Victoria University, Wellington, New Zealand, 11 January 2016

Priors for the frequentist, consistency beyond Schwartz

Bas Kleijn, KdV Institute for Mathematics

UNIVERSITEIT VAN AMSTERDAM

Part I

Introduction

Bayesian and Frequentist statistics

sample space	$(\mathscr{X},\mathscr{B})$	measurable space
i.i.d. data	$X^n = (X_1,\ldots,X_n) \in \mathscr{X}^n$	frequentist/Bayesian
model	$(\mathscr{P},\mathscr{G})$	model subsets $B, V \in \mathscr{G}$
prior	$\Pi : \mathscr{G} \to [0,1]$	probability measure
posterior	$\Pi(\,\cdot\,\vert X^n) : \mathscr{G} \to [0,1]$	Bayes's rule, inference

Frequentist	assume there is P_0	$X^n \sim P_0^n$
Bayes	assume $P \sim \Pi$	$X^n \,\vert\, P \sim P^n$

Definition of the posterior

Definition 4.1 *Assume that all $P \mapsto P^n(A)$ are $\mathscr{G}$-measurable. Given prior Π, a posterior is any $\Pi(\cdot\,|X^n = \cdot) : \mathscr{G} \times \mathscr{X}^n \to [0,1]$*

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

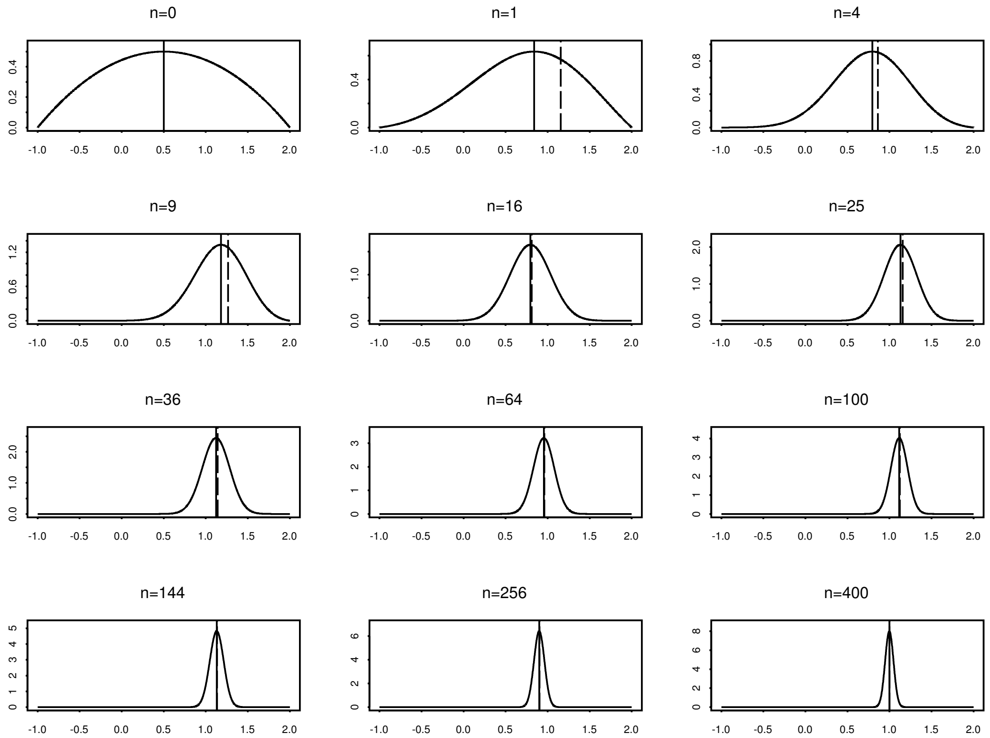
(ii) (Disintegration) For all $A \in \mathscr{B}^n$ and $G \in \mathscr{G}$

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

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

Remark 4.2 *For frequentists $(X_1, \ldots, X_n) \sim P_0^n$, so assume $P_0^n \ll P_n^\Pi$*

Asymptotic consistency of the posterior



Definition 5.1 *Given a model $\mathscr{P}$ with topology and a Borel prior Π, the posterior is consistent at $P \in \mathscr{P}$ if for every open nbd U of P*

$$\Pi(U|X^n) \xrightarrow{P} 1$$

Doob's and Schwartz's consistency theorems

Theorem 6.1 *(Doob (1948))*
Let $\mathscr{P}$ and $\mathscr{X}$ be Polish spaces. Assume that $P \mapsto P^n(A)$ is Borel measurable for all n, A. Then for any prior Π, the posterior is consistent at P, for Π-almost-all $P \in \mathscr{P}$

Remark 6.2 *(Schwartz (1961), Freedman (1963)) Not frequentist!*

Theorem 6.3 *(Schwartz (1965))*
Let $X_1, X_2, \ldots$ be an i.i.d.-sample from $P_0 \in \mathscr{P}$. Let $\mathscr{P}$ be Hellinger totally bounded and let Π be a Kullback-Leibler (KL-)prior, i.e.

