2022-2023

Bayesian nonparametric statistics

Vincent Rivoirard

Master 2 course

Contents

Chapter 1

Bayesian estimation in the parametric setting

The goal of this chapter is to provide a short introduction to the Bayesian paradigm. For this purpose, we consider the parametric setting.

1.1 Introduction

1.1.1 Recalls on the Bayesian paradigm

We assume that we are given a statistical model denoted $(\Omega, \mathcal{B}, \mathbb{P}_\theta, \theta \in \mathcal{T})$. We recall that if there exists $d \in \mathbb{N}^*$ such that $\mathcal{T} \subset \mathbb{R}^d$ then we speak about parametric statistics. Otherwise, the framework is nonparametric statistics.

In the Bayesian setting, we endow $\mathcal{T}$ with a σ-algebra $\mathcal{A}$ and with a probability distribution Π. This distribution is called **the prior distribution** on the parameter θ which is viewed as the realization of a variable Θ. Briefly speaking, the Bayesian setting offers several advantages.
- The first one is philosophical: As the observation, θ is random, whereas for frequentists, θ is just assumed to be unknown.
- It allows the statistician to incorporate an a priori information on the parameter θ to be estimated.
- it allows the statistician to enhance modeling.

In this course, we will often use the classical Bayesian abuse by not distinguishing between the notation of a random variable and that of its realization.

In the Bayesian setting, $\mathbb{P}_\theta$ is the distribution of a random variable X given θ. Given a prior distribution Π on θ, we are interested in the posterior distribution of θ given X.

For instance, in the discrete setting (i.e. Ω is a discrete set),

$$\mathbb{P}_\theta(x) := P(x|\theta) = P(X = x|\Theta = \theta),$$

and, if $\mathcal{T}$ is also discrete, we can compute $\Pi(\theta|x)$ with the following formula:

$$\Pi(\theta|x) = \frac{\mathbb{P}_\theta(x)\Pi(\theta)}{P(x)} = \frac{\mathbb{P}_\theta(x)\Pi(\theta)}{\sum_{\theta\in\mathcal{T}} \mathbb{P}_\theta(x)\Pi(\theta)}.$$

It is said that $\Pi(\cdot|x)$ is the posterior distribution of θ conditionally on $X = x$.

In the general setting, we consider the following measured space $(\Omega \times \mathcal{T}, \sigma(\mathcal{B} \times \mathcal{A}))$, where $\sigma(\mathcal{B} \times \mathcal{A})$ is the σ-algebra generated by all sets $B \times A$, with $B \in \mathcal{B}$ and $A \in \mathcal{A}$. We assume we are given:

- Π a probability measure on $(\mathcal{T}, \mathcal{A})$, the **prior distribution**, and Θ a random variable such that $\Theta \sim \Pi$.
- $(\mathbb{P}_\theta)_{\theta\in\mathcal{T}}$ a set of probability measures on $(\Omega, \mathcal{B})$, where for any $\theta \in \mathcal{T}$, $\mathbb{P}_\theta$ is the distribution of a random variable of X conditionally on $\Theta = \theta$.

We denote:

- $\mathbb{P}$ the **joint distribution of** (X, Θ):

$$\mathbb{P}(B \times A) := \int_A \mathbb{P}_\theta(B)\, \mathrm{d}\Pi(\theta), \quad \forall B \in \mathcal{B}, \ \ \forall A \in \mathcal{A}$$

- m the **marginal distribution of** X:

$$m(B) := \int_{\mathcal{T}} \mathbb{P}_\theta(B)\, \mathrm{d}\Pi(\theta) = \mathbb{P}(B \times \mathcal{T}), \quad \forall B \in \mathcal{B}$$

Definition 1.1. *We say that*

$$\Pi(\cdot|\cdot) : \mathcal{A} \times \Omega \longmapsto [0; 1]$$

is a posterior distribution given X if and only if

1. *for any $x \in \Omega$, $\Pi(\cdot|x)$ is a probability distribution on $(\mathcal{T}, \mathcal{A})$,*
2. *for any $A \in \mathcal{A}$, $\Pi(A|\cdot)$ is measurable on $(\Omega, \mathcal{B})$,*
3. *for any $A \in \mathcal{A}$, for any $B \in \mathcal{B}$,*

$$\mathbb{P}(B \times A) = \int_B \Pi(A|x)\, \mathrm{d}m(x).$$

Using an old result established by Dudley (1989), we can prove that "the" posterior distribution always exists:

Theorem 1.1. *If $\mathcal{T}$ and Ω are both complete and separable, (a version of) a posterior distribution always exists.*

Remark 1.1. *We often use the notation $\Pi(\cdot|X)$.*

When the $\mathbb{P}_\theta$'s are all dominated by a $\sigma-$finite measure μ, we can set:

$$f_\theta(x) = \frac{\mathrm{d}\mathbb{P}_\theta(x)}{\mathrm{d}\mu(x)}, \quad \forall x \in \Omega.$$

Then,

$$\Pi(A|X) = \frac{\int_A f_\theta(X)\,\mathrm{d}\Pi(\theta)}{\int_{\mathcal{T}} f_\theta(X)\,\mathrm{d}\Pi(\theta)}, \quad \forall A \in \mathcal{A}. \tag{1.1}$$

Indeed, for any $A \in \mathcal{A}$, for any $B \in \mathcal{B}$,

$$\begin{aligned}\mathbb{P}(B \times A) &= \int_A \mathbb{P}_\theta(B)\,\mathrm{d}\Pi(\theta)\\ &= \int_A \int_B f_\theta(x)\,\mathrm{d}\mu(x)\,\mathrm{d}\Pi(\theta)\\ &= \int_B \left[\int_A f_\theta(x)\,\mathrm{d}\Pi(\theta)\right]\mathrm{d}\mu(x).\end{aligned}$$

This yields

$$m(B) = \int_B \left[\int_{\mathcal{T}} f_\theta(x)\,\mathrm{d}\Pi(\theta)\right]\mathrm{d}\mu(x), \quad \forall B \in \mathcal{B}$$

and

$$\mathrm{d}\mu(x) = \frac{\mathrm{d}m(x)}{\int_{\mathcal{T}} f_\theta(x)\,\mathrm{d}\Pi(\theta)}$$

and

$$\mathbb{P}(B \times A) = \frac{\int_B \left[\int_A f_\theta(x)\,\mathrm{d}\Pi(\theta)\right]\mathrm{d}m(x)}{\int_{\mathcal{T}} f_\theta(x)\,\mathrm{d}\Pi(\theta)},$$

providing the announced result (1.1). Observe that

$$\frac{\mathrm{d}\Pi(\theta|X)}{\mathrm{d}\Pi(\theta)} = \frac{f_\theta(X)}{\int_{\mathcal{T}} f_\theta(X)\,\mathrm{d}\Pi(\theta)} \quad \forall \theta \in \mathcal{T}.$$

Furthermore, if Π has a density p with respect to a measure λ, then $\Pi(\cdot|X)$ has also a density with respect to λ which is:

$$p(\theta|X) = \frac{f_\theta(X)p(\theta)}{\int_{\mathcal{T}} f_\theta(X)\,\mathrm{d}\Pi(\theta)}.$$

Example 1.1. *We observe an n-sample $X = (X_1, \ldots, X_n)$ whose distribution is the Bernoulli distribution with parameter $\theta \in]0; 1[$. In this case, X has a density with respect to the counting measure on $\{0, 1\}^n$ and*

$$f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} \Big[\theta^{x_i}(1-\theta)^{1-x_i}\Big], \quad \forall (x_1, \ldots, x_n) \in \{0, 1\}^n.$$

We assume that Π is the uniform distribution on $]0; 1[$, so

$$p(\theta) := \frac{\mathrm{d}\Pi(\theta)}{\mathrm{d}\theta} = 1_{]0;1[}(\theta)$$

and

$$p(\theta|X) = \frac{\theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i} 1_{]0;1[}(\theta)}{\int_0^1 \theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}\,\mathrm{d}\theta},$$

meaning that

$$\Theta|X \sim Beta\Big(\sum_{i=1}^n X_i + 1, n + 1 - \sum_{i=1}^n X_i\Big).$$

Remark 1.2. *We recall that for $\alpha > 0$ and $\beta > 0$, the density of the $Beta(\alpha, \beta)$-distribution is*

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} 1_{]0;1[}(x),$$

with for $\alpha > 0$,

$$\Gamma(\alpha) = \int_0^{+\infty} e^{-t} t^{\alpha-1}\,\mathrm{d}t.$$

If $Y \sim Beta(\alpha, \beta)$, we have

$$\mathbb{E}[Y] = \frac{\alpha}{\alpha+\beta}, \quad var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Exercice 1.1. *We observe an n-sample $X = (X_1, \ldots, X_n)$ whose distribution is the Gaussian distribution $\mathcal{N}(\theta, \sigma^2)$ and we take $\Theta \sim \mathcal{N}(\mu, s^2)$. In this case,*

$$\Theta|X \sim \mathcal{N}\left(\frac{s^2\overline{X}_n + \frac{\sigma^2}{n}\mu}{s^2 + \frac{\sigma^2}{n}}, \frac{s^2\frac{\sigma^2}{n}}{s^2 + \frac{\sigma^2}{n}}\right).$$

We can again enhance the modeling by considering hierarchical models:

$$x \sim f_\theta, \quad \theta \sim p_\mu, \quad \mu \sim g,$$

where f_θ, p_μ and g are densities (and we don't use capital letters). We have (for instance):

- the density of $(x, \theta|\mu)$ is proportional to $f_\theta(x)p_\mu(\theta)$
- the density of (x, θ, μ) is proportional to $f_\theta(x)p_\mu(\theta)g(\mu)$
- the density of $(\mu|x, \theta)$ is $\dfrac{f_\theta(x)p_\mu(\theta)g(\mu)}{\int f_\theta(x)p_\mu(\theta)g(\mu)\,\mathrm{d}\mu}$
- the density of $(\mu|x)$ is $\dfrac{\int f_\theta(x)p_\mu(\theta)g(\mu)\,\mathrm{d}\theta}{\iint f_\theta(x)p_\mu(\theta)g(\mu)\,\mathrm{d}\mu\,\mathrm{d}\theta}$

1.1.2 Bayesian risk and Bayesian rules

To introduce the notion of risk, we need to define a loss function which measures the cost of an estimation error. The aim of the loss function is to measure the distance between the estimate and the target. It does not depend on the Bayesian setting. For $\theta \in \mathcal{T}$, let us assume that we aim at estimating $g(\theta)$ where $g : \mathcal{T} \longmapsto \mathcal{D}$ is a measurable function. Typically, $\mathcal{D} = g(\mathcal{T})$.

