definition 40.1** Let $Y$, $\mathcal{P}$, $\Pi$ be like before and let $\ell : \mathbb{R}^k \to [0, \infty)$ be a loss function. The posterior risk is defined almost-surely,

$$t \mapsto \int_{\Theta} \ell\left(\sqrt{n}(t - \theta)\right) d\Pi(\theta|Y).$$

A minimizer $\hat{\theta}_{3,n}$ of posterior risk is called the **formal Bayes estimator** associated with $\ell$ and $\Pi$.

**Theorem 40.2** (Le Cam (1953,1986) and van der Vaart (1998)) Assume that the BvM theorem holds and that $\ell$ is non-decreasing and $\ell(h) \leq 1 + \|h\|^p$ for some $p > 0$ such that $\int \|\theta\|^p d\Pi(\theta) < \infty$. Then $\sqrt{n}(\hat{\theta}_{3,n} - \theta_0)$ converges weakly to the minimizer of $\int \ell(t-h) dN_{Z,I_{\theta_0}^{-1}}(h)$ where $Z \sim N(0, I_{\theta_0}^{-1})$. 
Example Semiparametric regression

New Question
Observe $i.i.d. \ X_1, \ldots, X_n, \ X_i = \theta + e_i$ (or $Y_i = \theta \ X_i + e_i$, etcetera) with a symmetrically distributed error. Density for $X$’s is,

$$p_{\theta_0, \eta_0}(x) = \eta_0(x - \theta_0),$$

where $\eta \in H$ is a symmetric Lebesgue density on $\mathbb{R}$. We assume that $\eta$ is smooth and that the Fisher information for location is non-singular.

Adaptivity Stein (1956), Bickel (1982)
For inference on $\theta_0$ it does not matter whether we know $\eta_0$ or not!

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{P_{\theta_0, \eta_0}-w.} N(0, I_{\theta_0, \eta_0}^{-1})$$

where $I_{\theta_0, \eta_0}$ is the Fisher information.
Parametric/Semi-parametric analogy

Parametric posterior

The posterior density $\theta \mapsto d\Pi(\theta|X_1, \ldots, X_n)$

$$\prod_{i=1}^{n} p_\theta(X_i) d\Pi(\theta) \bigg/ \int_{\Theta} \prod_{i=1}^{n} p_\theta(X_i) d\Pi(\theta)$$

with LAN requirement on the likelihood.

Semiparametric analog

The marginal posterior density $\theta \mapsto d\Pi(\theta|X_1, \ldots, X_n)$

$$\int_{H} \prod_{i=1}^{n} p_{\theta,\eta}(X_i) d\Pi_H(\eta) d\Pi(\theta) \bigg/ \int_{\Theta} \int_{H} \prod_{i=1}^{n} p_{\theta,\eta}(X_i) d\Pi_H(\eta) d\Pi(\theta)$$

with integral LAN requirement on $\Pi_H$-integrated likelihood.
Integral local asymptotic normality ILAN

**Definition 43.1** Given a nuisance prior $\Pi_{H}$, the localized integrated likelihood is,

$$s_{n}(h) = \int_{H} \prod_{i=1}^{n} \frac{p_{\theta_{0}+n^{-1/2}h,\eta}(X_{i})}{p_{\theta_{0},\eta_{0}}} d\Pi_{H}(\eta),$$

**Definition 43.2** $s_{n}$ is said to have the ILAN property, if for every $h_{n} = O_{P_{0}}(1)$

$$\log \frac{s_{n}(h_{n})}{s_{n}(0)} = h_{n}^{T} \tilde{\Delta}_{n,\theta_{0},\eta_{0}} - \frac{1}{2} h_{n}^{T} \tilde{I}_{\theta_{0},\eta_{0}} h_{n} + o_{P_{0}}(1),$$

where the efficient $\tilde{\Delta}_{n,\theta_{0},\eta_{0}}$ is given by

$$\tilde{\Delta}_{n,\theta_{0},\eta_{0}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{\infty} \tilde{\ell}_{\theta_{0},\eta_{0}} \xrightarrow{P_{\theta_{0},\eta_{0}}-w.} N(0, \tilde{I}_{\theta_{0},\eta_{0}})$$
Consistency under $\sqrt{n}$-perturbation

Given $\rho_n \downarrow 0$ we speak of *consistency under $n^{-1/2}$-perturbation at rate $\rho_n$*, if for all $h_n = O_{P_0}(1)$,

\[
\Pi_n\left(D(\theta, \rho_n) \mid \theta = \theta_0 + n^{-1/2}h_n ; X_1, \ldots, X_n \right) \xrightarrow{P_0} 1
\]
reparametrize \((\theta, \zeta) \mapsto (\theta, \eta^*(\theta) + \zeta)\)
Semiparametric Bernstein-von Mises theorem

**Theorem 46.1** (Bickel and BK (2012))

Let \( \mathcal{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\} \) with **thick** prior \( \Pi_{\Theta} \) and nuisance prior \( \Pi_{H} \). Assume **ILAN** at \( P_{\theta_0,\eta_0} \) with **non-singular** \( \tilde{I}_{\theta_0,\eta_0} \). Assume that for every sequence of radii \( M_n \to \infty \),

\[
\Pi\left( \|h\| \leq M_n \mid X_1, \ldots, X_n \right) \xrightarrow{P_0} 1
\]

Then the posterior converges marginally to normality as follows

\[
\sup_{B} \left| \Pi\left( h \in B \mid X_1, \ldots, X_n \right) - N_{\Delta_{n,\theta_0,\eta_0},\tilde{I}_{\theta_0,\eta_0}^{-1}}(B) \right| \xrightarrow{P_0} 0
\]

**Both ILAN** and **\( \sqrt{n} \)-consistency** are sensitive to semiparametric bias!
Semiparametric bias

An estimator $\hat{\theta}_n$ for $\theta_0$ is regular but asymptotically biased if,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \tilde{\Delta}_{n,\theta_0,\eta_0} + \mu_{n,\theta_0,\eta_0} + o_{P_0}(1),$$

with $\tilde{\Delta}_{n,\theta_0,\eta_0} \xrightarrow{P_0-w.} N(0, \tilde{I}_{\theta_0,\eta_0}^{-1})$ and $\mu_{n,\theta_0,\eta_0} = O(1)$ or worse. Typically,

$$\left| \mu_{n,\theta_0,\eta_0} \right| \leq n^{-1/2} \sup_{\eta \in D_n} \left| \tilde{I}_{\theta_0,\eta_0}^{-1} P_{\theta_0,\eta} \tilde{\ell}_{\theta_0,\eta_0} \right|$$

where $D_n$ describes some form of localization for $\eta \in H$ around $\eta_0$.

**Theorem 47.1** (approximate, see Schick (1986), Klaassen (1987))

An efficient estimator for $\theta_0$ exists if and only if there exists an estimator $\hat{\Delta}_n$ for the influence function, whose asymptotic bias vanishes at a rate strictly faster than $\sqrt{n}$,

$$P_{\theta_0,\eta}^{n} \hat{\Delta}_n = o(n^{-1/2}).$$
Example Regression with symmetric errors

**Theorem 48.1 (Chae, Kim and BK (201?))**

Let $X_1, \ldots, X_n$ be i.i.d.-$P_{\theta_0, \eta_0}$, i.e. $X_i = \theta_0 + e_i$ with $e$ distributed as a symmetric normal location mixture $\eta_0$ from $H$ of the form,

$$\eta(x) = \int \phi(x - z) \, dF(z)$$

(where $F$ is symmetric and $\phi$ denotes the standard normal density).

With thick prior $\Pi_\Theta$ and nuisance prior $\Pi_H$ that has full weak support, the posterior converges marginally to normality

$$\sup_B \left| \prod \left( h \in B \mid X_1, \ldots, X_n \right) - N_{\Delta_{n, \theta_0, \eta_0}, \tilde{I}_{\theta_0, \eta_0}^{-1}}(B) \right| \xrightarrow{P_0} 0$$

where $\tilde{\ell}_{\theta_0, \eta_0}(X) = \dot{p}_{\theta_0, \eta_0} / p_{\theta_0, \eta_0}(X)$ and $\tilde{I}_{\theta_0, \eta_0} = P_0 \tilde{\ell}^2_{\theta_0, \eta_0}$. 
In the third lecture we consider application of Bayesian methods in non-parametric models: we do not focus on the construction of non-parametric priors but on the requirements for such priors to lead to consistent posteriors. After a review of the consequences of posterior consistency, we turn to Doob’s theorem, Freedman’s counterexamples and Schwartz’s theorem, which we prove. We also point out the limitations of Schwartz’s theorem.
Frequentist consistency

Let $X_1, \ldots, X_n$ be $i.i.d.-P_{\theta_0}$-distributed

Consider a point-estimator $\hat{\theta}_n(X)$.

