# Posterior Consistency in Bayesian Nonparametric Models

Flavio Argentieri

Bocconi University

April 20, 2022

# Framework

# Modeling the dataset

- **Dataset**: $(x_i)_{i=1}^{n}$, where $x_i \in \mathcal{X}$ for all $i$
- $\mathcal{X}$ is a **Polish space**, endowed such space with its Borel sigma algebra, $\mathcal{B}(\mathcal{X})$
- We consider a **random sequence** defined on an underlying experiment:

$$(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{X} (\mathcal{X}^{\infty}, \mathcal{B}(\mathcal{X}^{\infty})) \xrightarrow{X^{(n)}} (\mathcal{X}^{n}, \mathcal{B}(\mathcal{X}_{\rangle}{}^{n}))$$

where $X := (X_i)_{i \geq 1}$ and $X^{(n)} := (X_i)_{i=1}^{n}$

# Random Parameter

- We introduce a **random parameter**:

$$\theta : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Theta, \mathcal{B}(\Theta))$$

where $\Theta$ is Polish as well, and $\theta$ is distributed according to $\Pi$, the **prior**

- For all $n \in \mathbb{N}$ we define a **regular conditional probability** on $\mathcal{X}^n$ through the family:

$$\mathcal{M}_n = \{P_\theta^n : \theta \in \Theta\}$$

- We use a product structure for $P_\theta^n$, so that $X_i|\theta \overset{iid}{\sim} P_\theta$.

# Existence

- Under regularity conditions, there exists a **Kolmogorov extension** $P_\theta^\infty$ of $\{P_\theta^n\}$ on $\mathcal{X}^\infty$.
- In turn, **there exists a probability measure P** on $\mathcal{X}^\infty \times \Theta$, such that for all $A \in \mathcal{B}(\mathcal{X}^\infty)$, for all $B \in \mathcal{B}(\Theta)$

$$\mathbb{P}(X \in A, \theta \in B) = \mathbf{P}(A \times B) = \int_B P_\theta^\infty(A) d\Pi(\theta)$$

- In turn, there exists a regular conditional probability $\bar{\Pi}_n(\cdot | X^{(n)})$ on $\mathcal{X}^\infty \times \Theta$ from which we can define both a **posterior** $\Pi_n$ on $\Theta$ and a **predictive** $P_{\Pi_n}$ on $\mathcal{X}^\infty$, integrating out $\mathcal{X}^\infty$ or $\Theta$, respectively.

# Exchangeability

# Exchangeability

**Definition**

A random sequence $X = (X_i)_{i \geq 1}$ is exchangeable if and only if for all $n \in \mathbb{N}$, for all bijective functions $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ the distribution of $(X_1, \dots, X_n)$ is the same as the distribution of $(X_{\sigma(1)}, \dots, X_{\sigma(n)})$.

# De Finetti Theorem

### Theorem

*Let $\mathcal{P}_{\mathcal{X}}$ be the space of probability measures on $\mathcal{X}$, and let $\mathcal{B}(\mathcal{P}_{\mathcal{X}})$ be its Borel sigma-algebra. $X$ is exchangeable if and only if there exists a* **random probability measure** *$\theta$ on $(\mathcal{P}_{\mathcal{X}}, \mathcal{B}(\mathcal{P}_{\mathcal{X}}))$ distributed according to $\Pi$ such that, for all $B \in \mathcal{B}(\mathcal{X}^{\infty})$:*

$$\mathbb{P}(X \in B) = \int_{\Theta} \theta^{\infty}(B) d\Pi(\theta)$$

*where $\theta^{\infty}$ is a probability measure on $(\mathcal{X}^{\infty}, \mathcal{B}(\mathcal{X}^{\infty}))$ such that for all rectangles $C = \times_{i=1}^{\infty} C_i$, with $C_i \in \mathcal{B}(\mathcal{X})$ for all $i$:*

$$\theta^{\infty}(C) = \prod_{i=1}^{\infty} \theta(C_i)$$

# Remarks

- In light of de Finetti's theorem, it is very natural to set $\Theta = \mathcal{P}_{\mathcal{X}}$.
- However, we could also use a **different parametric space**: $\hat{\Theta}$, endowed with its Borel sigma-algebra $\mathcal{B}(\hat{\Theta})$.
- Let $\tilde{g} : \hat{\Theta} \to \mathcal{P}_{\mathcal{X}}$ be defined by $\tilde{g}(\hat{\theta}) = P_{\hat{\theta}}$. Once we restrict the codominium to $\Theta = \tilde{g}(\hat{\Theta})$, $g : \hat{\theta} \mapsto P_{\hat{\theta}}$ is **bijective**. We will assume that $g$ is also **bimeasurable**, i.e. measurable and with a measurable inverse.

# Consistency: definition and justification

# Consistency: definition

**Definition**

Let $\hat{\Theta}$ be Polish. The posterior $\hat{\Pi}_n(\cdot|X^{(n)})$ is said to be consistent at $\hat{\theta}_0$ if

$$\hat{\Pi}_n(\cdot|X^{(n)}) \Longrightarrow \delta_{\hat{\theta}_0}$$

either $\theta_0^{\infty}$-almost surely or in $\theta_0^n$-probability.