Definition 1.2. *A loss function is a function L such that $L : \mathcal{D} \times \mathcal{D} \longmapsto \mathbb{R}_+$ that is measurable.*

Remark 1.3. *Most of the time, L is symmetric.*

Example 1.2. *Classical examples of loss functions are:*

$$L(\cdot,\cdot) = \|\cdot - \cdot\|_2^2, \quad L(\cdot,\cdot) = \|\cdot - \cdot\|_1, \quad L(\cdot,\cdot) = 1_{\{\cdot \neq \cdot\}}.$$

Definition 1.3. *The* ***frequentist risk*** *of an estimate $T(X)$ of $g(\theta)$ associated with a loss function L is the function*

$$\begin{array}{rccl} R(\cdot, T): & \mathcal{T} & \longrightarrow & \overline{\mathbb{R}}_+ \\ & \theta & \longmapsto & \mathbb{E}_\theta[L(T(X), g(\theta))]. \end{array}$$

In particular, for any $\theta \in \mathcal{T}$,

$$R(\theta, T) = \int_\Omega L(T(x), g(\theta))\,\mathrm{d}\mathbb{P}_\theta(x).$$

In the Bayesian setting, once we are given a prior distribution Π, we can integrate the risk R with respect to Π. It is natural since we do not assume that there is a true value for the unknown parameter θ. We obtain the Bayesian risk.

Definition 1.4. *The Bayesian risk (with respect to Π and T) is*

$$r(\Pi, T) = \int_{\mathcal{T}} R(\theta, T)\,\mathrm{d}\Pi(\theta).$$

Remark 1.4. *We have*

$$\begin{aligned} r(\Pi,T) &= \int_{\mathcal{T}} R(\theta,T)\,\mathrm{d}\Pi(\theta) \\ &= \int_{\mathcal{T}}\int_{\Omega} L(T(x),g(\theta))\,\mathrm{d}\mathbb{P}_\theta(x)\,\mathrm{d}\Pi(\theta) \\ &= \int_{\mathcal{T}}\int_{\Omega} L(T(x),g(\theta))\,\mathrm{d}\Pi(\theta|x)\,\mathrm{d}m(x) \\ &= \int_{\Omega} \mathrm{d}m(x) \int_{\mathcal{T}} L(T(x),g(\theta))\,\mathrm{d}\Pi(\theta|x). \end{aligned}$$

Using $r(\Pi,T)$, we can define what is a good estimate given Π. It will be called a Bayesian (or Bayes) estimate.

Definition 1.5. *A Bayesian estimate associated with a prior distribution Π is an estimate that minimizes the function $T \longmapsto r(\Pi,T)$ equivalently, the function $T(x) \longmapsto \int_{\mathcal{T}} L(T(x),g(\theta))\,\mathrm{d}\Pi(\theta|x)$.*

The Bayesian rule can be characterized for some specific loss functions.

Theorem 1.2. *We assume that $g(\theta)\in\mathbb{R}$.*

1. *If $\mathbb{E}[g^2(\theta)|X]<\infty$ and $L(T(X),g(\theta)) = (T(X)-g(\theta))^2$, then the Bayesian estimate is the posterior mean $\mathbb{E}[g(\theta)|X]$.*

2. *If $\mathbb{E}[|g(\theta)||X]<\infty$ and $L(T(X),g(\theta)) = |T(X)-g(\theta)|$, then the Bayesian estimate is any $a\in\mathbb{R}$ such that*

$$\Pi(g(\theta)\leq a|X) \geq \frac{1}{2}, \quad \Pi(g(\theta)\geq a|X)\geq \frac{1}{2}. \tag{1.2}$$

Proof. We have to minimize the function

$$u \longmapsto G(u) := \int_{\mathcal{T}} L(u,g(\theta))\,\mathrm{d}\Pi(\theta|X).$$

1. For the first case, we have

$$\begin{aligned} G(u) &= \int_{\mathcal{T}} (u-g(\theta))^2\,\mathrm{d}\Pi(\theta|X) \\ &= u^2 - 2u\mathbb{E}[g(\theta)|X] + \mathbb{E}[g^2(\theta)|X], \end{aligned}$$

which is minimum for $u=\mathbb{E}[g(\theta)|X]$.

2. For the second case, we have

$$G(u)=\int_{\mathcal{T}}|u-g(\theta)|\,\mathrm{d}\Pi(\theta|X)=:\mathbb{E}_{\theta|X}[|u-g(\theta)|].$$

We take $a\in\mathbb{R}$ so that (1.2) is satisfied and we show that, given $c\in\mathbb{R}$, $G(a)\leq G(c)$. We first assume that $c>a$. We now have that $(a+c)/2\in(a;c)$ and

$$\begin{array}{rcl} |g(\theta)-c|= & |g(\theta)-a|+(c-a) & \text{on } \{g(\theta)\leq a\}\\ |g(\theta)-c|\geq & |g(\theta)-a| & \text{on } \{a<g(\theta)<(a+c)/2\}\\ |g(\theta)-c|\geq & |g(\theta)-a|-(c-a) & \text{on } \{g(\theta)\geq (a+c)/2\}. \end{array}$$

Consequently,

$$|g(\theta)-c|\geq|g(\theta)-a|+(c-a)1_{\{g(\theta)\leq a\}}-(c-a)1_{\{g(\theta)\geq(a+c)/2\}}.$$

Therefore,

$$\mathbb{E}_{\theta|X}[|g(\theta)-c|]\geq\mathbb{E}_{\theta|X}[|g(\theta)-a|]+(c-a)\big(\Pi(g(\theta)\leq a|X)-\Pi(g(\theta)\geq(a+c)/2|X)\big).$$

But

$$\Pi(g(\theta)\geq(a+c)/2|X)\leq\Pi(g(\theta)>a|X)=1-\Pi(g(\theta)\leq a|X),$$

which implies that

$$\mathbb{E}_{\theta|X}[|g(\theta)-c|]\geq\mathbb{E}_{\theta|X}[|g(\theta)-a|]+(c-a)\big(2\Pi(g(\theta)\leq a|X)-1\big)\geq\mathbb{E}_{\theta|X}[|g(\theta)-a|]$$

and then $G(a)\leq G(c)$. The case $c<a$ is similar except that we use $\Pi(g(\theta)\geq a|X)\geq\frac{1}{2}\geq 0$.

□

Remark 1.5. *We recall that if $F_{g,X}$ is the c.d.f of $g(\theta)$ under the posterior distribution $\Pi(\cdot|X)$, then, setting for any $t\in(0;1)$,*

$$F_{g,X}^{(-1)}(t):=\inf\{\theta:F_{g,X}(\theta)\geq t\},$$

$q_{g,X,0.5}:=F_{g,X}^{(-1)}\big(\frac{1}{2}\big)$ *satisfies*

$$\Pi(g(\theta)\leq q_{g,X,0.5}|X)\geq\frac{1}{2},\quad\Pi(g(\theta)\geq q_{g,X,0.5}|X)\geq\frac{1}{2}.$$

Indeed, since $F_{g,X}$ is right-continuous, we have that for any $t\in(0;1)$, $F_{g,X}(F_{g,X}^{(-1)}(t))\geq t$. Then,

$$\Pi(g(\theta)\leq q_{g,X,0.5}|X)=F_{g,X}(q_{g,X,0.5})=F_{g,X}\Big(F_{g,X}^{(-1)}\Big(\frac{1}{2}\Big)\Big)\geq\frac{1}{2}.$$

And

$$\begin{aligned}1-\Pi(g(\theta)\geq q_{g,X,0.5}|X)&=\Pi(g(\theta)<q_{g,X,0.5}|X)\\&=\lim_{n\to+\infty}\Pi(g(\theta)\leq q_{g,X,0.5}-n^{-1}|X)\\&\leq\frac{1}{2}.\end{aligned}$$

This yields $\Pi(g(\theta)\geq q_{g,X,0.5}|X)\geq\frac{1}{2}$.