An estimator is said to be (strongly) consistent if

$$\hat{\theta}_n \xrightarrow{P_{\theta_0}(-a.s.)} \theta_0.$$ 

E.g. if the topology is metric, a consistent estimator $\hat{\theta}_n$ is found at a distance from $\theta_0$ greater than some $\epsilon > 0$ with $P_{\theta_0}^n$-probability arbitrarily small, if we make the sample large enough.

Since $\theta_0$ is unknown, we have to prove this for all $\theta \in \Theta$ before it is useful.
Frequentist rate of convergence

Next, suppose that $\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$. Let $(r_n)$ be a sequence $r_n \downarrow 0$.

We say that $\hat{\theta}_n$ converges to $\theta_0$ at rate $r_n$ if

$$r_n^{-1}\|\hat{\theta}_n - \theta_0\| = O_{P_{\theta_0}}(1)$$

So $r_n$ compensates the decrease in distance between $\hat{\theta}_n$ and $\theta_0$, such that the fraction is bounded in probability.

Or: the $r_n$ are the radii of balls around $\hat{\theta}_n$ that shrink (just) slowly enough to still capture $\theta_0$ with high probability.
Frequentist limit distribution

Suppose that \( \hat{\theta}_n \) converges to \( \theta_0 \) at rate \( r_n \).

Let \( L_{\theta_0} \) be a non-degenerate but tight distribution. If

\[
r_n^{-1} (\hat{\theta}_n - \theta_0) \overset{P_{\theta_0}}{\longrightarrow} L_{\theta_0},
\]

we say that \( \hat{\theta}_n \) converges to \( \theta_0 \) at rate \( r_n \) with limit-distribution \( L_{\theta_0} \).

So if we blow up the difference between \( \hat{\theta}_n \) and \( \theta_0 \) by exactly the right factors \( r_n^{-1} \), we keep up with convergence and arrive at a stable distribution \( L_{\theta_0} \).
Posterior consistency

Given \( P_0 \)-i.i.d. \( X^n \), \( \mathcal{P} \) with prior \( \Pi \), do posteriors concentrate on \( P_0 \)?

**Definition 53.1** Given a model \( \mathcal{P} \) with Borel prior \( \Pi \), the posterior is (strongly) consistent at \( P \in \mathcal{P} \) if for every neighbourhood \( U \) of \( P \)

\[
\Pi(U|X^n) \xrightarrow{P(-a.s.)} 1
\]
Consistency is Prokhorov’s tight convergence

**Theorem 54.1** Let $\mathcal{P}$ be a uniform model with Borel prior $\Pi$. The posterior is strongly consistent, if and only if, for every bounded, continuous $f : \mathcal{P} \rightarrow \mathbb{R}$,

$$\int f(P) d\Pi(P|X^n) \xrightarrow{P_0\text{-a.s.}} f(P_0),$$

which we denote by $\Pi(\cdot|X_1,\ldots,X_n) \xrightarrow{w} \delta_{P_0}$.

**Remark 54.2** All weak, polar and metric topologies are uniform:

$$U = \{P \in \mathcal{P} : |(P - P_0)f| < \epsilon\}, V = \{P \in \mathcal{P} : \sup_{f \in B} |(P - P_0)f| < \epsilon\},$$

$$W = \{P \in \mathcal{P} : d(P, P_0) < \epsilon\},$$

for $\epsilon > 0$ and functions $0 \leq f \leq 1$ measurable (or smaller class).
Proof

Assume (1). Let \( f : \mathcal{P} \to \mathbb{R} \) be bounded (\(|f| \leq M\)) and continuous. Let \( \eta > 0 \) be given. Let \( U \) be a neighbourhood of \( P_0 \) s.t. \(|f(P) - f(P_0)| \leq \eta\) for all \( P \in U \). Integrate \( f \) with respect to the posterior and to \( \delta_{P_0} \):

\[
\left| \int_{\mathcal{P}} f(P) d\Pi_n(P|X_1, \ldots, X_n) - f(P_0) \right|
\]

\[
\leq \int_{\mathcal{P} \setminus U} |f(P) - f(P_0)| d\Pi_n(P|X_1, \ldots, X_n)
\]

\[
+ \int_U |f(P) - f(P_0)| d\Pi_n(P|X_1, \ldots, X_n)
\]

\[
\leq 2M \Pi_n(\mathcal{P} \setminus U | X_1, X_2, \ldots, X_n)
\]

\[
+ \sup_{P \in U} |f(P) - f(P_0)| \Pi_n(U | X_1, X_2, \ldots, X_n)
\]

\[
\leq \eta + o(1), \quad (n \to \infty).
\]

Consequently, (2) holds.
Proof

Conversely, assume (2) holds. Let $U$ be an open neighbourhood of $P_0$. Because $\mathcal{P}$ is completely regular, there exists a continuous $f : \mathcal{P} \to [0, 1]$ that separates $\{P_0\}$ from $\mathcal{P} \setminus U$, i.e. $f = 1$ at $\{P_0\}$ and $f = 0$ on $\mathcal{P} \setminus U$.

\[
\liminf_{n \to \infty} \prod_n (U | X_1, X_2, \ldots, X_n) = \liminf_{n \to \infty} \int_{\mathcal{P}} 1_U(P) \, d\prod_n (P|X_1, \ldots, X_n) \\
\geq \liminf_{n \to \infty} \int_{\mathcal{P}} f(P) \, d\prod_n (P|X_1, \ldots, X_n) = \int_{\mathcal{P}} f(P) \, d\delta_{P_0}(P) = 1,
\]

$P_0$-almost-surely. Consequently, (1) holds.
Consistency of Bayesian point estimators

**Theorem 57.1** Suppose that $\mathcal{P}$ is endowed with the topology of total variation. Assume that the posterior is strongly consistent. Then the posterior mean $\hat{P}_n$ is a $P_0$-almost-surely consistent point-estimator in total-variation.
Proof

Extend $P \mapsto \|P - P_0\|$ to the convex hull of $\mathcal{P}$. Since $P \mapsto \|P - P_0\|$ is convex, Jensen says,

$$\|\hat{P}_n - P_0\| = \left\| \int_{\mathcal{P}} P\,d\Pi_n(\,P \mid X_1, \ldots, X_n ) - P_0 \right\| \leq \int_{\mathcal{P}} \|P - P_0\|\,d\Pi_n(\,P \mid X_1, \ldots, X_n ).$$

Since $P \xrightarrow{\Pi_n-w.} P_0$ under $\Pi_n = \Pi_n(\cdot \mid X_1, \ldots, X_n )$ and $P \mapsto \|P - P_0\|$ is bounded and continuous, the r.h.s. converges to the expectation of $\|P - P_0\|$ under the limit law $\delta_{P_0}$, which equals zero. Hence

$$\hat{P}_n \xrightarrow{P_0-a.s.} P_0,$$

in total variation.
Doob’s theorem

**Theorem 59.1** *(Doob (1948))*
Suppose that the parameter space $\Theta$ and the sample space $\mathcal{X}$ are Polish spaces endowed with their respective Borel $\sigma$-algebras. Assume that $\Theta \rightarrow \mathcal{P} : \theta \mapsto P_\theta$ is one-to-one. Then for any prior $\Pi$ on $\Theta$ the posterior is consistent, $\Pi$-almost-surely.

**Proof** An application of Doob’s martingale convergence theorem (see van der Vaart (1998) or Ghosh and Ramamoorthi (2003)), combined with a difficult argument on existence of a measurable $f : \mathcal{X}^\infty \rightarrow \Theta$ s.t. $f(X_1, X_2, \ldots) = \theta$, $P_\theta^\infty$ – a.s. for all $\theta \in \Theta$ (Le Cam’s accessibility (Breiman, Le Cam, Schwartz (1964), Le Cam (1986))).
Freedman’s point

Remark 60.1  Doob’s theorem says nothing about specific points: it is always possible that $P_0$ belongs to the null-set for which inconsistency occurs.