# Merging of opinions

**Why is consistency relevant** for a subjective Bayesian?

Let $(\alpha_n)_{n\geq 1}$ and $(\beta_n)_{n\geq 1}$ be sequences of probability measures defined on $\hat{\Theta}$.
$\alpha_n$ and $\beta_n$ **merge weakly** if $R(\alpha_n, \beta_n) \to 0$ for all $R : \mathcal{P}_{\hat{\Theta}} \times \mathcal{P}_{\hat{\Theta}} \to \mathbb{R}$ such that

- $R(\alpha, \alpha) = 0$ for all $\alpha \in \mathcal{P}_{\hat{\Theta}}$
- $R$ is continuous.

Let $G_{\hat{\Pi}} = \{(\hat{\theta}_0, x) \in \hat{\Theta} \times \mathcal{X}^\infty : \hat{\Pi}_n(\cdot|X^{(n)} = x^{(n)}) \implies \delta_{\hat{\theta}_0})\} \subset \hat{\Theta} \times \mathcal{X}^\infty$ be the good set where consistency holds.

**Theorem**

*Let $g : \hat{\theta} \mapsto \theta = P_{\hat{\theta}}$ be an omeomorphism. Let $\hat{\Pi}$ be a prior on $\hat{\Theta}$, and let us pick a specific version of the posterior $\hat{\Pi}_n$.*
*The following are equivalent.*

1. $\hat{\Pi}_n$ *is consistent for all* $\hat{\theta}_0 \in \hat{\Theta}$
2. *for all* $\hat{\nu} \in \mathcal{P}_{\hat{\Theta}}$, $\mathbf{P}_{\hat{\nu}}(G_{\hat{\Pi}}) = 1$
3. *for all* $\hat{\nu} \in \mathcal{P}_{\hat{\Theta}}$, $\hat{\Pi}_n$ *and* $\hat{\nu}_n$ *merge weakly* $\mathbf{P}_{\hat{\nu}}$*-almost surely.*
4. *for all* $\hat{\nu} \in \mathcal{P}_{\hat{\Theta}}$, $P_{\hat{\Pi}_n}$ *and* $P_{\hat{\nu}_n}$ *merge* $\mathbf{P}_{\hat{\nu}}$*-almost surely.*

# Consistency: Schwartz

# Background

- Let $\Theta = \{\theta \in \mathcal{P}_{\mathcal{X}} \,:\, \theta \ll \nu\} \subset \mathcal{P}_{\mathcal{X}}$
- Let $\hat{\Theta} = \{[f_\theta] \,:\, \theta \in \Theta\}$, where $[f_\theta]$ is the **class** of Radon-Nikodym derivatives of $\theta$ w.r.t. $\nu$.
- Let $\tau$ be the restriction of the weak topology on $\Theta$. Let $h : \theta \mapsto [f_\theta]$ and $g := h^{-1}$. We **induce** a topology on $\hat{\Theta}$ through $g$ and $\tau$:

$$\hat{\tau} := \{A \subset \hat{\Theta} \,:\, A = g^{-1}(B),\, B \in \tau\}$$

By construction, $g$ is an isomorphism and so **bimeasurable**. Thus, priors defined on $(\Theta, \mathcal{B}(\Theta))$ induce priors on $(\hat{\Theta}, \mathcal{B}(\hat{\Theta}))$ and viceversa.

- For all $\hat{\theta}_0 \in \hat{\Theta}$, for all $\epsilon > 0$ we define a **Kullback-Leibler neighborhood** of $\hat{\theta}_0$ to be:

$$KL(\hat{\theta}_0.\epsilon) := \{\hat{\theta} \in \hat{\Theta} \,:\, KL(\hat{\theta}_0, \hat{\theta}) < \epsilon\}$$

- Then, the **Kullback-Leibler support** of $\hat{\Pi}$ is:

$$KL(\hat{\Pi}) \stackrel{\mathrm{d}}{=} \{\hat{\theta} \in \hat{\Theta} \,:\, \forall\, \epsilon > 0,\ \hat{\Pi}(KL(\hat{\theta}, \epsilon)) > 0\}$$

- Does $KL(\hat{\theta}_0, \epsilon) \in \mathcal{B}(\hat{\Theta})$ for all $\epsilon$ and for all $\theta_0$? We will **assume** so.

# Schwartz Theorem

**Theorem**

*Let* $\hat{\Pi}$ *be a prior on* $\hat{\Theta}$*, let* $\hat{\theta}_0 \in \hat{\Theta}$*. If* $\hat{\theta}_0 \in KL(\hat{\Pi})$ *then* $\hat{\Pi}_n(\cdot|X^{(n)}) \implies \delta_{\hat{\theta}_0}$, $\theta_0^\infty$*-almost surely.*