Theorem 1.2 and Remark 1.5 introduce two classical Bayes rules:
- the **posterior mean** (i.e. the mean of the posterior distribution): $\mathbb{E}[g(\theta)|X]$
- the **posterior median** (i.e. the median of the posterior distribution): $q_{g,X,0.5}$.
We can also consider the **posterior mode** (i.e. the mode of the posterior distribution). When $g(\theta)=\theta$ it is defined by

$$\hat{\theta}_{\text{mode}}\in\arg\max_{\theta\in\mathcal{T}}\Pi(\theta|X).$$

Example 1.3. *We consider the setting of Example 1.1 and we estimate* $\theta\in\mathcal{T}=(0;1)$. *The maximum likelihood estimate is*

$$\hat{\theta}_{emv}=\frac{1}{n}\sum_{i=1}^{n}X_i.$$

With $p(\theta)=1_{(0;1)}(\theta)$, *the posterior mode is* $\hat{\theta}_{emv}$, *the posterior mean for* θ *is*

$$\mathbb{E}[\theta|X^n]=\frac{1+\sum_{i=1}^{n}X_i}{n+2}$$

and note that

$$\hat{\theta}_{emv}\leq\mathbb{E}[\theta|X^n]\iff\frac{1}{n}\sum_{i=1}^{n}X_i\leq\frac{1}{2}.$$

Observe that with the improper prior Π *such that its density with respect to the Lebesgue measure is proportional to the function* $\theta\longmapsto\theta^{-1}(1-\theta)^{-1}$ *then,*

$$\mathbb{E}[\theta|X^n]=\hat{\theta}_{emv}.$$

Finally, if $\theta\sim Beta(\alpha,\beta)$,

$$\mathbb{E}[\theta|X^n]=\frac{\alpha+\sum_{i=1}^{n}X_i}{\alpha+\beta+n}.$$

1.1.3 Admissible estimates

Definition 1.6. *An estimate $T(X)$ is* ***non-admissible*** *is there exists another estimate $T'(X)$ such that*

$$R(\theta, T) \geq R(\theta, T'), \quad \forall \theta \in \mathcal{T}$$

and there exists $\theta \in \mathcal{T}$ such that

$$R(\theta, T) > R(\theta, T').$$

An estimate is ***admissible*** *if it is not non-admissible.*

Proposition 1.1. *If Π is absolutely continuous with respect to the Lebesgue measure and if the density p of Π is positive on $\mathcal{T}$ and $\theta \longmapsto R(\theta, T)$ is continuous for any estimate $T(X)$, then the Bayes rule is admissible.*

Proof. We denote $\delta_\Pi(X)$ the Bayes rule. If $\delta_\Pi(X)$ is non-admissible then there exists an estimate $T(X)$ such that $R(\theta, T) \leq R(\theta, \delta_\Pi)$ for any $\theta \in \mathcal{T}$ and $R(\theta, T) < R(\theta, \delta_\Pi)$ for any $\theta \in C$, where C is an open set of $\mathcal{T}$ with positive Lebesgue measure ($\theta \longmapsto R(\theta, T)$ and $\theta \longmapsto R(\theta, \delta_\Pi)$ are continuous). Then

$$\begin{aligned} r(\Pi, T) = \mathbb{E}_\Pi[R(\theta, T)] &= \int_{\mathcal{T}} R(\theta, T) p(\theta) \, \mathrm{d}\theta \\ &< \int_{\mathcal{T}} R(\theta, \delta_\Pi) p(\theta) \, \mathrm{d}\theta = \mathbb{E}_\Pi[R(\theta, \delta_\Pi)] = r(\Pi, \delta_\Pi), \end{aligned}$$

which cannot occur. So, $\delta_\Pi(X)$ is admissible. □

1.2 Consistency in the parametric setting

We keep the same notations. We consider the asymptotic setting, meaning that we assume that the number of observations goes to $+\infty$. In the sequel, we assume that we are given $X^n = (X_1, \ldots, X_n)$ where the X_i's are i.i.d. We denote $\mathbb{P}_\theta^n$ the distribution of X^n and $\mathbb{P}_\theta^\infty$ the distribution of $X^\infty = (X_1, \ldots, X_n, \ldots)$. Before introducing the definition of consistency, let us give some elementary recalls on convergence.

1.2.1 Recalls on convergence

Let S a metric space and $\mathcal{B}(S)$ the Borelian σ-algebra on S. We denote $C(S)$ the set of all $\mathbb{R}$-valued bounded continuous functions on S.

Definition 1.7. *A sequence $(P_n)_n$ of probability measures on S is said to converge weakly to a probability measure P, written as $\mathbb{P}_n \overset{n \to +\infty}{\rightsquigarrow} P$ if*

$$\int f \, \mathrm{d}P_n \overset{n \to +\infty}{\longrightarrow} \int f \, \mathrm{d}P, \quad \forall f \in C(S).$$

Theorem 1.3 (Portmanteau theorem)**.** *The following facts are equivalent.*

1. $P_n \overset{n\to+\infty}{\rightsquigarrow} P$
2. $\int f \,\mathrm{d}P_n \overset{n\to+\infty}{\longrightarrow} \int f \,\mathrm{d}P, \quad \forall f \in C(S)$ *uniformly continuous.*
3. $\limsup_{n\to+\infty} P_n(F) \leq P(F), \quad \forall F$ *closed.*
4. $\liminf_{n\to+\infty} P_n(U) \geq P(U), \quad \forall U$ *open.*
5. $\lim_{n\to+\infty} P_n(B) = P(B), \quad \forall B$ *such that* $P(\delta B) = 0$.

Theorem 1.4 (Prohorov's theorem)**.** *If S is a complete separable metric space, then every subsequence of $(P_n)_n$ has a weakly convergent subsequence if and only if $(P_n)_n$ is tight (i.e. $\forall\varepsilon > 0$ $\exists K_\varepsilon$ a compact set such that $\forall n$, $P_n(K_\varepsilon) \geq 1-\varepsilon$).*

Remark 1.6. *Theorem 2.3 of Billingsley (1995) states that $P_n \overset{n\to+\infty}{\rightsquigarrow} P$ if and only if every subsequence $(P_{n'})_{n'}$ of $(P_n)_n$ as a subsequence $(P_{n''})_{n''}$ such that $P_{n''} \overset{n''\to+\infty}{\rightsquigarrow} P$. So, the weak convergence on a complete separable metric space is proved by using tightness and the unicity of the limit.*

Remark 1.7. *Remember that in a complete metric space, a subset is a relatively compact set if and only if it is a precompact set (i.e. its closure is compact). Therefore, Prohorov's theorem is equivalent to the following statement: Let $\Gamma \subset S$; then, Γ is precompact if and only if Γ is tight.*

We refer the reader to Appendix A of Ghosal and van der Vaart (2017) and Billingsley (1995, 1999) for further details.

1.2.2 Consistency properties

Consistency is one of the most elementary asymptotic properties that can be satisfied by a sequence of posterior distributions. It expresses the fact that if θ_0 is the true value of the parameter, the posterior learns more and more from the data and puts more and more mass close to 0.

Definition 1.8. *For each n, let $\Pi(\cdot|X^n)$ be a posteriori distribution given X^n. The sequence $(\Pi(\cdot|X^n))_n$ is said to be consistent at $\theta_0 \in \mathcal{T}$ if there exists $\Omega_0 \in \mathcal{B}$ with $\mathbb{P}^\infty_{\theta_0}(\Omega_0) = 1$ such that if $\omega \in \Omega_0$ then for each open set U containing θ_0, we have*

$$\Pi(U|X^n(w)) \overset{n\to+\infty}{\longrightarrow} 1.$$

Most of the time, consistency can be characterized by using the following result.

Theorem 1.5. *If $\mathcal{T}$ is a separable metric space, then*

$$(\Pi(\cdot|X^n))_n \textbf{ is consistent at } \theta_0 \in \mathcal{T} \iff \Pi(\cdot|X^n) \overset{n\to+\infty}{\rightsquigarrow} \delta_{\theta_0}\ \mathbb{P}_{\theta_0} - a.e.$$

Proof. We denote d the distance that makes $\mathcal{T}$ metrizable. We denote $B_d(\theta_0, r)$ the ball of center θ_0 with radius $r > 0$ (for the distance d). We first prove the following lemma.

Lemma 1.1. *Assume that $\mathcal{T}$ is a metric space. Then,*

$$(\Pi(\cdot|X^n))_n \textbf{ is consistent at } \theta_0 \in \mathcal{T} \iff \forall p \in \mathbb{N}^*\ \exists \Omega_p \in \mathcal{B} \text{ with } \mathbb{P}_{\theta_0}^{\infty}(\Omega_p) = 1$$

$$\text{such that if } \omega \in \Omega_p\ \Pi(B_d(\theta_0, p^{-1})|X^n(\omega)) \to 1.$$

Proof of the lemma:

- $\Rightarrow$: Let $p \in \mathbb{N}^*$. We take $\Omega_p = \Omega_0$ and $U = B_d(\theta_0, p^{-1})$.
- $\Leftarrow$: We take $\Omega_0 = \cap_{p\in\mathbb{N}^*}\Omega_p$, so that $\mathbb{P}_{\theta_0}^{\infty}(\Omega_0) = 1$. Furthermore, for each open set U containing θ_0, there exists $p \in \mathbb{N}^*$ such that $B_d(\theta_0, p^{-1}) \subset U$ and if we take $\omega \in \Omega_0$ we have that $\omega \in \Omega_p$ and

$$\Pi(U|X^n(w)) \geq \Pi(B_d(\theta_0, p^{-1})|X^n(\omega)) \to 1.$$

In particular, the previous lemma shows that

$$\begin{aligned} & (\Pi(\cdot|X^n))_n \text{ is consistent at } \theta_0 \in \mathcal{T} \\ \iff & \text{for any open set } U \text{ containing } \theta_0,\ \Pi(U|X^n) \overset{n\to+\infty}{\longrightarrow} 1\ \mathbb{P}_{\theta_0} - a.e. \\ \iff & \text{for any open set } U\ \liminf_{n\to+\infty} \Pi(U|X^n) \geq \delta_{\theta_0}(U)\ \mathbb{P}_{\theta_0} - a.e. \\ \iff & \Pi(\cdot|X^n) \overset{n\to+\infty}{\rightsquigarrow} \delta_{\theta_0}\ \mathbb{P}_{\theta_0} - a.e. \end{aligned}$$

For the last equivalence, we use the Portmanteau Theorem 1.3 applied with $P_n = \Pi(\cdot|X^n)$. □

As expressed by the previous result, consistency means that if θ_0 is the true value of the parameter, the posterior concentrates around θ_0. This reconciliates Bayesian and frequentist approaches.