Remark 60.2 (Non-parametric counterexamples)

Example 60.3  Let $X_1, X_2 \ldots \in \mathbb{N}$ be i.i.d.-$P_0$. The full model is the unit simplex in $\ell^1$, $\mathcal{P} = \{(p_i) \in [0,1] : p_i \geq 0, \sum_i p_i = 1\}$. Let $\mathcal{P}_i = \{P \in \mathcal{P} : p_i = 0\}$ with prior $\Pi_i$ of full support $\mathcal{P}_i$. Define $\Pi' = \sum_i \lambda_i \Pi_i$ for $(\lambda_i)$ s.t. $\lambda_i > 0$, $\sum_i \lambda_i = 1$. For some fixed $P$, choose $\Pi = \frac{1}{2} \Pi' + \frac{1}{2} \delta_P$. $\Pi$ has full support on $\mathcal{P}$. Nonetheless, if $P_0(X = i) > 0$ for all $i \geq 1$, then the posterior is inconsistent (it converges to $\delta_P$).
Schwartz’s theorem

\textbf{Theorem 61.1} \textit{(Schwartz (1965))}

Assume that

(i) For every \( \epsilon > 0 \), there is a test sequence \((\phi_n)\) s.t.

\[ P_0^n \phi_n \to 0, \quad \sup \{P : d(P, P_0) > \epsilon\} \]

\[ P^n (1 - \phi_n) \to 0. \]

(ii) Let \( \Pi \) be a KL-prior, i.e. for every \( \eta > 0 \),

\[ \Pi \left( P \in \mathcal{P} : -P_0 \log \frac{p}{p_0} \leq \eta \right) > 0, \]

Then the posterior is strongly consistent at \( P_0 \).

\textbf{Theorem 61.2} Let \( \mathcal{P} \) be Hellinger totally bounded and let \( \Pi \) a KL-prior. Then the posterior is Hellinger consistent at \( P_0 \).
Proof of Schwartz’s theorem

Let $\epsilon, \eta > 0$ be given. Define

$$V = \{ P \in \mathcal{P} : d(P, P_0) \geq \epsilon \}.$$

Split the $n$-th posterior (of $V$) with the test functions $\phi_n$ and take the lim sup:

$$\limsup_{n \to \infty} \prod_n (V | X_1, \ldots, X_n) \leq \limsup_{n \to \infty} \prod_n (V | X_1, \ldots, X_n) (1 - \phi_n) + \limsup_{n \to \infty} \prod_n (V | X_1, \ldots, X_n) \phi_n. \quad (3)$$

Define $K_\eta = \{ P \in \mathcal{P} : -P_0 \log(p/p_0) \leq \eta \}$. For every $P \in K_\eta$, LLN

$$\left| \mathbb{P}_n \log \frac{p}{p_0} - P_0 \log \frac{p}{p_0} \right| \to 0, \quad (P_0 - a.s.)$$
Proof of Schwartz’s theorem

So for every $\alpha > \eta$ and all $P \in K_\eta$,

$$\prod_{i=1}^{n} \frac{p}{p_0}(X_i) \geq e^{-n\alpha},$$

$P^n_0$-almost-surely. Use this to lower-bound the denominator

$$\liminf_{n \to \infty} e^{n\alpha} \int_{\mathcal{P}} \prod_{i=1}^{n} \frac{p}{p_0}(X_i) d\Pi(P) \geq \liminf_{n \to \infty} e^{n\alpha} \int_{K_\eta} \prod_{i=1}^{n} \frac{p}{p_0}(X_i) d\Pi(P)$$

$$\geq \int_{K_\eta} \liminf_{n \to \infty} e^{n\alpha} \prod_{i=1}^{n} \frac{p}{p_0}(X_i) d\Pi(P) \geq \Pi(K_\eta) > 0.$$
Proof of Schwartz’s theorem

The first term in (3) can be bounded as follows

\[
\limsup_{n \to \infty} \prod_n (V | X_1, \ldots, X_n)(1 - \phi_n)(X_1, \ldots, X_n)
\]

\[
\leq \limsup_{n \to \infty} e^{n\alpha} \int_V \prod_{i=1}^n \left(\frac{p}{p_0}\right)(X_i) (1 - \phi_n)(X_1, \ldots, X_n) \, d\Pi(P)
\]

\[
\leq \liminf_{n \to \infty} e^{n\alpha} \int_{\mathcal{P}} \prod_{i=1}^n \left(\frac{p}{p_0}\right)(X_i) \, d\Pi(P)
\]

\[
\leq \frac{1}{\Pi(K_\eta)} \limsup_{n \to \infty} f_n(X_1, \ldots, X_n),
\]

where we use the (non-negative)

\[
f_n(X_1, \ldots, X_n) = e^{n\alpha} \int_V \prod_{i=1}^n \frac{p}{p_0}(X_i) (1 - \phi_n)(X_1, \ldots, X_n) \, d\Pi(P).
\]
Proof of Schwartz’s theorem, interlude

At this stage in the proof we need the following lemma, which says that uniform consistency of testing can be assumed to be of exponential power without loss of generality.

**Lemma 65.1** Suppose that for given $\epsilon > 0$ there exists a sequence of tests $(\phi_n)$ such that:

$$P_0^n \phi_n \to 0, \quad \sup_{P \in V_\epsilon} P^n(1 - \phi_n) \to 0,$$

where $V_\epsilon = \{P \in \mathcal{P} : d(P, P_0) \geq \epsilon\}$. Then there exists a sequence of tests $(\omega_n)$ and positive constants $C, D$ such that:

$$P_0^n \omega_n \leq e^{-nC}, \quad \sup_{P \in V_\epsilon} P^n(1 - \omega_n) \leq e^{-nD} \quad (5)$$
Proof of Schwartz’s theorem

The previous lemma guarantees that there exists a constant $\beta > 0$ such that for large enough $n,$

$$P_0^\infty f_n = P_0^n f_n = e^{n\alpha} \int_V P_0^n \left( \prod_{i=1}^n \frac{p}{p_0} (X_i) (1 - \phi_n)(X_1, \ldots, X_n) \right) d\Pi(P)$$

\[
\leq e^{n\alpha} \int_V P^n (1 - \phi_n) d\Pi(P) \leq e^{-n(\beta - \alpha)}.
\]

Choose $\eta < \beta$ and $\alpha$ such that $\eta < \alpha < \frac{1}{2}(\beta + \eta).$ Markov’s inequality

$$P_0^\infty \left( f_n > e^{-\frac{n}{2}(\beta - \eta)} \right) \leq e^{\frac{n}{2}(\beta - \eta)} P_0^\infty f_n \leq e^{n(\alpha - \frac{1}{2}(\beta + \eta))}.$$
Proof of Schwartz’s theorem

Hence \( \sum_{n=1}^{\infty} P_0^\infty (f_n > \exp -\frac{n}{2} (\beta - \eta)) \) converges. By the first Borel-Cantelli lemma

\[
0 = P_0^\infty \left( \bigcap_{N=1}^{\infty} \bigcup_{n \geq N} \{ f_n > e^{-\frac{n}{2}(\beta-\eta)} \} \right) \geq P_0^\infty \left( \limsup_{n \to \infty} (f_n - e^{-\frac{n}{2}(\beta-\eta)}) > 0 \right)
\]

So \( f_n \to 0, \) \( (P_0 - a.s.) \) and hence

\[
\Pi_n(V|X_1, \ldots, X_n) (1 - \phi_n)(X_1, \ldots, X_n) \xrightarrow{P_0\text{-a.s.}} 0.
\]

The other term in (3) is treated similarly: \( P_0^n \Pi(V|X_1, \ldots, X_n) \phi_n \leq P_0^n \phi_n \leq e^{-nC}; \) use Markov’s inequality and the first Borel-Cantelli lemma again to show that:

\[
\Pi(V|X_1, \ldots, X_n) \phi_n(X_1, \ldots, X_n) \xrightarrow{P_0\text{-a.s.}} 0.
\]

Combination of (4) and (7) proves that (3) equals zero.
... but there are very nasty examples

**Example 68.1** Consider $P_0$ on $\mathbb{R}$ with Lebesgue density $p_0$ supported on an interval of width one but unknown location. With $\eta(x) > 0$, if $x \in (0, 1)$ and $\eta(x) = 0$ otherwise, and $\theta \in \mathbb{R}$:

$$p_\theta(x) = \eta(x - \theta) \mathbf{1}_{[\theta, \theta+1]}(x)$$

Note that if $\theta \neq \theta'$,

$$-P_{\theta, \eta} \log \frac{p_{\theta', \eta}}{p_{\theta, \eta}} = \infty$$

Kullback-Leibler neighbourhoods are singletons: no prior can be a Kullback-Leibler prior in this model!
Lecture IV
More posterior consistency

In the fourth lecture, we delve deeper into the theory on posterior convergence, motivated by examples that show the limitations of Schwartz's prior mass condition. We prove an alternative consistency theorem that does not rely on KL-priors. We also make contact with Barron's theorem, Walker's theorem and the Ghosal-Ghosh-van der Vaart theorem on the rate of posterior convergence. Particularly, we indicate that GGV-priors suffer from limitations as well, and we derive a theorem on posterior rates of convergence with weaker prior-mass condition.