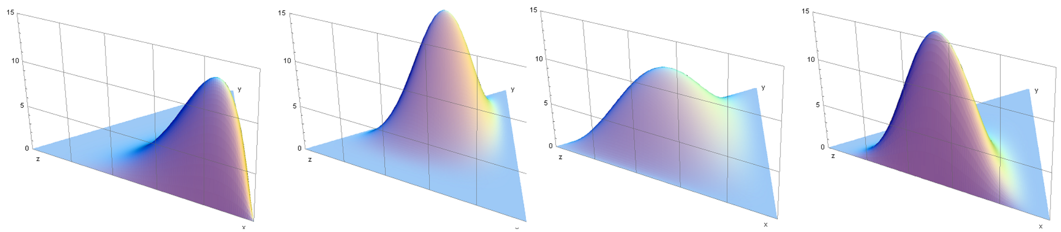
$$\Pi\Big(P \in \mathscr{P} \,:\, -P_0 \log dP/dP_0 < \epsilon \Big) > 0$$

for all $\epsilon > 0$. Then the posterior is consistent at P_0 in the Hellinger topology

The Dirichlet process

Definition 7.1 *(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 7.2 *(Dirichlet process, Ferguson 1973-74)*
Let α *be a finite measure on* $(\mathscr{X}, \mathscr{B})$. *The Dirichlet process* $P \sim D_\alpha$ *is defined by, (for all finite msb partitions* $A = \{A_1, \ldots, A_k\}$ *of* $\mathscr{X}$*)*

$$\big(P(A_1), \ldots, P(A_k)\big) \sim D_{(\alpha(A_1), \ldots, \alpha(A_k))}$$

Weak consistency with Dirichlet priors

Theorem 8.1 *(Dirichlet consistency)*
Let $X_1, X_2, \ldots$ be an i.i.d.-sample from P_0 If Π is a Dirichlet prior D_α with finite α such that $\mathrm{supp}(P_0) \subset \mathrm{supp}(\alpha)$, the posterior is consistent at P_0 in the weak model topology

Remark 8.2 *Priors are not necessarily KL for consistency*

Remark 8.3 *(Freedman (1965))*
Dirichlet distributions are tailfree: if A' refines A and $A'_{i1} \cup \ldots \cup A'_{il_i} = A_i$, then $(P(A'_{i1}|A_i), \ldots, P(A'_{il_i}|A_i) : 1 \leq i \leq k)$ is independent of $(P(A_1), \ldots, P(A_k))$.

Remark 8.4 *$X^n \mapsto \Pi(P(A)|X^n)$ 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 $B_i = \{X_i \notin A\}$.*

Bayesian and Frequentist testability

For B, V be two (disjoint) model subsets

Definition 9.1 *Uniform (or minimax) testability*

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

Definition 9.2 *Pointwise testability for all $P \in B$, $Q \in V$*

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

Definition 9.3 *Bayesian testability for Π-almost-all $P \in B$, $Q \in V$*

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

Part II

Bayesian testability

and prior-a.s.-consistency

A posterior concentration inequality

Lemma 11.1 *Let* $(\mathscr{P}, \mathscr{G})$ *be given. For any prior* Π*, any test function* ϕ *and any* $B, V \in \mathscr{G}$ *such that* $B \cap V = \varnothing$ *and* $\Pi(B) > 0$*,*

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

Corollary 11.2 *Consequently, in i.i.d.-context, for any sequences* (Π_n)*,* (B_n)*,* (V_n) *such that* $B_n \cap V_n = \varnothing$ *and* $\Pi_n(B_n) > 0$*, we have,*

$$\int P^n \Pi(V_n|X^n)\, d\Pi_n(P|B_n)$$
$$\leq \frac{1}{\Pi(B_n)} \Bigl(\int_{B_n} P^n \phi_n\, d\Pi_n(P) + \int_{V_n} Q^n (1-\phi_n)\, d\Pi_n(Q) \Bigr)$$