Example 1.4. *We consider $X^n = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} \mathbb{P}_\theta = Ber(\theta)$ and $\theta \sim Beta(\alpha, \beta)$, with $\alpha > 0$ and $\beta > 0$. Then, for all i, $\mathbb{P}_\theta(X_i = 1) = \theta$ and $\mathbb{P}_\theta(X_i = 0) = 1 - \theta$ and the density of Π is*

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}1_{(0;1)}(\theta).$$

Then,

$$p(\theta|X^n) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+r)\Gamma(\beta+n+r)}\theta^{\alpha+r-1}(1-\theta)^{\beta+n-r-1}1_{(0;1)}(\theta),$$

where $r = \sum_{i=1}^n 1_{\{X_i=1\}}$*. Then, computing the expectation and the variance of* θ *under* $\Pi(\cdot|X^n)$*, we obtain*

$$\mathbb{E}[\theta|X^n] = \frac{\alpha+r}{\alpha+\beta+n} = \frac{\alpha}{\alpha+\beta} \times \frac{\alpha+\beta}{\alpha+\beta+n} + \frac{n}{\alpha+\beta+n} \times \frac{r}{n} \overset{n\to+\infty}{\longrightarrow} \theta_0 \ \mathbb{P}_{\theta_0} - a.e.$$

and

$$\begin{aligned} var(\theta|X^n) &= \mathbb{E}[\theta^2|X^n] - (\mathbb{E}[\theta|X^n])^2 \\ &= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+r)\Gamma(\beta+n+r)} \times \frac{\Gamma(\alpha+r+2)\Gamma(\beta+n+r)}{\Gamma(\alpha+\beta+2+n)} - \left(\frac{\alpha+r}{\alpha+\beta+n}\right)^2 \\ &= \frac{(\alpha+r)(\beta+n-r)}{(\alpha+\beta+n+1)(\alpha+\beta+n)^2} \end{aligned}$$

and

$$var(\theta|X^n) \overset{n\to+\infty}{\longrightarrow} 0 \ \mathbb{P}_{\theta_0} - a.e.$$

Now, for any $\varepsilon > 0$*,*

$$\begin{aligned} \Pi(|\theta-\theta_0| \geq \varepsilon|X^n) &\leq \Pi\Big(|\theta - \mathbb{E}[\theta|X^n]| \geq \frac{\varepsilon}{2}|X^n\Big) + \Pi\Big(|\mathbb{E}[\theta|X^n] - \theta_0| \geq \frac{\varepsilon}{2}|X^n\Big) \\ &\leq \frac{4}{\varepsilon^2} var(\theta|X^n) + \mathbb{P}\Big(|\mathbb{E}[\theta|X^n] - \theta_0| \geq \frac{\varepsilon}{2}|X^n\Big). \end{aligned}$$

Observe that $|\mathbb{E}[\theta|X^n] - \theta_0|$ *is not random under* $\Pi(\cdot|X^n)$ *and for n large enough, the second term of the right hand side is equal to 0. Therefore, with* $U = \{\theta : \ |\theta-\theta_0| < \varepsilon\}$*,*

$$\lim_{n\to+\infty} \Pi(U|X^n) = 1 \ \mathbb{P}_{\theta_0} - a.e.$$

We now state Doob's theorem whose proof is based on martingale convergence theorems and sophisticated arguments of measure theory.

Theorem 1.6 (Doob's theorem)**.** *Suppose that* Ω *and* $\mathcal{T}$ *are both complete separable spaces and let us assume that* $\theta \to \mathbb{P}_\theta$ *is injective. Let* Π *a prior. Then, there exists* $\mathcal{T}_0 \in \mathcal{A}$ *with* $\Pi(\mathcal{T}_0) = 1$ *such that* $\Pi(\cdot|X^n)$ *is consistent at any* $\theta_0 \in \mathcal{T}_0$*.*

Doob's theorem is a very nice result but its proof is not constructive. So, we have no idea about $\mathcal{T}_0$. Consequently, given $\theta_0 \in \mathcal{T}$, we do not know if consistency holds at θ_0. See for instance Freedman (1963) for counter-examples.

The next theorem shows that posterior consistency is connected to posterior robustness.

Theorem 1.7. *Assume that the family $\{\mathbb{P}_\theta\}_{\theta\in\mathcal{T}}$ is dominated by a σ-finite measure μ. Let $f_\theta = \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}$ for any $\theta \in \mathcal{T}$. Let θ_0 belonging to the interior of $\mathcal{T}$ and let p_1 and p_2 two prior densities with respect to a σ-finite measure ν such that p_1 and p_2 are continuous and positive at θ_0. If the posterior densities $p_1(\cdot|X^n)$ and $p_2(\cdot|X^n)$ are both consistent at θ_0, then*

$$\lim_{n\to+\infty}\int_{\mathcal{T}} |p_1(\theta|X^n) - p_2(\theta|X^n)|\,\mathrm{d}\nu(\theta) = 0 \ \mathbb{P}_{\theta_0} - a.e.$$

Proof. We want to show that with $\mathbb{P}_{\theta_0}-$probability 1,

$$\lim_{n\to+\infty}\int_{\mathcal{T}} \left|p_2(\theta|X^n)\left(1 - \frac{p_1(\theta|X^n)}{p_2(\theta|X^n)}\right)\right|\mathrm{d}\nu(\theta) = 0 \ \mathbb{P}_{\theta_0} - a.e.$$

Let $\delta > 0$ fixed later. Since p_1 and p_2 are positive and continuous at θ_0 there exists an open set $U \subset \mathcal{T}$ such that for all $\theta \in U$,

$$\left|\frac{p_1(\theta)}{p_2(\theta)} - \frac{p_1(\theta_0)}{p_2(\theta_0)}\right| < \delta, \quad |p_1(\theta) - p_1(\theta_0)| < \delta, \quad |p_2(\theta) - p_2(\theta_0)| < \delta.$$

There exists Ω_0 with $\mathbb{P}_{\theta_0}^\infty(\Omega_0) = 1$ such that $\forall\omega \in \Omega_0$, $\forall j \in \{1,2\}$

$$\Pi_j(U|X^n(\omega)) := \frac{\int_U \prod_{i=1}^n f_\theta(X_i(\omega))p_j(\theta)\,\mathrm{d}\nu(\theta)}{\int_{\mathcal{T}} \prod_{i=1}^n f_\theta(X_i(\omega))p_j(\theta)\,\mathrm{d}\nu(\theta)} \overset{n\to+\infty}{\longrightarrow} 1.$$

Let $\eta > 0$ and $\omega \in \Omega_0$ be fixed. Then, there exists n_0 such that $\forall n \geq n_0$, $\forall j \in \{1,2\}$

$$\Pi_j(U|X^n(\omega)) \geq 1 - \eta. \tag{1.3}$$

We fix $\theta \in U$. Since $\forall j \in \{1,2\}$,

$$p_j(\theta|X^n(\omega)) = \frac{p_j(\theta)\prod_{i=1}^n f_\theta(X_i(\omega))}{\int_{\mathcal{T}} p_j(\theta)\prod_{i=1}^n f_\theta(X_i(\omega))\,\mathrm{d}\nu(\theta)},$$

then

$$\frac{p_1(\theta|X^n(\omega))}{p_2(\theta|X^n(\omega))} = \frac{p_1(\theta)}{p_2(\theta)}\frac{\int_{\mathcal{T}} p_2(\theta)\prod_{i=1}^n f_\theta(X_i(\omega))\,\mathrm{d}\nu(\theta)}{\int_{\mathcal{T}} p_1(\theta)\prod_{i=1}^n f_\theta(X_i(\omega))\,\mathrm{d}\nu(\theta)}$$

and

$$\frac{p_1(\theta|X^n(\omega))}{p_2(\theta|X^n(\omega))} \leq \left(\frac{p_1(\theta_0)}{p_2(\theta_0)} + \delta\right)(1-\eta)^{-1}\frac{\int_U p_2(\theta)\prod_{i=1}^n f_\theta(X_i(\omega))\,\mathrm{d}\nu(\theta)}{\int_U p_1(\theta)\prod_{i=1}^n f_\theta(X_i(\omega))\,\mathrm{d}\nu(\theta)}$$
$$\frac{p_1(\theta|X^n(\omega))}{p_2(\theta|X^n(\omega))} \geq \left(\frac{p_1(\theta_0)}{p_2(\theta_0)} - \delta\right)(1-\eta)\frac{\int_U p_2(\theta)\prod_{i=1}^n f_\theta(X_i(\omega))\,\mathrm{d}\nu(\theta)}{\int_U p_1(\theta)\prod_{i=1}^n f_\theta(X_i(\omega))\,\mathrm{d}\nu(\theta)}.$$