[arxiv: 1308.1263v3]
Recall Schwartz

**Theorem 70.1** *(Schwartz (1965))*

Let $\mathcal{P}$ be *Hellinger totally bounded* and let $\Pi$ a *KL-prior*, i.e. for $\eta > 0$,

$$\Pi\left( P \in \mathcal{P} : -P_0 \log \frac{p}{p_0} \leq \eta \right) > 0,$$

*Then the posterior is Hellinger consistent at $P_0$.*

**Example 70.2** *Consider $P_0$ on $\mathbb{R}$ with density,*

$$p_\theta(x) = \eta (x - \theta) 1_{[\theta, \theta+1]}(x),$$

*for some $\theta \in \mathbb{R}$. Note that if $\theta \neq \theta'$,*

$$-P_{\theta, \eta} \log \frac{p_{\theta', \eta'}}{p_{\theta, \eta}} = \infty$$

*no prior can be a Kullback-Leibler prior in this model!*
Walker’s theorem

**Theorem 71.1 (Walker (2004))**

Let $\mathcal{P}$ be Hellinger separable. Let $\{V_i : i \geq 1\}$ be a countable cover of $\mathcal{P}$ by balls of radius $\epsilon$. If $\Pi$ is a Kullback-Leibler prior and,

$$
\sum_{i \geq 1} \Pi(V_i)^{1/2} < \infty
$$

then $\Pi( H(P, P_0) > \epsilon \mid X_1, \ldots, X_n ) \xrightarrow{P_0\text{-a.s.}} 0$. 

Theorem 72.1 (Ghosal, Ghosh and van der Vaart, 2000)

Let \((\epsilon_n)\) be such that \(\epsilon_n \downarrow 0\) and \(n\epsilon_n^2 \rightarrow \infty\). Let \(C > 0\) and \(\mathcal{P}_n \subset \mathcal{P}\) be such that, for large enough \(n\),

(i) \(N(\epsilon_n, \mathcal{P}_n, H) \leq e^{-n\epsilon_n^2}\)

(ii) \(\Pi(\mathcal{P} \setminus \mathcal{P}_n) \leq e^{-n\epsilon_n^2(C+4)}\)

(iii) the prior \(\Pi\) is a GGV-prior, i.e.

\[
\Pi\left( P \in \mathcal{P} : -P_0 \log \frac{dP}{dP_0} < \epsilon_n^2, P_0 \left( \log \frac{dP}{dP_0} \right)^2 < \epsilon_n^2 \right) \geq e^{-Cn\epsilon_n^2}
\]

Then \(\Pi( P \in \mathcal{P} : H(P, P_0) > M\epsilon_n \mid X_1, \ldots, X_n) \xrightarrow{P_0} 0\) for some \(M > 0\).
... but here’s another tricky example

**Example 73.1** Consider the distributions $P_a$, $(a \geq 1)$, defined by,

$$p_a(k) = P_a(X = k) = \frac{1}{Z_a} k^a (\log k)^{-3}$$

for all $k \geq 2$, with $Z_a = \sum_{k \geq 2} k^{-a} (\log k)^{-3} < \infty$. For $a = 1$, $b > 1$,

$$-P_a \log \frac{p_b}{p_a} < \infty, \quad P_a \left( \log \frac{p_b}{p_a} \right)^2 = \infty$$

Schwartz’s KL-condition for the prior for the parameter $a$ can be satisfied but GGV priors do not exist.

**Remark 73.2** With $(\log k)^2$ instead of $(\log k)^2$, KL-priors also fail.
Posterior convergence

Recall the prior predictive distribution $P_n^\Pi(A) = \int \varnothing P^n(A) \, d\Pi(P)$.

**Theorem 74.1** Assume that $P_0^n \ll P_n^\Pi$ for all $n \geq 1$. Let $V_1, \ldots, V_N$ be a finite collection of model subsets. If there exist constants $D_i > 0$ and test sequences $(\phi_{i,n})$ for all $1 \leq i \leq N$ such that,

$$P_0^n \phi_{i,n} + \sup_{P \in V_i} P_0^n \frac{dP^n}{dP_0^n}(1 - \phi_{i,n}) \leq e^{-nD_i},$$

for large enough $n$, then any $V \subset \bigcup_{1 \leq i \leq N} V_i$ receives posterior mass zero asymptotically,

$$\Pi(V|X_1, \ldots, X_n) \xrightarrow{P_0-a.s.} 0.$$
Proof

If $\prod(V_i|X_1,\ldots,X_n) \xrightarrow{P_0\text{-a.s.}} 0$ for all $1 \leq i \leq N$ then the assertion is proved. So pick some $i$ and consider,

$$P_0^n \prod(V_i|X_1,\ldots,X_n) \leq P_0^n \phi_n + P_0^n \prod(V_i|X_1,\ldots,X_n)(1 - \phi_n)$$

By Fubini,

$$P_0^n \prod(V_i|X_1,\ldots,X_n)(1 - \phi_n) = \int_V \frac{dP^n}{P^n}(1 - \phi_n) d\prod(P)$$

$$\leq \prod(V_i) \sup_{P \in V_i} P_0 \left( \frac{dP^n}{dP^n_P} \right)(1 - \phi_n) \leq e^{-nD_i}$$

Apply Markov and Borel-Cantelli to conclude that,

$$\limsup_{n \to \infty} \prod(V_i|X_1,\ldots,X_n) = 0.$$
Lemma 76.1 Let $V \subset \mathcal{P}$ be given and assume that $P_0^n(dP^n/dP^n_\Pi) < \infty$ for all $P \in V$. For every $B$ there exists a test sequence $(\phi_n)$ such that,

$$P_0^n\phi_n + \sup_{P \in V} P_0^n dP^n_{\Pi}(1 - \phi_n)$$

$$\leq \inf_{0 \leq \alpha \leq 1} \Pi(B)^{-\alpha} \int \left( \sup_{P \in \text{co}(V)} P_0 \left( \frac{dP}{dQ} \right)^\alpha \right)^n d\Pi(Q|B).$$

i.e. testing power is bounded in terms of Hellinger transforms.

The construction is technically close to that needed for the analysis of posteriors for misspecified models, i.e. when $P_0 \notin \mathcal{P}$ (see, Kleijn and van der Vaart (2006)).
Sketch of the proof

Let $Q_n^n(A)$ be the prior predictive with $\Pi(\cdot|B)$: $P_n^n(A) \geq \Pi(B) Q_n^n(A)$ and using Jensen’s inequality,

$$P_0^n\left(\frac{dP(n)}{dP^n}\right)^\alpha \leq \Pi(B)^{-\alpha} P_0^n\left(\frac{dP(n)}{dQ^n}\right)^\alpha \leq \Pi(B)^{-\alpha} P_0^n \int \left(\frac{dP(n)}{dQ^n}\right)^\alpha d\Pi(Q|B),$$

Hellinger transforms “sub-factorize” over convex hulls of products

$$\sup_{P(n) \in \text{co}(V^n)} \int P_0^n\left(\frac{dP(n)}{dQ^n}\right)^\alpha d\Pi(Q|B) \leq \int \sup_{P(n) \in \text{co}(V^n)} P_0^n\left(\frac{dP(n)}{dQ^n}\right)^\alpha d\Pi(Q|B) \leq \int \left(\sup_{P \in V} P_0\left(\frac{dP}{dQ}\right)^\alpha\right)^n d\Pi(Q|B).$$

(see lemma 3.14 in Kleijn (2003))
A new consistency theorem

For $\alpha \in [0, 1]$, model subsets $B, W$ and a given $P_0$, define,

$$\pi_{P_0}(W, B; \alpha) = \sup_{P \in W} \sup_{Q \in B} P_0 \left( \frac{dP}{dQ} \right)^\alpha$$

**Theorem 78.1** Assume that $P_0^n \ll P_n^\Pi$ for all $n \geq 1$. Let $V_1, \ldots, V_N$ be model subsets. If there exist subsets $B_1, \ldots, B_N$ such that $\Pi(B_i) > 0$,

$$\pi_{P_0}(\text{co}(V_i), B_i) < 1$$

and $\sup_{Q \in B_i} P_0(dP/dQ) < \infty$ for all $P \in V_i$, then,

$$\Pi(V \mid X_1, \ldots, X_n) \xrightarrow{P_0\text{-a.s.}} 0$$

for any $V \subset U_{1 \leq i \leq N} V_i$.