Martingale convergence

Proposition 12.1 *Let* $(\mathscr{P}, \mathscr{G}, \Pi)$ *be given. For any* $B, V \in \mathscr{G}$, *the following are equivalent,*

(i) There exist Bayesian tests (ϕ_n) *for* B *versus* V*;*

(ii) There exist tests (ϕ_n) *such that,*

$$\int_B P^n \phi_n \, d\Pi(P) \to 0 + \int_V Q^n (1 - \phi_n) \, d\Pi(Q) \to 0,$$

(iii) For Π*-almost-all* $P \in B$, $Q \in V$,

$$\Pi(V|X^n) \xrightarrow{P\text{-a.s.}} 0, \qquad \Pi(B|X^n) \xrightarrow{Q\text{-a.s.}} 0$$

Remark 12.2 *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*

Prior-almost-sure consistency

Theorem 13.1 *Let Hausdorff $\mathscr{P}$ with Borel prior Π be given. Assume that for Π-almost-all $P \in \mathscr{P}$ and any open nbd U of P, there exist a $B \subset U$ with $\Pi(B) > 0$ and Bayesian tests (ϕ_n) for B versus $\mathscr{P} \setminus U$. Then the posterior is consistent at Π-almost-all $P \in \mathscr{P}$*

Remark 13.2 *Let $\mathscr{P}$ be a Polish space and assume that all $P \mapsto P^n(A)$ are Borel measurable. Then, for any prior Π, any Borel set $V \subset \mathscr{P}$ is Bayesian testable versus $\mathscr{P} \setminus V$.*

Corollary 13.3 *Doob's theorem (1948), and much more!*

Part III

Frequentism

Le Cam's inequality

Definition 15.1 *For* $B \in \mathscr{G}$ *such that* $\Pi(B) > 0$, *the local prior predictive distribution is* $P_n^{\Pi|B} = \int P^n \, d\Pi(P|B)$.

Remark 15.2 *(Le Cam, unpublished (197?) and (1986))*
Rewrite the posterior concentration inequality

$$P_0^n \Pi(V_n|X^n) \leq \left\| P_0^n - P_n^{\Pi|B_n} \right\| + \int P^n \phi_n \, d\Pi(P|B_n) + \frac{\Pi(V_n)}{\Pi(B_n)} \int Q^n (1 - \phi_n) \, d\Pi(Q|V_n)$$

Remark 15.3 *For some* $b_n \downarrow 0$, $B_n = \{P \in \mathscr{P} : \|P^n - P_0^n\| \leq b_n\}$,

$$a_n^{-1} \, \Pi(B_n) \to \infty$$

Remark 15.4 *Useful in parametric models but "a considerable nuisance" [sic] (Le Cam (1986)) in non-parametric context*

Schwartz's theorem revisited

Remark 16.1 *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

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

Theorem 16.2 *Let* $\mathscr{P}$ *be a model with KL-prior* Π; $P_0 \in \mathscr{P}$. *Let* $B, V \in \mathscr{G}$ *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 16.3 *(Schwartz's theorem (1965))*

Remote contiguity

Definition 17.1 *Given* (P_n), (Q_n) *of prob msr's,* Q_n *is contiguous w.r.t.* P_n ($Q_n \lhd P_n$)*, if for any* (ψ_n), $\psi_n : \mathscr{X}^n \to [0,1]$

$$P_n\psi_n = o(1) \quad \Rightarrow \quad Q_n\psi_n = o(1)$$

Definition 17.2 *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 ($Q_n \lhd a_n^{-1} P_n$)*, if for any sequence* (ψ_n), $\psi_n : \mathscr{X}^n \to [0,1]$

$$P_n\psi_n = o(a_n) \quad \Rightarrow \quad Q_n\psi_n = o(1)$$

Remark 17.3 *Contiguity is stronger than remote contiguity*
note that $Q_n \lhd P_n$ *iff* $Q_n \lhd a_n^{-1} P_n$ *for all* $a_n \downarrow 0$.