By the choice of U, $\forall j \in \{1,2\}$

$$(p_j(\theta_0) - \delta)\int_U \prod_{i=1}^n f_\theta(X_i(\omega))\,\mathrm{d}\nu(\theta) \leq \int_U p_j(\theta)\prod_{i=1}^n f_\theta(X_i(\omega))\,\mathrm{d}\nu(\theta)$$

and

$$\int_U p_j(\theta) \prod_{i=1}^n f_\theta(X_i(\omega))\, \mathrm{d}\nu(\theta) \leq (p_j(\theta_0) + \delta) \int_U \prod_{i=1}^n f_\theta(X_i(\omega))\, \mathrm{d}\nu(\theta).$$

We obtain

$$\left(\frac{p_1(\theta_0)}{p_2(\theta_0)} - \delta\right)(1-\eta)\left(\frac{p_2(\theta_0)-\delta}{p_1(\theta_0)+\delta}\right) \leq \frac{p_1(\theta|X^n(\omega))}{p_2(\theta|X^n(\omega))} \leq \left(\frac{p_1(\theta_0)}{p_2(\theta_0)} + \delta\right)(1-\eta)^{-1}\left(\frac{p_2(\theta_0)+\delta}{p_1(\theta_0)-\delta}\right).$$

Now, for any $\varepsilon > 0$ for δ and η small enough, $\forall\theta \in U$,

$$\left|\frac{p_1(\theta|X^n(\omega))}{p_2(\theta|X^n(\omega))} - 1\right| \leq \varepsilon.$$

Finally, for $n \geq n_0$, using (1.3),

$$\begin{aligned}\int_{\mathcal{T}} |p_1(\cdot|X^n) - p_2(\cdot|X^n)|\, \mathrm{d}\nu(\theta) &= \int_U |p_1(\cdot|X^n) - p_2(\cdot|X^n)|\, \mathrm{d}\nu(\theta) \\ &\qquad + \int_{U^c} |p_1(\cdot|X^n) - p_2(\cdot|X^n)|\, \mathrm{d}\nu(\theta) \\ &\leq \varepsilon + 2\eta.\end{aligned}$$

This ends the proof of the theorem. □

In view of previous results, posterior consistency can be viewed as:

1. a frequentist validation of the Bayes approach
2. merging of posteriors arising from different priors
3. an expression of ”data eventually swamps the prior.”

Chapter 2

Prior distributions in the nonparametric setting - Dirichlet processes

In this chapter, we denote $\mathcal{M}(\Omega)$ the set of all probability measures on Ω. We shall focus on two cases: $\mathrm{card}(\Omega) < \infty$ and $\Omega = \mathbb{R}$. Before describing the most classical probability distributions on $\mathcal{M}(\Omega)$, we provide a short study of this space.

2.1 Short study of the space $\mathcal{M}(\Omega)$

In this section, Ω is a complete separable metric space (i.e. a Polish space); $\mathcal{B}$ is the corresponding Borel σ-algebra on Ω. In this case, it was proved (also by Prohorov) that $\mathcal{M}(\Omega)$ is also metrizable, complete and separable under the weak convergence. Such properties are important to characterize the weak convergence through tightness properties (see Prohorov's Theorem 1.4) So, there is a metric ρ on $\mathcal{M}(\Omega)$ such that $\rho(P_n, P)$ if and only if $P_n \overset{n\to+\infty}{\rightsquigarrow} P$. See Appendix A of Ghosal and van der Vaart (2017) or Billingsley (1995, 1999) for further details. The set $\mathcal{M}(\Omega)$ has also other natural metrics.

2.1.1 Metrics on the space $\mathcal{M}(\Omega)$

We describe the most classical metrics on the space $\mathcal{M}(\Omega)$. In the sequel, we consider P and Q two elements of $\mathcal{M}(\Omega)$.

- The **Total Variation metric:**

$$\|P - Q\|_{TV} := \sup_{B\in\mathcal{B}} |P(B) - Q(B)|.$$

- For the case $\Omega = \mathbb{R}$, the **supremum metric:**

$$d_{\sup}(P, Q) := \sup_{t \in \mathbb{R}} |P(-\infty; t] - Q(-\infty; t]|.$$

We have the following proposition.

Proposition 2.1. *The metric space $(\mathcal{M}(\Omega), \| \cdot \|_{TV})$ is not separable if Ω is not countable. Furthermore, $(\mathcal{M}(\mathbb{R}), d_{\sup})$ is complete but not separable.*

Proof. We only prove the first point. We recall that the space $\mathcal{M}(\Omega)$ is separable if and only if it contains a dense countable subspace. To prove that $\mathcal{M}(\Omega)$ is not separable if Ω is not countable, for any $x \in \Omega$, we can characterize the open ball centered at δ_x of radius $\varepsilon > 0$ defined by

$$U(x, \varepsilon) := \{P : \ \|P - \delta_x\|_{TV} < \varepsilon\}.$$

It is easy to observe that

$$U(x, \varepsilon) := \{P : \ P(\{x\}) > 1 - \varepsilon\}.$$

Therefore if $\varepsilon < \frac{1}{2}$, and if $x \neq x'$, $U(x, \varepsilon) \cap U(x', \varepsilon) = \emptyset$. Consequently, if there exists a dense subset $(P_n)_{n \in \mathbb{N}^*}$, for any $x \in \Omega$, there exists $n_x \in \mathbb{N}^*$ such that $P_{n_x} \in U(x, \varepsilon)$ and $P_{n_x} \notin U(x', \varepsilon)$ for $x' \neq x$. Therefore, Ω is countable. □

The previous proposition shows that it seems interesting to restrict to subsets of $\mathcal{M}(\Omega)$. We consider μ a σ-finite measure on Ω and we set

$$\mathcal{L}_\mu := \{P \in \mathcal{M}(\Omega) : \ P \ll \mu\}.$$

In the sequel, we consider $P \in \mathcal{L}_\mu$ and $Q \in \mathcal{L}_\mu$ and we define

$$p := \frac{\mathrm{d}P}{\mathrm{d}\mu}, \quad q := \frac{\mathrm{d}Q}{\mathrm{d}\mu}.$$

We now introduce:

- The $\mathbb{L}_1$**-metric:**

$$\|P - Q\|_1 := \int_\Omega |p(x) - q(x)| \,\mathrm{d}\mu(x).$$

- The **Hellinger metric:**

$$H(P, Q) := \left(\int_\Omega \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 \mathrm{d}\mu(x) \right)^{\frac{1}{2}}.$$

Remark 2.1. *Observe that $\mathbb{L}_1$ and Hellinger metrics do not depend on the reference measure μ.*

The space $\mathcal{L}_\mu$ endowed with the $\mathbb{L}_1$-metric or the Hellinger metric is a Polish space.

- The **Kullback-Leibler divergence:**

$$K(P,Q) := \begin{cases} \int_\Omega \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P & \text{if } P \ll Q \\ +\infty & \text{otherwise} \end{cases}$$

Observe that $KL(P,Q) \neq KL(Q,P)$, so the Kullback-Leibler divergence is not a distance. However, we have that $K(P,Q) \geq 0$ and if $P \ll \mu$ and $Q \ll \mu$, $K(P,Q) = 0 \iff \frac{\mathrm{d}P}{\mathrm{d}\mu} = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$ $\mu-$ a.e. (as proved later).

Exercice 2.1. *If $P = \otimes_{i=1}^n P_i$ and $Q = \otimes_{i=1}^n Q_i$, then*

$$K(P,Q) = \sum_{i=1}^n K(P_i, Q_i).$$

2.1.2 Connections between metrics

Proposition 2.2. *Let P and Q two elements of $\mathcal{M}(\Omega)$. If p and q are respectively the densities of P and Q with respect to a measure μ, then*

$$\|P-Q\|_{TV} = \frac{1}{2}\|P-Q\|_1 = \frac{1}{2}\int_\Omega |p(x)-q(x)|\,\mathrm{d}\mu(x).$$

Proof. We denote for any $y \in \mathbb{R}$, $y_+ = \max(y,0)$ and $y_- = \max(-y,0)$. So $y = y_+ - y_-$ and $|y| = y_+ + y_-$. Let $B := \{x \in \Omega :\ p(x) \geq q(x)\} \in \mathcal{B}$. So, for any $A \in \mathcal{B}$,

$$\begin{aligned} P(A) - Q(A) &= \int_A (p(x)-q(x))\,\mathrm{d}\mu(x) \\ &\leq \int_{A\cap B} (p(x)-q(x))\,\mathrm{d}\mu(x) \\ &\leq \int_B (p(x)-q(x))\,\mathrm{d}\mu(x) \\ &= \int_\Omega (p(x)-q(x))_+\,\mathrm{d}\mu(x) \\ &= \frac{1}{2}\int_\Omega |p(x)-q(x)|\,\mathrm{d}\mu(x). \end{aligned}$$

We have used that $\int_\Omega (p(x)-q(x))\,\mathrm{d}\mu(x) = 1-1 = 0$ so,

$$\int_\Omega (p(x)-q(x))_+\,\mathrm{d}\mu(x) = \int_\Omega (p(x)-q(x))_-\,\mathrm{d}\mu(x) = \frac{1}{2}\int_\Omega |p(x)-q(x)|\,\mathrm{d}\mu(x).$$

Taking the supremum with respect to $A \in \mathcal{B}$, we obtain

$$\|P - Q\|_{TV} \leq \frac{1}{2} \int_{\Omega} |p(x) - q(x)| \,\mathrm{d}\mu(x).$$