With theorem 78.1 consistency in the fixed-width domain example (for priors of full support on $\mathbb{R}$) is demonstrated without problems.
Flexibility

Given a consistency question, i.e. given \( \mathcal{P} \) and \( V \), the approach is uncommitted regarding the prior and \( B \). We look for neighbourhoods \( B \) of \( P_0 \) (of course such that \( \sup_{Q \in B} P_0(dP/dQ) < \infty \) for all \( P \in V \)), which

(i) allow (uniform) control of \( P_0(p/q)^\alpha \),

(ii) allow convenient choice of a prior such that \( \Pi(B) > 0 \).

The two requirements on \( B \) leave room for a trade-off between being ‘small enough’ to satisfy (i), but ‘large enough’ to enable a choice for \( \Pi \) that leads to (ii).

So we are no longer committed to KL-priors!
Relation with Schwartz’s KL condition

**Lemma 80.1** Let $P_0 \in B \subset \mathcal{P}$ and $W \subset \mathcal{P}$ be given. Assume there is an $\alpha \in (0, 1)$ such that for all $Q \in B$ and $P \in W$, $P_0(dP/dQ)^\alpha < \infty$. Then,

$$\pi_{P_0}(W, B) < 1$$

if and only if,

$$\sup_{Q \in B} -P_0 \log \frac{dQ}{dP_0} < \inf_{P \in W} -P_0 \log \frac{dP}{dP_0}$$
Consistency in KL-divergence

**Theorem 81.1** Let $\Pi$ be a Kullback-Leibler prior. Define $V = \{P \in \mathcal{P} : -P_0 \log(dP/dP_0) \geq \epsilon\}$ and assume that for some KL neighbourhood $B$ of $P_0$, $\sup_{Q \in B} P_0(dP/dQ) < \infty$ for all $P \in V$. Also assume that $V$ is covered by subsets $V_1, \ldots, V_N$ such that,

$$\inf_{P \in \text{co}(V_i)} -P_0 \log \frac{dP}{dP_0} > 0$$

for all $1 \leq i \leq N$. Then,

$$\Pi( -P_0 \log(dP/dP_0) < \epsilon | X_1, \ldots, X_n ) \xrightarrow{P_0 \text{-a.s.}} 1$$
Relation with priors that charge metric balls

Note that if we choose $\alpha = 1/2$,

$$P_0\left(\frac{p}{q}\right)^{1/2} = \int \left(\frac{p_0}{q}\right)^{1/2} p_0^{1/2} p^{1/2} d\mu$$

$$= \int p_0^{1/2} p^{1/2} d\mu + \int \left(\left(\frac{p_0}{q}\right)^{1/2} - 1\right) \left(\frac{p_0}{q}\right)^{1/2} \left(\frac{p}{q}\right)^{1/2} dQ$$

$$\leq 1 - \frac{1}{2} H(P_0, P)^2 + H(P_0, Q) \left\| \frac{p_0}{q} \right\|_{2,Q}^{1/2} \left\| \frac{p}{q} \right\|_{2,Q}^{1/2}.$$ 

So if $\|p/q\|_{2,Q}$ is bounded, a lower bound to $H(\text{co}(V), P_0)$ and an upper bound for $H(Q, P_0)$ guarantee $\pi(\text{co}(V), B; \frac{1}{2}) < 1.$
Borel priors of full support

**Theorem 83.1** Suppose that $\mathcal{P}$ is Hellinger totally bounded. Assume an $L > 0$ and a Hellinger ball $B'$ centred on $P_0$ such that,

$$\|p\|_{2,Q} = \left( \int \frac{p^2}{q} \, d\mu \right)^{1/2} < L, \quad \text{for all } P \in \mathcal{P} \text{ and } Q \in B'$$

If $\Pi(B) > 0$ for all Hellinger neighbourhoods of $P_0$, the posterior is Hellinger consistent, $P_0$-almost-surely.

**Lemma 83.2** If the KL divergence $\mathcal{P} \to \mathbb{R} : Q \mapsto -P \log(dQ/dP)$ is continuous, then a Borel prior of full support is a KL prior. If $\mathcal{P}$ is metrizable, all net priors of full support are KL priors.
Separable models and Barron’s sieves

**Theorem 84.1** Let $V$ be given. Assume that there are $K, L > 0$, submodels $(P_n)_{n \geq 1}$ and a $B$ with $\Pi(B) > 0$, such that,

(i) there is a cover $V_1, \ldots, V_{N_n}$ for $V \cap P_n$ of order $N_n \leq \exp(\frac{1}{2}Ln)$, such that for every $1 \leq i \leq N_n$,

$$\pi_{P_0}(\text{co}(V_i), B) \leq e^{-L}$$

and $\sup_{Q \in B} P_0(dP/dQ) < \infty$ for all $P \in V_i$;

(ii) $\Pi(P \setminus P_n) \leq \exp(-nK)$ and,

$$\sup_{P \in V \setminus P_n} \sup_{Q \in B} P_0\left(\frac{dP}{dQ}\right) \leq e^{\frac{K}{2}}$$

Then $\Pi(V | X_1, \ldots, X_n) \xrightarrow{P_0\text{-a.s.}} 0$. 


A new theorem for separable models

**Theorem 85.1** Assume that \( P_0^n \ll P_n^n \) for all \( n \geq 1 \). Let \( V \) be a model subset with a countable cover \( V_1, V_2, \ldots \) and \( B_1, B_2, \ldots \) such that \( \prod(B_i) > 0 \) and for \( P \in V_i \), we have \( \sup_{Q \in B_i} P_0(dP/dQ) < \infty \). Then,

\[
P^n_0 \prod(V|X_1, \ldots, X_n) \leq \sum_{i \geq 1} \inf_{0 \leq \alpha \leq 1} \frac{\prod(V_i)^\alpha}{\prod(B_i)^\alpha} \pi(\text{co}(V_i), B_i; \alpha)^n.
\]
Relation with Walker’s condition

**Corollary 86.1** Assume that $P_0^n \ll P_n^n$ for all $n \geq 1$. Let $V$ be a subset with a *countable cover* $V_1, V_2, \ldots$ and a $B$ such that $\Pi(B) > 0$ and for all $i \geq 1$, $P \in V_i$, $\sup_{Q \in B} P_0(dP/dQ) < \infty$. Also assume,

$$\sup_{i \geq 1} \pi P_0 \left( \text{co}(V_i), B \right) < 1$$

If the prior satisfies Walker’s condition,

$$\sum_{i \geq 1} \Pi(V_i)^{1/2} < \infty$$

Then $\Pi(V|X_1, \ldots, X_n) \xrightarrow{P_0-\text{a.s.}} 0$. 

86
Posterior rates of convergence

**Theorem 87.1** Assume that $P^n_0 \ll P^n_\Pi$ for all $n \geq 1$. Let $(\epsilon_n)$ be such that $\epsilon_n \downarrow 0$ and $n\epsilon_n^2 \to \infty$. Define $V_n = \{P \in \mathcal{P} : d(P, P_0) > \epsilon_n\}$, submodels $\mathcal{P}_n \subset \mathcal{P}$ and subsets $B_n$ such that $\sup_{Q \in B_n} P_0(p/q) < \infty$ for all $P \in V_n$. Assume that,

(i) there is an $L > 0$ such that $V_n \cap \mathcal{P}_n$ has a cover $V_{n,1}, V_{n,2}, \ldots, V_{n,N_n}$ of order $N_n \leq \exp(\frac{1}{2}Ln\epsilon_n^2)$, such that,

$$\pi_{P_0}(\text{co}(V_{n,i}), B_n) \leq e^{-Ln\epsilon_n^2}$$

for all $1 \leq i \leq N_n$.

(ii) there is a $K > 0$ such that $\Pi(\mathcal{P} \setminus \mathcal{P}_n) \leq e^{-Kn\epsilon_n^2}$ and $\Pi(B_n) \geq e^{-\frac{K}{2}n\epsilon_n^2}$, while also,

$$\sup_{P \in \mathcal{P} \setminus \mathcal{P}_n} \sup_{Q \in B_n} P_0\left(\frac{dP}{dQ}\right) < e^{\frac{K}{4}\epsilon_n^2}$$

Then $\Pi( P \in \mathcal{P} : d(P, P_0) > \epsilon_n \mid X_1, \ldots, X_n) \xrightarrow{P_0} 0$. 