Definition 17.4 *Hellinger transform* $\rho_n(\alpha) = \int (dP_n)^\alpha (dQ_n)^{1-\alpha}$

Le Cam's first lemma

Lemma 18.1 *Given* (P_n), (Q_n) *like above,* $Q_n \lhd P_n$ *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* $\|Q_n - Q_n \wedge c\,P_n\| < \epsilon$

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

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

(vi) $\liminf_n \phi_n(\alpha) \to 1$ *as* $\alpha \uparrow 1$

Criteria for remote contiguity

Lemma 19.1 *Given* (P_n), (Q_n), $a_n \downarrow 0$, $Q_n \triangleleft a_n^{-1} P_n$ *if any of the following holds:*

(i) *If, for all* $\epsilon > 0$, $P_n(T_n > \epsilon) = o(a_n)$, *then* $T_n \xrightarrow{Q_n} 0$

(ii) *Given* $\epsilon > 0$, *there is a* $b > 0$ *such that* $Q_n(dQ_n/dP_n > b\, a_n^{-1}) < \epsilon$

(iii) *Given* $\epsilon > 0$, *there is a* $c > 0$ *such that* $\|Q_n - Q_n \wedge c\, a_n^{-1} P_n\| < \epsilon$

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

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

(vi) $\limsup_n \inf_\alpha a_n{}^{-\alpha} \rho_n(\alpha) \to 0$

Beyond Schwartz

Theorem 20.1 *Let* $(\mathscr{P}, \mathscr{G})$ *with priors* (Π_n) *and* $(X_1, \ldots, X_n) \sim P_0^n$ *be given. Assume there are* $B_n, V_n \in \mathscr{G}$ *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_0^n *satisfies* $P_0^n \lhd b_n a_n^{-1} \, P_n^{\Pi_n | B_n}$

Then $\Pi_n(V_n | X^n) \xrightarrow{P_0} 0$

Application to consistency I

Remark 21.1 *(Schwartz (1965))*
Take $P_0 \in \mathscr{P}$, and define

$$V_n = V := \{P \in \mathscr{P} : H(P, P_0) \geq \epsilon\}$$
$$B_n = B := \{P : -P_0 \log dP/dP_0 < \epsilon^2\}$$

with a_n and b_n of form $\exp(-nK)$. With $N(\epsilon, \mathscr{P}, H) < \infty$, the theorem proves Hellinger consistency with KL-priors.

Remark 21.2 *(Ghosal-Ghosh-vdVaart (2000))*
Take $P_0 \in \mathscr{P}$, and define

$$V_n = \{P \in \mathscr{P} : H(P, P_0) \geq \epsilon_n\}$$
$$B_n = B := \{P : -P_0 \log dP/dP_0 < \epsilon_n^2,\ P_0 \log^2 dP/dP_0 < \epsilon_n^2\}$$

with a_n and b_n of form $\exp(-Kn\epsilon_n^2)$. With $\log N(\epsilon_n, \mathscr{P}, H) \leq n\epsilon_n^2$, the theorem then proves Hellinger consistency at rate ϵ_n with GGV-priors.
Other B_n are possible! (see Kleijn and Zhao (201x))

Application to consistency II

Remark 22.1 *Dirichlet posteriors* $X^n \mapsto \Pi(P(A)|X^n)$ *are msb* $\sigma_n(A)$ *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* $B_i = \{X_i \notin A\}$.

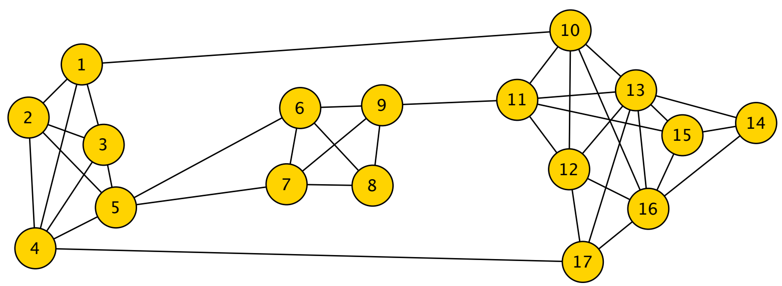
Remark 22.2 *(Freedman (1965), Ferguson (1973), Lo (1984), ...)*
Take $P_0 \in \mathscr{P}$, *and define*

$$V_n = V := \{P \in \mathscr{P} : |(P_0 - P)f| \geq 2\epsilon\}$$
$$B_n = B := \{P : |(P_0 - P)f| < \epsilon\}$$

for some bounded, measurable f. *Impose remote contiguity only for* ψ_n *that are* $\sigma_n(A)$*-measurable! Take* a_n *and* b_n *of form* $\exp(-nK)$. *The theorem then proves* $\mathscr{T}_1$ *consistency with a Dirichlet prior* D_α, *if* $\mathrm{supp}(P_0) \subset \mathrm{supp}(\alpha)$.