The equality is obtained with $A = B$. □

Proposition 2.3. *Let P and Q two elements of $\mathcal{L}_\mu$ for μ a measure on Ω. Then,*

$$H^2(P, Q) \leq K(P, Q)$$

and

$$\frac{1}{4}\|P - Q\|_1^2 \leq H^2(P, Q) \leq \|P - Q\|_1.$$

Proof. We prove the first point. We only have to consider the case $P \ll Q$. We denote p and q the densities of P and Q with respect to μ such that $P \ll \mu$ and $Q \ll \mu$. Since $Q(\{q = 0\}) = 0$, then $P(\{q = 0\}) = 0$ and $\int p(x) 1_{\{q(x)=0\}} \,\mathrm{d}\mu(x) = 0$. This implies that

$$p(x)1_{\{q(x)=0\}} = 0 \ \mu - a.e.$$

And except on a μ-negligible set, if $q(x) = 0$ then $p(x) = 0$ or if $p(x) > 0$ then $q(x) > 0$. Hence, by using $-\log(y + 1) \geq -y$ for $y > -1$, we have:

$$\begin{aligned}
K(P, Q) &= \int_{\Omega} \log\left(\frac{p(x)}{q(x)}\right) p(x) \,\mathrm{d}\mu(x) \\
&= \int_{\{x:\ p(x)q(x)>0\}} \log\left(\frac{p(x)}{q(x)}\right) p(x) \,\mathrm{d}\mu(x) \\
&= -2 \int_{\{x:\ p(x)q(x)>0\}} \log\left(\sqrt{\frac{q(x)}{p(x)}} - 1 + 1\right) p(x) \,\mathrm{d}\mu(x) \\
&\geq -2 \int_{\{x:\ p(x)q(x)>0\}} \left(\sqrt{\frac{q(x)}{p(x)}} - 1\right) p(x) \,\mathrm{d}\mu(x) \\
&= 2 - 2 \int_{\{x:\ p(x)q(x)>0\}} \sqrt{p(x)q(x)} \,\mathrm{d}\mu(x).
\end{aligned}$$

Observing that

$$H^2(P, Q) = \int_{\Omega} (\sqrt{p(x)} - \sqrt{q(x)})^2 \,\mathrm{d}\mu(x) = 2 - 2 \int_{\{x:\ p(x)q(x)>0\}} \sqrt{p(x)q(x)} \,\mathrm{d}\mu(x),$$

we conclude that

$$H^2(P, Q) \leq K(P, Q).$$

We prove the second point. We have

$$\begin{aligned}\|P-Q\|_1^2 &= \left(\int_\Omega |p(x)-q(x)|\,\mathrm{d}\mu(x)\right)^2\\ &= \left(\int_\Omega |\sqrt{p(x)}-\sqrt{q(x)}||\sqrt{p(x)}+\sqrt{q(x)}|\,\mathrm{d}\mu(x)\right)^2\end{aligned}$$

Applying the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\|P-Q\|_1^2 &\leq \int_\Omega |\sqrt{p(x)}-\sqrt{q(x)}|^2\,\mathrm{d}\mu(x)\times\int_\Omega |\sqrt{p(x)}+\sqrt{q(x)}|^2\,\mathrm{d}\mu(x)\\ &\leq H^2(P,Q)\times\Big(2+2\int_\Omega\sqrt{p(x)q(x)}\,\mathrm{d}\mu(x)\Big)\\ &\leq 4H^2(P,Q).\end{aligned}$$

Since for any $x\in\Omega$,

$$(\sqrt{p(x)}-\sqrt{q(x)})^2\leq p(x)+q(x)-2\min(p(x,q(x))=|p(x)-q(x)|,$$

which provides the last upper bound of the proposition. □

2.2 Probability measures on $\mathcal{M}(\Omega)$ when Ω is finite

In this section, we assume that Ω is finite and without loss of generality, we assume that

$$\Omega=\{1,2,\dots,k\}.$$

In this case, denoting for any $i\in\{1,\dots,k\}$, p_i the probability of i, $\mathcal{M}(\Omega)$ can be identified with the simplex

$$S_k:=\left\{(p_1,\dots,p_k):\ p_i\geq 0\,\forall i\in\{1,\dots,k\}\text{ and }\sum_{i=1}^k p_i=1\right\}.$$

To define a probability measure on $\mathcal{M}(\Omega)$, we just have to define a probability measure on S_k.

2.2.1 Polya-tree construction

We assume that $k=2^\ell$. We build a dyadic partition of $\Omega=\{1,\dots,2^\ell\}$. We set

$$B_0=\{1,\dots,2^{\ell-1}\},\quad B_1=\{2^{\ell-1}+1,\dots,2^\ell\},$$

$$B_{00}=\{1,\dots,2^{\ell-2}\},\quad B_{01}=\{2^{\ell-2}+1,\dots,2^{\ell-1}\},$$

$$B_{10} = \{2^{\ell-1}+1, \ldots, 2^{\ell-1}+2^{\ell-2}\}, \quad B_{11} = \{2^{\ell-1}+2^{\ell-2}+1, 2^{\ell}\},$$

and so on. In this case, if $x \in \Omega$, $\exists!(\varepsilon_1, \ldots, \varepsilon_\ell) \in \{0,1\}^\ell$ such that

$$B_{\varepsilon_1 \ldots \varepsilon_\ell} = \{x\}.$$

Conversely, any sequence $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_\ell) \in \{0,1\}^\ell$ corresponds to the point $\cap_{i=1}^{\ell} B_{\varepsilon_1 \ldots \varepsilon_i}$. So, we have a correspondence between $\mathcal{M}(\Omega)$ and $\mathcal{M}(\{0,1\}^\ell)$. A way of building a prior on $\mathcal{M}(\Omega)$ is to define a prior on the random variables

$$Y_{\varepsilon_1 \ldots \varepsilon_{i-1}} := P(\varepsilon_i = 0 | \varepsilon_1, \ldots, \varepsilon_{i-1})$$

for $i = 2, \ldots, \ell$ and on

$$Y_\emptyset = P(\varepsilon_1 = 0).$$

Indeed, we have $\forall(\varepsilon_1, \ldots, \varepsilon_\ell) \in \{0,1\}^\ell$

$$\begin{aligned} P(\varepsilon_1, \ldots, \varepsilon_\ell) &= P(\varepsilon_\ell | \varepsilon_1, \ldots, \varepsilon_{\ell-1}) P(\varepsilon_1, \ldots, \varepsilon_{\ell-1}) \\ &= Y_{\varepsilon_1 \ldots \varepsilon_{\ell-1}}^{1_{\{\varepsilon_\ell = 0\}}} (1 - Y_{\varepsilon_1 \ldots \varepsilon_{\ell-1}})^{1_{\{\varepsilon_\ell = 1\}}} P(\varepsilon_1, \ldots, \varepsilon_{\ell-1}) \\ &= \prod_{i=1}^{\ell} \left(Y_{\varepsilon_1 \ldots \varepsilon_{i-1}}^{1_{\{\varepsilon_i = 0\}}} (1 - Y_{\varepsilon_1 \ldots \varepsilon_{i-1}})^{1_{\{\varepsilon_i = 1\}}} \right). \end{aligned}$$

Actually, by using the correspondence $\varepsilon_1, \ldots, \varepsilon_i \leftrightarrow P(B_{\varepsilon_1 \ldots \varepsilon_i})$ a prior on $\mathcal{M}(\Omega)$ is parametrized as a prior on $(P(B_0), P(B_{00}|B_0), P(B_{10}|B_1), \ldots, \mathbb{P}(B_{\varepsilon_1 \ldots \varepsilon_{i-1} 0} | B_{\varepsilon_1 \ldots \varepsilon_{i-1}}), \ldots)$. A special case of interest is the case where the variables $Y_\varepsilon = P(B_{\varepsilon 0} | B_\varepsilon)$ are all independent. These priors are called **tail free priors**. When they are independent Beta variables, they are called **Polya free priors**.

2.2.2 Finite-dimensional Dirichlet distributions

Dirichlet processes were introduced by Ferguson (1973) to give a Bayesian interpretation of nonparametric estimation problems. They have tractable distributions and nice consistency properties. They are also easy to elicit and provide a natural interpretation. They constitute starting points for more complex prior distributions. Dirichlet processes are a natural extension of finite-dimensional Dirichlet priors.

1. Case $\Omega = \{1, 2\}$.

In this case, we have

$$\mathcal{M}(\Omega) = \{p = (p_1, p_2) : \; p_1 \geq 0, \, p_2 \geq 0, \, p_1 + p_2 = 1\}.$$

For $p = (p_1, p_2) \in \mathcal{M}(\Omega)$, since $p_2 = 1 - p_1$, we just have to define a prior on p_1 to have a prior on p.

Definition 2.1. *We say that $p = (p_1, p_2)$ has a $Beta(\alpha_1, \alpha_2)$-**prior** for $\alpha_1 > 0$ and $\alpha_2 > 0$ if the density (with respect to the Lebesgue measure) of the prior distribution of p_1 is*

$$g(p_1) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} p_1^{\alpha_1 - 1}(1 - p_1)^{\alpha_2 - 1} 1_{(0,1)}(p_1), \quad p_1 \in \mathbb{R}.$$

We denote

$$p \sim Beta(\alpha_1, \alpha_2) \quad (or \quad p_1 \sim Beta(\alpha_1, \alpha_2)).$$

Proposition 2.4. *If $p \sim Beta(\alpha_1, \alpha_2)$, we have:*

$$\mathbb{E}[p_1] = \frac{\alpha_1}{\alpha_1 + \alpha_2}, \quad var(p_1) = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)}.$$

Remark 2.2. *Observe that if p_{1n} has a $Beta(\alpha_{1n}, \alpha_{2n})$-distribution with $\alpha_{1n} \overset{n \to +\infty}{\longrightarrow} 0$ and $\alpha_{2n} \overset{n \to +\infty}{\longrightarrow} c > 0$, then the Markov inequality implies that $p_{1n} \overset{n \to +\infty}{\longrightarrow} 0$ in probability. So, we adopt the following convention: if the prior distribution of p_1 is $Beta(0, \alpha_2)$ with $\alpha_2 > 0$, then we force $p_1 \equiv 0$ almost everywhere.*

We now characterize the Beta-distributions by using Gamma-distributions.