87
Posterior rates with Schwartz’s KL priors

**Theorem 88.1** Let $\epsilon_n$ be such that $\epsilon_n \downarrow 0$ and $n\epsilon_n^2 \to \infty$. For $M > 0$, define $V_n = \{ P \in \mathcal{P} : H(P_0, P) > M\epsilon_n \}$, $B_n = \{ Q \in \mathcal{P} : -P_0 \log(dQ/dP_0) < \epsilon_n^2 \}$. Assume that,

(i) for all $P \in V_n$, $\sup\{ P_0(dP/dQ) : Q \in B_n \} < \infty$

(ii) there is an $L > 0$, such that $N(\epsilon_n, \mathcal{P}, H) \leq e^{Ln\epsilon_n^2}$

(iii) there is a $K > 0$, such that for large enough $n \geq 1$,

$$\Pi\left( P \in \mathcal{P} : -P_0 \log \left( \frac{dP}{dP_0} \right) < \epsilon_n^2 \right) \geq e^{-Kn\epsilon_n^2}$$

then $\Pi( P \in \mathcal{P} : H(P, P_0) > M\epsilon_n | X_1, \ldots, X_n ) \xrightarrow{P_0} 0$, for some $M > 0$.

With theorem 88.1 $\sqrt{n}$-consistency in the heavy-tailed example 73.1 obtains (for uniform priors on bounded intervals in $\mathbb{R}$).
Estimation of support boundary I: model

Model
Define $\Theta = \{(\theta_1, \theta_2) \in \mathbb{R}^2 : 0 < \theta_2 - \theta_1 < \sigma\}$ (for some $\sigma > 0$) and let $H$ be a convex collection of Lebesgue probability densities $\eta : [0, 1] \to [0, \infty)$ with a function $f : (0, a) \to \mathbb{R}$, $f > 0$ such that,

$$\inf_{\eta \in H} \min \left\{ \int_0^\epsilon \eta \, d\mu, \int_{1-\epsilon}^1 \eta \, d\mu \right\} \geq f(\epsilon), \quad (0 < \epsilon < a)$$

The semi-parametric model $\mathcal{P} = \{P_{\theta, \eta} : \theta \in \Theta, \eta \in H\},$

$$p_{\theta, \eta}(x) = \frac{1}{\theta_2 - \theta_1} \eta \left( \frac{x - \theta_1}{\theta_2 - \theta_1} \right) 1\{\theta_1 \leq x \leq \theta_2\}.$$ 

Question
We are interested in marginal consistency for $\theta$. Define the pseudo-metric $d : \mathcal{P} \times \mathcal{P} \to [0, \infty)$,

$$d(P_{\theta, \eta}, P_{\theta', \eta'}) = \max \{|\theta_1 - \theta'_1|, |\theta_2 - \theta'_2|\}.$$ 

We want posterior consistency with $V = \{P_{\theta, \eta} : d(P, P_0) \geq \epsilon\}$. 

89
Estimation of support boundary II: construction

Lemma 90.1 Suppose that $P_0(p/q) < \infty$. Then

$$P_0(p/q)^\alpha|_{\alpha=0} = P_0(p > 0), \quad P_0(p/q)^\alpha|_{\alpha=1} = \int \frac{p_0}{q} 1\{p_0>0\} \, dP.$$ 

Take $B = \{Q : \| (p_0/q) - 1 \|_\infty < \delta \}$,

$$\inf_{0 \leq \alpha \leq 1} P_0 \left( \frac{p}{q} \right)^\alpha \leq (1 + \delta) \min \{ P_0(p > 0), P(p_0 > 0) \}$$

The supports of $p$ and $p_0$ differ by an interval of length $\geq \epsilon$,

$$\min \{ P_0(p > 0), P(p_0 > 0) \} \leq 1 - \frac{f(\epsilon)}{\sigma}.$$ 

Conclude: for every $\epsilon, \delta > 0$,

$$\sup \sup_{Q \in B} \sup_{P \in V} \inf_{0 \leq \alpha \leq 1} P_0 \left( \frac{p}{q} \right)^\alpha \leq (1 + \delta) \left( 1 - \frac{f(\epsilon)}{\sigma} \right) < 1.$$
Estimation of support boundary III: theorem

**Theorem 91.1** Let \( \Theta = \{ (\theta_1, \theta_2) \in \mathbb{R}^2 : 0 < \theta_2 - \theta_1 < \sigma \} \) (for some \( \sigma > 0 \)) and convex \( H \) with associated \( f \) be given. Let \( \Pi \) be a prior on \( \Theta \times H \) such that,

\[
\Pi \left( Q : \| (p_0/q) - 1 \|_\infty < \delta \right) > 0,
\]

for all \( \delta > 0 \). If \( X_1, X_2, \ldots \) form an i.i.d.-\( P_0 \) sample, where \( P_0 = P_{\theta_0, \eta_0} \), then,

\[
\Pi \left( \| \theta - \theta_0 \| < \epsilon \left| X_1, \ldots, X_n \right. \right) \xrightarrow{P_0\text{-a.s.}} 1,
\]

for every \( \epsilon > 0 \).

**Remark 91.2** The \( \sigma \)-restriction on \( \theta_1 - \theta_2 \) can be eliminated with theorem 84.1.
Lecture V

Remote contiguity and Bayes factors

To conclude, we turn to weak consistency for the Dirichlet distribution and to non-\emph{i.i.d.} data with parameter spaces that grow with the sample size. To prove consistency of the posterior, we require the existence of tests, sufficiency of prior mass and a property similar to, but weaker than Le Cam’s notion of contiguity, generalising Schwartz’s Kullback-Leibler condition for the prior. We also consider the consistency of Bayes factors for model selection and hypothesis testing.

[arxiv:1606.XXXX]
The Dirichlet process

**Definition 93.1 (Dirichlet distribution)**

A random variable \( p = (p_1, \ldots, p_k) \) with \( p_l \geq 0 \) and \( \sum_l p_l = 1 \) is Dirichlet distributed with parameter \( \alpha = (\alpha_1, \ldots, \alpha_k) \), \( p \sim D_\alpha \), if it has density

\[
f_\alpha(p) = C(\alpha) \prod_{l=1}^k p_l^{\alpha_l - 1}
\]

**Definition 93.2 (Dirichlet process, Ferguson 1973-74)**

Let \( \alpha \) be a finite measure on \((\mathcal{X}, \mathcal{B})\). The Dirichlet process \( P \sim D_\alpha \) is defined by, (for all finite msb partitions \( A = \{A_1, \ldots, A_k\} \) of \( \mathcal{X} \))

\[
\left( P(A_1), \ldots, P(A_k) \right) \sim D_{(\alpha(A_1), \ldots, \alpha(A_k))}
\]
Weak consistency with Dirichlet priors

**Theorem 94.1** *(Dirichlet consistency)*

Let $X_1, X_2, \ldots$ be an i.i.d.-sample from $P_0$. If $\Pi$ is a Dirichlet prior $D_\alpha$ with finite $\alpha$ such that $\text{supp}(P_0) \subset \text{supp}(\alpha)$, the posterior is consistent at $P_0$ in the weak model topology.

**Remark 94.2** Priors are not necessarily KL for consistency

**Remark 94.3** *(Freedman (1965))*

*Dirichlet distributions are tailfree:* if $A'$ refines $A$ and $A'_{i_1} \cup \ldots \cup A'_{i_l} = A_i$, then $(P(A'_{i_1}|A_i), \ldots, P(A'_{i_l}|A_i) : 1 \leq i \leq k)$ is independent of $(P(A_1), \ldots, P(A_k))$.

**Remark 94.4** $X^n \mapsto \prod (P(A|X^n))$ is $\sigma_n(A)$-measurable where $\sigma_n(A)$ is generated by products of the form $\prod_{i=1}^n B_i$ with $B_i = \{X_i \in A\}$ or $B_i = \{X_i \notin A\}$. 


Stochastic Block Model

**Definition 95.1** At step $n$, nodes belong to one of $K_n$ unobserved classes: $\theta_i$. We estimate $\theta = (\theta_1, \ldots, \theta_n) \in \Theta_n$ upon observation of $X^n = \{X_{ij} : 1 \leq 1 < j \leq n\}$. Edges $X_{ij}$ occur independently with probabilities $Q_{ij}(\theta) = Q(\theta_i, \theta_j)$. The (expected) degree is denoted $\lambda_n$.