Stochastic Block Model

Definition 23.1 *At step n, nodes belong to one of K_n unobserved classes: θ_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 λ_n.*



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

Testing in the Stochastic Block Model

Assume there is a q s.t. $0 < q < Q_{ij} < 1 - q < 1$

Lemma 24.1 *For given $\theta, \theta' \in \Theta_n$, there exists a test ϕ_n s.t.*

$$P_{\theta,n}\phi_n \vee P_{\theta',n}(1 - \phi_n) \leq e^{-8q(1-q)\sum_{i<j}(Q_{ij}(\theta)-Q_{ij}(\theta'))^2}$$

Lemma 24.2 *For given, $B_n, V_n \subset \Theta_n$, there exists a test ϕ_n s.t.*

$$\sup_{\theta \in B_n} P_{\theta,n}\phi_n \leq e^{-8q(1-q)\,a_n^2 + \log \#(V_n)}$$

$$\sup_{\theta' \in V_n} P_{\theta',n}(1 - \phi_n) \leq e^{-8q(1-q)\,a_n^2 + \log \#(B_n)}$$

where $a_n^2 = \inf_{\theta \in B_n} \inf_{\theta \in V_n} \sum_{i<j}(Q_{ij}(\theta) - Q_{ij}(\theta'))^2$

Note: $\log \#(V_n), \log \#(B_n) \leq n \log(K_n)$

Consistency in the Stochastic Block Model

Theorem 25.1 *(Bickel, Chen (2009))*
If $K_n = K$, $\lambda_n / \log(n) \to \infty$, the ML estimator for θ is consistent.

Consistency requires not a single mistake in $\theta = (\theta_1, \theta_2, \ldots)$.

Conjecture 25.2 *Give Θ_n a prior Π_n s.t. $\Pi(\{\theta\}) > 0$ for all $\theta \in \Theta_n$. Unless very special conditions for q, K_n, λ_n are satisfied, the posterior is not consistent.*

Theorem 25.3 *(see also Choi, Wolfe, Airoldi (2011))*
Given uniform priors Π_n on Θ_n, posteriors are consistent for hypotheses $B_n, V_n \subset \Theta_n$, if,

$$4q_n(1-q_n)a_n^2 \geq n \log(K_n)$$

for all n large enough.

Open questions

Tailfreeness is too strong for (weak or $\mathscr{T}_1$) consistency; the Dirichlet example allows generalization. Can we show that Pitman-Yor is inconsistent? What about inverse Gaussian? Gibbs-type? Can we construct a family of consistent priors around Dirichlet?

Which pairs of (FDR-type) hypotheses in SBM are testable and which are not? Can the method be applied to other network models?

What is the extent of the usefulness of remote contiguity?

The weak and $\mathscr{T}_1$ topologies

Uniformity $\mathscr{U}_n$: basis is finite intersections of $W_{n,f}$'s

$$W_{n,f} = \{(P,Q) : |(P^n - Q^n)f| < \epsilon\}, \quad (\text{bnd msb } f : \mathscr{X}^n \to [0,1])$$

and $\mathscr{U}_\infty = \cup_n \mathscr{U}_n \subset \mathscr{U}_H$. Weak $\mathscr{U}_C \subset \mathscr{U}_1$ for (bnd cont $f : \mathscr{X} \to [0,1]$)

$\mathscr{T}_C \subset \mathscr{T}_1 \subset \mathscr{T}_n \subset \mathscr{T}_\infty \subset \mathscr{T}_H$ are completely regular

$(\mathscr{P}, \mathscr{T}_1)$ sep $\Leftrightarrow$ $(\mathscr{P}, \mathscr{T}_\infty)$ sep $\Leftrightarrow$ $(\mathscr{P}, \mathscr{T}_H)$ sep $\Leftrightarrow$ $\mathscr{P}$ is dominated

Any model $\mathscr{P}$ is pre-cpt in $\mathscr{T}_C, \mathscr{T}_1, \mathscr{T}_n, \mathscr{T}_\infty$: any complete $\mathscr{P}$ is cpt

$\mathscr{T}_C$ is metrizable, even Polish if $\mathscr{X}$ is Polish

$\mathscr{T}_1, \mathscr{T}_n, \mathscr{T}_\infty$ are metrizable iff $\mathscr{X}$ discrete