Definition 2.2. *We say that a random variable Z has a **Gamma-distribution with parameter** $\alpha > 0$ if its density (with respect to the Lebesgue measure) is*

$$g(z) = \frac{1}{\Gamma(\alpha)} z^{\alpha - 1} e^{-z} 1_{(0;+\infty)}(z), \quad z \in \mathbb{R}.$$

We denote

$$Z \sim \Gamma(\alpha).$$

Proposition 2.5. *Let $\alpha_1 > 0$ and $\alpha_2 > 0$. If $Z_1 \sim \Gamma(\alpha_1)$, $Z_1 \sim \Gamma(\alpha_2)$ with Z_1 and Z_2 independent, then $Z_1/(Z_1 + Z_2)$ and $Z_1 + Z_2$ are independent and*

$$\frac{Z_1}{Z_1 + Z_2} \sim Beta(\alpha_1, \alpha_2), \quad Z_1 + Z_2 \sim \Gamma(\alpha_1 + \alpha_2).$$

Proof. The densities of Z_1 and Z_1 are respectively

$$f_1(z) = \frac{1}{\Gamma(\alpha_1)} z^{\alpha_1 - 1} e^{-z} 1_{(0;+\infty)}(z), \quad z \in \mathbb{R},$$

and

$$f_2(z) = \frac{1}{\Gamma(\alpha_2)} z^{\alpha_2 - 1} e^{-z} 1_{(0;+\infty)}(z), \quad z \in \mathbb{R}.$$

So, for any bounded Borelian functions ϕ and ψ,

$$\begin{aligned}\mathbb{E}\Big[\phi\Big(\frac{Z_1}{Z_1+Z_2}\Big)\psi(Z_1+Z_2)\Big] &= \int_0^{+\infty}\int_0^{+\infty}\phi\Big(\frac{x}{x+y}\Big)\psi(x+y)\frac{x^{\alpha_1-1}}{\Gamma(\alpha_1)}\frac{y^{\alpha_2-1}}{\Gamma(\alpha_2)}e^{-(x+y)}\,\mathrm{d}x\,\mathrm{d}y\\ &= \int_0^1 \mathrm{d}t\int_0^{+\infty}\phi(t)\psi(u)\frac{(tu)^{\alpha_1-1}}{\Gamma(\alpha_1)}\frac{(u(1-t))^{\alpha_2-1}}{\Gamma(\alpha_2)}e^{-u}\times u\,\mathrm{d}u\\ &= \int\phi(t)\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}t^{\alpha_1-1}(1-t)^{\alpha_2-1}1_{(0;1)}(t)\,\mathrm{d}t\\ &\qquad\times\int\psi(u)\frac{1}{\Gamma(\alpha_1+\alpha_2)}u^{\alpha_1+\alpha_2-1}e^{-u}1_{(0;+\infty)}(u)\,\mathrm{d}u,\end{aligned}$$

where we have used the change of variables $t = x/(x+y)$ and $u = x+y$. □

We recall the following definition.

Definition 2.3. *We say that the sequence of variables $(X_i)_{i\in\mathbb{N}^*}$ is* ***exchangeable*** *if for any $n\in\mathbb{N}^*$ and any permutation σ of the first n integers, we have*

$$(X_1,\dots,X_n)\sim(X_{\sigma(1)},\dots,X_{\sigma(n)}).$$

Proposition 2.6. *Assume that conditionally on p, $(X_1,\dots,X_n)$ is an n-sample of $Ber(p)$-variables; assume also that p is distributed according to a $Beta(\alpha_1,\alpha_2)$-distribution with $\alpha_1\geq 0$ and $\alpha_2\geq 0$ with $\alpha_1+\alpha_2>0$. Then,*

$$p|X_1,\dots,X_n\sim Beta\Big(\alpha_1+\sum_{i=1}^n\delta_{X_i}(1),\alpha_2+\sum_{i=1}^n\delta_{X_i}(2)\Big).$$

This result means that the Beta distribution is a conjugate prior of the Bernoulli distribution. Furthermore, $(X_1,\dots,X_n)$ are exchangeable and

$$m(i) := P(X_1=i) = \frac{\alpha_i}{\alpha_1+\alpha_2},\quad i=1,2.$$

Proof. The density of the joint distribution of $(X_1,\dots,X_n,p_1)$ is the function

$$(X_1,\dots,X_n,p_1)\mapsto p_1^{\sum_{i=1}^n\delta_{X_i}(1)}(1-p_1)^{\sum_{i=1}^n\delta_{X_i}(2)}\times\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}p_1^{\alpha_1-1}(1-p_1)^{\alpha_2-1}1_{(0;1)}(p_1).$$

Therefore, the conditional distribution of $p_1|X_1,\dots,X_n$ has a density proportional to the function

$$p_1\longmapsto\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}p_1^{\sum_{i=1}^n\delta_{X_i}(1)+\alpha_1-1}(1-p_1)^{\sum_{i=1}^n\delta_{X_i}(2)+\alpha_2-1}1_{(0;1)}(p_1),$$

which proves the first point and the density of the marginal distribution of $(X_1, \ldots, X_n)$ is the function

$$(X_1, \ldots, X_n) \longmapsto \int_0^1 \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} p_1^{\sum_{i=1}^n \delta_{X_i}(1)+\alpha_1-1}(1-p_1)^{\sum_{i=1}^n \delta_{X_i}(2)+\alpha_2-1} 1_{(0;1)}(p_1)\, \mathrm{d}p_1.$$

In particular, with $n = 1$,

$$\begin{aligned} m(1) := P(X_1 = 1) &= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^1 p_1^{\alpha_1}(1-p_1)^{\alpha_2-1}\, \mathrm{d}p_1 \\ &= \frac{\alpha_1}{\alpha_1 + \alpha_2}. \end{aligned}$$

Similar computations can be done for $m(2)$. □

2. Case $\Omega = \{1, 2, \ldots, k\}$.

We have to specify a prior on the simplex

$$\mathcal{S}_k := \left\{ p = (p_1, \ldots, p_k) : \; p_i \geq 0, \forall\, i, \; \sum_{i=1}^k p_i = 1 \right\}.$$

We set for all $A \subset \Omega$, $p(A) = \sum_{i \in A} p_i$. In particular $p(\{i\}) = p_i$ for any $i \in \Omega$. We have the following definition.

Definition 2.4. *Let $\alpha = (\alpha_1, \ldots, \alpha_k) \in (\mathbb{R}_+^*)^k$. We say that p has a Dirichlet distribution with parameter α if the density of p with respect to the Lebesgue measure is*

$$g(p) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} p_1^{\alpha_1-1} p_2^{\alpha_2-1} \cdots p_k^{\alpha_k-1} 1_{\mathcal{S}_k}(p).$$

We denote $p \sim \mathcal{D}(\alpha)$.

Remark 2.3. *Observe that if $k = 2$, with $\alpha = (\alpha_1, \alpha_2)$, $\mathcal{D}(\alpha) = Beta(\alpha_1, \alpha_2)$.*

Remark 2.4. *Exactly as before, we extend the previous definition to the case where $\alpha_i \geq 0, \forall\, i$ and $\sum_{i=1}^k \alpha_i > 0$. In this case, if $\alpha_i = 0$ we force $p_i \equiv 0$ almost everywhere and we interpret the previous density as a density on a lower-dimensional set.*

Remark 2.5. *If for all $i \in \{1, \ldots, k\}$, $\alpha_i = 1$, then g is constant on $\mathcal{S}_k$ and*

$$g(p) = \Gamma(k) 1_{\mathcal{S}_k}(p) = (k-1)! 1_{\mathcal{S}_k}(p).$$

Proposition 2.7. *Assume that we are given k independent variables $Z_i \sim \Gamma(\alpha_i)$ with for all i, $\alpha_i > 0$. Then, with $\alpha = (\alpha_1, \dots, \alpha_k)$,*

$$\left(\frac{Z_1}{\sum_{i=1}^k Z_i}, \dots, \frac{Z_k}{\sum_{i=1}^k Z_i}\right) \sim \mathcal{D}(\alpha)$$

and the previous vector is independent of $\sum_{i=1}^k Z_i$.