A SBM network realisation: $n = 17$, $K_n = 3$, $\lambda_n \approx 2.24$
Bayesian and Frequentist testability

For $B,V$ be two (disjoint) model subsets

**Definition 96.1** Uniform (or minimax) testability

$$\sup_{P \in B} P^n \phi_n \to 0, \quad \sup_{Q \in V} Q^n (1 - \phi_n) \to 0$$

**Definition 96.2** Pointwise testability for all $P \in B, Q \in V$

$$\phi_n \xrightarrow{P\text{-a.s.}} 0, \quad \phi_n \xrightarrow{Q\text{-a.s.}} 1$$

**Definition 96.3** Bayesian testability for $\Pi$-almost-all $P \in B, Q \in V$

$$\phi_n \xrightarrow{P\text{-a.s.}} 0, \quad \phi_n \xrightarrow{Q\text{-a.s.}} 1$$
A posterior concentration inequality

**Lemma 97.1** Let \((\mathcal{P}, \mathcal{G})\) be given. For any prior \(\Pi\), any test function \(\phi\) and any \(B, V \in \mathcal{G}\) such that \(B \cap V = \emptyset\),

\[
\int_B P\Pi(V|X)\,d\Pi(P) \leq \int_B P\phi\,d\Pi(P) + \int_V Q(1 - \phi)\,d\Pi(Q)
\]

**Corollary 97.2** Consequently, in i.i.d.-context, for any sequences \((\Pi_n)\), \((B_n)\), \((V_n)\) such that \(B_n \cap V_n = \emptyset\) and \(\Pi_n(B_n) > 0\), we have,

\[
\int P^n\Pi_n(V_n|X^n)\,d\Pi_n(P|B_n) \\
\leq \frac{1}{\Pi(B_n)} \left( \int_{B_n} P^n\phi_n\,d\Pi_n(P) + \int_{V_n} Q^n(1 - \phi_n)\,d\Pi_n(Q) \right)
\]
Proof

Disintegration: for all $A \in \mathcal{B}^n$ and $V \in \mathcal{G}$,
\[
\int_\mathcal{X} 1_A(X) \Pi(V|X) \, dP^\Pi = \int_V \int_\mathcal{X} 1_A(X) \, dP \, d\Pi(P)
\]
So for any $\mathcal{B}^n$-measurable, simple $f(X) = \sum_{j=1}^J c_j 1_{A_j}(X)$,
\[
\int_\mathcal{X} f(X) \Pi(V|X) \, dP^\Pi = \int_V \int_\mathcal{X} f(X) \, dP \, d\Pi(P)
\]
Taking monotone limits, we see this equality also holds for any positive, measurable $f : \mathcal{X} \to \mathbb{R}$. In particular, with $f(X) = (1 - \phi(X))$,
\[
\int_{\mathcal{D}} P\left((1 - \phi(X)) \Pi(V|X)\right) \, d\Pi(P) = \int_V P(1 - \phi(X)) \, d\Pi(P)
\]
Proof

Since $B \subset \mathcal{P}$ and the integrand is positive,

$$\int_B P((1 - \phi)(X)\Pi(V|X)) d\Pi(P)$$

$$\leq \int_{\mathcal{P}} P((1 - \phi(X))\Pi(V|X)) d\Pi(P) = \int_V P(1 - \phi(X)) d\Pi(P)$$

bring the 2nd term on the l.h.s. to the r.h.s. and divide by $\Pi(B) > 0$,

$$\int P\Pi(V|X) d\Pi(P|B)$$

$$\leq \frac{1}{\Pi(B)} \left( \int_B P\phi(X)\Pi(V|X) d\Pi(P) + \int_V P(1 - \phi)(X) d\Pi(P) \right)$$

$$\leq \frac{1}{\Pi(B)} \left( \int_B P\phi(X) d\Pi(P) + \int_V P(1 - \phi)(X) d\Pi(P) \right)$$
Martingale convergence

**Proposition 100.1** Let $(\mathcal{P}, \mathcal{G}, \Pi)$ be given. For any $B, V \in \mathcal{G}$, the following are equivalent,

(i) There exist Bayesian tests $(\phi_n)$ for $B$ versus $V$;

(ii) There exist tests $(\phi_n)$ such that,
\[
\int_B P^n \phi_n \, d\Pi(P) + \int_V Q^n (1 - \phi_n) \, d\Pi(Q) \to 0,
\]

(iii) For $\Pi$-almost-all $P \in B$, $Q \in V$,
\[
\Pi(V|X^n) \xrightarrow{P\text{-a.s.}} 0, \quad \Pi(B|X^n) \xrightarrow{Q\text{-a.s.}} 0
\]

**Remark 100.2** Interpretation distinctions between model subsets are Bayesian testable, iff they are picked up by the posterior asymptotically, if(f), the Bayes factor for $B$ versus $V$ is consistent
Proof

Condition \((i)\) implies \((ii)\) by dominated convergence. Assume \((ii)\) and note that by the previous lemma,

\[
\int P^n \Pi(V|X^n) \, d\Pi(P|B) \to 0.
\]

Martingale convergence (in \(L^1(\mathcal{X}^\infty \times \mathcal{P})\)) implies that there is a \(g : \mathcal{X}^\infty \to [0, 1]\) such that,

\[
\int P^\infty \left| \Pi(V|X^n) - g(X^n) \right| \, d\Pi(P, B) \to 0,
\]

So \(\int P^\infty g \, d\Pi(P|B) = 0\), so \(g = 0, P^\infty\)-almost-surely for \(\Pi\)-almost-all \(P \in B\). Using martingale convergence again (now in \(L^\infty(\mathcal{X}^\infty \times \mathcal{P})\)), conclude \(\Pi(V|X^n) \to 0\) \(P^\infty\)-almost-surely for \(\Pi\)-almost-all \(P \in B\), i.e. \((iii)\) follows.

Choose \(\phi(X^n) = \Pi(V|X^n)\) to conclude that \((i)\) follows from \((iii)\).
Prior-almost-sure consistency

**Theorem 102.1** Let Hausdorff $\mathcal{P}$ with Borel prior $\Pi$ be given. Assume that for $\Pi$-almost-all $P \in \mathcal{P}$ and any open nbd $U$ of $P$, there exist a $B \subset U$ with $\Pi(B) > 0$ and Bayesian tests $(\phi_n)$ for $B$ versus $\mathcal{P} \setminus U$. Then the posterior is consistent at $\Pi$-almost-all $P \in \mathcal{P}$.

**Remark 102.2** Let $\mathcal{P}$ be a Polish space and assume that all $P \mapsto P^n(A)$ are Borel measurable. Then, for any prior $\Pi$, any Borel set $V \subset \mathcal{P}$ is Bayesian testable versus $\mathcal{P} \setminus V$.

**Corollary 102.3** Doob’s theorem (1948), and much more!
Le Cam’s inequality

**Definition 103.1** For $B \in \mathcal{G}$ such that $\prod(B) > 0$, the local prior predictive distribution is $P_n^{\prod|B} = \int P^n \, d\prod(P|B)$.

**Remark 103.2** *(Le Cam, unpublished (197?) and (1986))*
Rewrite the posterior concentration inequality

$$P^n_0 \cap(V_n|X^n) \leq \left\|P^n_0 - P^n_{\prod|B_n}\right\|$$

$$+ \int P^n \phi_n \, d\prod(P|B_n) + \frac{\prod(V_n)}{\prod(B_n)} \int Q^n(1 - \phi_n) \, d\prod(Q|V_n)$$

**Remark 103.3** For some $b_n \downarrow 0$, $B_n = \{P \in \mathcal{P} : \|P^n - P^n_0\| \leq b_n\}$,

$$a_n^{-1} \prod(B_n) \rightarrow \infty$$

**Remark 103.4** Useful in parametric models but “a considerable nuisance” [sic] *(Le Cam (1986)) in non-parametric context*
Schwartz's theorem revisited

**Remark 104.1** Suppose that for all $\delta > 0$, there is a $B$ s.t. $\Pi(B) > 0$ and for all $P \in B$ and large enough $n$

$$P_0^n \Pi(V|X^n) \leq e^{n\delta} P^n \Pi(V|X^n)$$

then (by Fatou) for large enough $m$

$$\sup_{n \geq m} \left[ (P_0^n - e^{n\delta} P_n^{\Pi|B}) \Pi(V|X^n) \right] \leq 0$$

**Theorem 104.2** Let $\mathcal{P}$ be a model with $KL$-prior $\Pi; P_0 \in \mathcal{P}$. Let $B, V \in \mathcal{G}$ be given and assume that $B$ contains a $KL$-neighbourhood of $P_0$. If there exist Bayesian tests for $B$ versus $V$ of exponential power then

$$\Pi(V|X^n) \xrightarrow{P_0-a.s.} 0$$

**Corollary 104.3** (*Schwartz's theorem*)
Remote contiguity

Definition 105.1 Given \((P_n), (Q_n)\) of prob msr's, \(Q_n\) is contiguous w.r.t. \(P_n\) \((Q_n \prec P_n)\), if for any \((\psi_n)\), \(\psi_n : \mathcal{X}^n \to [0, 1]\)

\[ P_n \psi_n = o(1) \implies Q_n \psi_n = o(1) \]

Definition 105.2 Given \((P_n), (Q_n)\) of prob msr's and a \(a_n \downarrow 0\), \(Q_n\) is \(a_n\)-remotely contiguous w.r.t. \(P_n\) \((Q_n \prec a_n^{-1} P_n)\), if for any sequence \((\psi_n)\), \(\psi_n : \mathcal{X}^n \to [0, 1]\)

\[ P_n \psi_n = o(a_n) \implies Q_n \psi_n = o(1) \]

Remark 105.3 Contiguity is stronger than remote contiguity
note that \(Q_n \prec P_n\) iff \(Q_n \prec a_n^{-1} P_n\) for all \(a_n \downarrow 0\).

Definition 105.4 Hellinger transform \(\psi(P, Q; \alpha) = \int (dP)^\alpha (dQ)^{1-\alpha}\)
Le Cam’s first lemma

**Lemma 106.1** Given \((P_n), (Q_n)\) like above, \(Q_n \prec P_n\) iff any of the following holds:

(i) If \(T_n \xrightarrow{P_n} 0\), then \(T_n \xrightarrow{Q_n} 0\)

(ii) Given \(\epsilon > 0\), there is a \(b > 0\) such that \(Q_n(dQ_n/dP_n > b) < \epsilon\)

(iii) Given \(\epsilon > 0\), there is a \(c > 0\) such that \(\|Q_n - Q_n \wedge cP_n\| < \epsilon\)

(iv) If \(dP_n/dQ_n \xrightarrow{Q_n-w.} f\) along a subsequence, then \(P(f > 0) = 1\)

(v) If \(dQ_n/dP_n \xrightarrow{P_n-w.} g\) along a subsequence, then \(Eg = 1\)

(vi) \(\lim \inf_n \psi(P_n, Q_n; \alpha) \to 1\) as \(\alpha \uparrow 1\)
Criteria for remote contiguity

Lemma 107.1 Given \((P_n), (Q_n), a_n \downarrow 0\), \(Q_n \prec a_{n}^{-1}P_n\) if any of the following holds:

(i) For any bnd msb \(T_n : \mathcal{X}^n \to \mathbb{R}\), \(a_{n}^{-1}T_n \overset{P_n}{\longrightarrow} 0\), implies \(T_n \overset{Q_n}{\longrightarrow} 0\)

(ii) Given \(\epsilon > 0\), there is a \(\delta > 0\) s.t. \(Q_n (dP_n/dQ_n > \delta a_n) < \epsilon\) f.l.e.n.

(iii) There is a \(b > 0\) s.t. \(\lim \inf_{n \to \infty} ba_{n}^{-1} P_n (dQ_n/dP_n > ba_{n}^{-1}) = 1\)

(iv) Given \(\epsilon > 0\), there is a \(c > 0\) such that \(\|Q_n - Q_n \wedge c a_{n}^{-1}P_n\| < \epsilon\)

(v) Under \(Q_n\), \((a_n dQ_n/dP_n)\) are r.v.’s and every subseq has a weakly convergent subseq

(vi) \(\lim \inf_n \lim_{\alpha \uparrow 1} a_{n}^{-\alpha} \psi(P_n, Q_n; \alpha) > 0\)
Beyond Schwartz

**Theorem 108.1** Let $(\mathcal{P}, \mathcal{G})$ with priors $(\Pi_n)$ and $(X_1, \ldots, X_n) \sim P_0^n$ be given. Assume there are $B_n, V_n \in \mathcal{G}$ and $a_n, b_n \geq 0$, $a_n \downarrow 0$ s.t.

(i) There exist Bayesian tests for $B_n$ versus $V_n$ of power $a_n$,

$$\int_{B_n} P^n \phi_n d\Pi_n(P) + \int_{V_n} Q^n(1 - \phi_n) d\Pi_n(Q) \leq a_n$$

(ii) The prior mass of $B_n$ is lower-bounded by $b_n$, $\Pi_n(B_n) \geq b_n$

(iii) The sequence $P^n_0$ satisfies $P^n_0 \prec b_n a_n^{-1} P^n_0|B_n$

Then $\Pi_n(V_n|X^n) \xrightarrow{P_0} 0$
Remark 109.1 (Schwartz (1965))
Take $P_0 \in \mathcal{P}$, and define
\[
V_n = V := \{ P \in \mathcal{P} : H(P, P_0) \geq \epsilon \}
\]
\[
B_n = B := \{ P : -P_0 \log dP/dP_0 < \epsilon^2 \}
\]
with $a_n$ and $b_n$ of form $\exp(-nK)$. With $N(\epsilon, \mathcal{P}, H) < \infty$, the theorem proves Hellinger consistency with KL-priors.

Remark 109.2 (Ghosal-Ghosh-vdVaart (2000))
Take $P_0 \in \mathcal{P}$, and define
\[
V_n = \{ P \in \mathcal{P} : H(P, P_0) \geq \epsilon_n \}
\]
\[
B_n = B := \{ P : -P_0 \log dP/dP_0 < \epsilon_n^2, \ P_0 \log^2 dP/dP_0 < \epsilon_n^2 \}
\]
with $a_n$ and $b_n$ of form $\exp(-Kn\epsilon_n^2)$. With $\log N(\epsilon_n, \mathcal{P}, H) \leq n\epsilon_n^2$, the theorem then proves Hellinger consistency at rate $\epsilon_n$ with GGV-priors. Other $B_n$ are possible! (see Kleijn and Zhao (201x))
Application to consistency II

Remark 110.1  \textit{Dirichlet posteriors} $X^n \mapsto \prod(P(A)|X^n)$ \textit{are msb} $\sigma_n(A)$ where $\sigma_n(A)$ \textit{is generated by products of the form} $\prod_{i=1}^n B_i$ \textit{with} $B_i = \{X_i \in A\}$ \textit{or} $B_i = \{X_i \notin A\}$.

Remark 110.2 (Freedman (1965), Ferguson (1973), Lo (1984), ...) \textit{Take} $P_0 \in \mathcal{P}$, \textit{and define}

$$V_n = V := \{P \in \mathcal{P} : |(P_0 - P)f| \geq 2\epsilon\}$$

$$B_n = B := \{P : |(P_0 - P)f| < \epsilon\}$$

\textit{for some bounded, measurable} $f$. \textit{Impose remote contiguity only for} $\psi_n$ \textit{that are} $\sigma_n(A)$-\textit{measurable!} \textit{Take} $a_n$ \textit{and} $b_n$ \textit{of form} $\exp(-nK)$. \textit{The theorem then proves} $\mathcal{T}_1$ \textit{consistency with a Dirichlet prior} $D_\alpha$, \textit{if} $\text{supp}(P_0) \subset \text{supp}(\alpha)$.
Consistent Bayes factors

**Theorem 111.1** Let \((\mathcal{P}, \mathcal{G})\) with priors \((\Pi_n)\) and \((X_1, \ldots, X_n) \sim P_0^n\) be given. Assume there are \(B, V \in \mathcal{G}\) with \(\Pi(B), \Pi(V) > 0\) and \(a_n \geq 0, a_n \downarrow 0\) s.t.

\(\text{(i) There exist Bayesian tests for } B_n \text{ versus } V_n \text{ of power } a_n,\)

\[
\int_{B_n} P^n \phi_n d\Pi_n(P) + \int_{V_n} Q^n (1 - \phi_n) d\Pi_n(Q) \leq a_n
\]

\(\text{(ii) For every } P \in B, \ P^n \prec a_{n}^{-1} P_n^{\Pi_n|B}\)

\(\text{(iii) For every } Q \in V, \ Q^n \prec a_{n}^{-1} P_n^{\Pi_n|V}\)

Then the posterior odds or Bayes factors,

\[
B_n = \frac{\Pi(B|X^n) \Pi(V)}{\Pi(V|X^n) \Pi(B)}
\]

for \(B\) versus \(V\) are consistent.