Proof. The proof is tedious but similar to the one for the case $k = 2$. □

Proposition 2.8. *If $p = (p_1, \dots, p_k) \sim \mathcal{D}(\alpha_1, \dots, \alpha_k)$, then for any partition $A_1, \dots, A_m$ of Ω we have*

$$(p(A_1), \dots, p(A_m)) \sim \mathcal{D}\bigg(\sum_{i \in A_1} \alpha_i, \dots, \sum_{i \in A_m} \alpha_i\bigg).$$

Proof. We use the characterization of Dirichlet distributions based on Gamma-variables. We take k independent variables $Z_i \sim \Gamma(\alpha_i)$ so that

$$p \sim \left(\frac{Z_1}{\sum_{i=1}^k Z_i}, \dots, \frac{Z_k}{\sum_{i=1}^k Z_i}\right).$$

Then,

$$\begin{aligned}(p(A_1), \dots, p(A_m)) &\sim \Big(\sum_{i \in A_1} p_i, \dots, \sum_{i \in A_m} p_i\Big) \\ &\sim \left(\frac{\sum_{i \in A_1} Z_i}{\sum_{i=1}^k Z_i}, \dots, \frac{\sum_{i \in A_m} Z_i}{\sum_{i=1}^k Z_i}\right)\end{aligned}$$

and we observe that for all $j \in \{1, \dots, m\}$,

$$\sum_{i \in A_j} Z_i \sim \Gamma\Big(\sum_{i \in A_j} \alpha_i\Big),$$

which is independent from $\sum_{i \in A_{j'}} Z_i$ for any $j' \neq j$. □

Definition 2.5. *Let $\alpha = (\alpha_1, \dots, \alpha_k)$ a vector of non-negative components. We define the **measure** α on Ω by:*

$$\alpha(A) = \sum_{i \in A} \alpha_i, \quad A \subset \Omega.$$

Observe that $\alpha(\{i\}) = \alpha_i \; \forall \, i \in \{1, \dots, k\}$.

We now prove successively several useful results.

Proposition 2.9. *Let $A_1, \ldots, A_m$ be a partition of Ω. Let $i \in \{1, \ldots, k\}$. If $\alpha(A_i) > 0$ then we set:*

$$p(j|A_i) := \frac{p(\{j\})}{p(A_i)}, \quad j \in A_i$$

If $\alpha(A_i) = 0$ then we set:

$$p(j|A_i) := \frac{1}{|A_i|}, \quad j \in A_i.$$

Now, if $p \sim \mathcal{D}(\alpha)$, then

1. *For any i, if $\alpha(A_i) > 0$, then $p(\cdot|A_i) \sim \mathcal{D}(\alpha/A_i)$, where α/A_i is the measure α restricted to A_i.*

2. *For any $i \neq i'$, for any $j \in A_i$, $p(j|A_i)$ is independent from $(p(A_i), p(A_{i'}))$. Therefore, for any i, for any $j_i \in A_i$ $(p(A_1), \ldots, p(A_m)), (p(j_i|A_i))_i$ are independent vectors.*

Proof. We prove the results for the case where $\alpha(A_i) > 0$ for any $i \in \Omega$. As before, we use the characterization of Dirichlet distributions based on Gamma-variables. We take k independent variables $Z_i \sim \Gamma(\alpha_i)$ so that

$$p \sim \left(\frac{Z_1}{\sum_{i=1}^k Z_i}, \ldots, \frac{Z_k}{\sum_{i=1}^k Z_i} \right).$$

We have for any $j \in A_i$,

$$p(j|A_i) = \frac{p(\{j\})}{p(A_i)} \sim \frac{\frac{Z_j}{\sum_{i=1}^k Z_i}}{\frac{\sum_{j' \in A_i} Z_{j'}}{\sum_{i=1}^k Z_i}} = \frac{Z_j}{\sum_{j' \in A_i} Z_{j'}},$$

which proves the first point. Furthermore, for i fixed and $j \in A_i$, $\frac{Z_j}{\sum_{j' \in A_i} Z_{j'}}$ is independent from $\sum_{j' \in A_i} Z_{j'}$ and then independent from

$$\sum_{j' \in A_i} Z_{j'} + \sum_{i' \neq i} \sum_{j' \in A_{i'}} Z_{j'} = \sum_{j'=1}^k Z_{j'}.$$

This leads to the second point. □

Proposition 2.10. *Let α and α' two measures on Ω. If p and q are two independent k-dimensional Dirichlet random vectors with parameters α and α' and if w is independent of p and q and $w \sim Beta(\alpha(\Omega), \alpha'(\Omega))$, then*

$$wp + (1 - w)q \sim \mathcal{D}(\alpha + \alpha').$$

Proof. We take $2k$ independent variables $Z_i \sim \Gamma(\alpha_i)$ and $Z_{i+k} \sim \Gamma(\alpha'_i)$ so that

$$p \sim \left(\frac{Z_1}{\sum_{i=1}^k Z_i}, \dots, \frac{Z_k}{\sum_{i=1}^k Z_i} \right), \quad q \sim \left(\frac{Z_{1+k}}{\sum_{i=k+1}^{2k} Z_i}, \dots, \frac{Z_{2k}}{\sum_{i=k+1}^{2k} Z_i} \right)$$

and

$$w \sim \frac{\sum_{i=1}^k Z_i}{\sum_{i=1}^{2k} Z_i}.$$

Then,

$$\begin{aligned} wp + (1-w)q &\sim \frac{\sum_{i=1}^k Z_i}{\sum_{i=1}^{2k} Z_i} \times \left(\frac{Z_1}{\sum_{i=1}^k Z_i}, \dots, \frac{Z_k}{\sum_{i=1}^k Z_i} \right) \\ &\quad + \frac{\sum_{i=k+1}^{2k} Z_i}{\sum_{i=1}^{2k} Z_i} \times \left(\frac{Z_{1+k}}{\sum_{i=k+1}^{2k} Z_i}, \dots, \frac{Z_{2k}}{\sum_{i=k+1}^{2k} Z_i} \right) \\ &\sim \left(\frac{Z_1 + Z_{1+k}}{\sum_{i=1}^{2k} Z_i}, \dots, \frac{Z_k + Z_{2k}}{\sum_{i=1}^{2k} Z_i} \right) \\ &\sim \mathcal{D}(\alpha + \alpha'). \end{aligned}$$

□

Proposition 2.11. *Assume that $p \sim \mathcal{D}(\alpha)$ and conditionally on p, $(X_1, \dots, X_n)$ is an n-sample distributed according to p. Then,*

$$p | X_1, \dots, X_n \sim \mathcal{D}\Big(\alpha + \sum_{j=1}^n \delta_{X_j} \Big).$$

This result means that the Dirichlet distribution is a conjugate prior of the multivariate distribution.

Proof. the density of the joint distribution of $(X_1, \dots, X_n, p)$ is proportional, on $\Omega^n \times \mathcal{S}_k$, to the function

$$(X_1, \dots, X_n, p) \mapsto \prod_{i=1}^k p_i^{\alpha_i - 1} \times \prod_{i=1}^k p_i^{\sum_{j=1}^n 1_{\{X_j = i\}}} = \prod_{i=1}^k p_i^{\alpha_i + \sum_{j=1}^n 1_{\{X_j = i\}} - 1}.$$

Therefore,

$$p | X_1, \dots, X_n \sim \mathcal{D}\Big(\alpha + \sum_{j=1}^n \delta_{X_j} \Big).$$

□

Proposition 2.12. *Assume that $p \sim \mathcal{D}(\alpha)$ and conditionally on p, $(X_1,\dots,X_n)$ is an n-sample distributed according to p. Then, for any $A \subset \Omega$,*

$$m(A) = \frac{\alpha(A)}{\alpha(\Omega)}.$$

Proof. The 2-dimensional vector $(p(A), p(A^c))$ is distributed according to $\Pi = \mathcal{D}(\alpha(A), \alpha(A^c)) = Beta(\alpha(A), \alpha(A^c))$. Then,

$$m(A) := P(X_1 \in A) = \int p(A)\,\mathrm{d}\Pi(p) = \mathbb{E}[p(A)] = \frac{\alpha(A)}{\alpha(A)+\alpha(A^c)} = \frac{\alpha(A)}{\alpha(\Omega)}.$$

□

Remark 2.6. *The previous results imply that the* ***predictive distribution*** *is given for all $A \subset \Omega$ by*

$$P(X_{n+1} \in A | X_1,\dots,X_n) = \mathbb{E}[p(A)|X_1,\dots,X_n] = \frac{\alpha(A) + \sum_{j=1}^n 1_{\{X_j \in A\}}}{\alpha(\Omega)+n}.$$

Observe that the X_i's are not independent.

Remark 2.7. *If we denote $\bar\alpha$ the renormalized measure $\alpha/\alpha(\Omega)$, previous results say that if $p \sim \mathcal{D}(\alpha)$ and conditionally on p, $X_1 \sim p$, then we have proved that*

$$p|X_1 \sim \mathcal{D}(\alpha + \delta_{X_1}) \quad and \quad X_1 \sim \bar\alpha.$$

Therefore if $p \sim \Pi = \mathcal{D}(\alpha)$, with $m = \bar\alpha$ the marginal distribution of X_1, for any Borelian function Φ,

$$\begin{aligned}\mathbb{E}_\Pi[\Phi(p)] = \int \Phi(p) d\mathcal{D}(\alpha)(p) &= \int \Phi(p) \int_{x\in\Omega} d\mathcal{D}(\alpha)\big(p|X_1 = x\big) d\bar\alpha(x)\\ &= \int \Phi(p) \sum_{i=1}^k \bar\alpha(i) d\mathcal{D}(\alpha+\delta_i)(p).\end{aligned}$$

Since the previous equality is true for any function Φ, it means that

$$\mathcal{D}(\alpha) \sim \sum_{i=1}^k \bar\alpha(i)\mathcal{D}(\alpha+\delta_i). \tag{2.1}$$

2.2.3 Dirichlet distributions via Polya urn schemes

In this section, we still consider the finite case $\Omega = \{1, 2, \dots, k\}$ and we describe a scheme corresponding to drawing samples from $p \sim \mathcal{D}(\alpha)$ even if p is not observable (since it is a realization of $\mathcal{D}(\alpha)$). This is the so-called **Polya urn scheme**. To describe it, we first assume that for any $i \in \Omega$, $\alpha(i)$ is an integer.

We consider a (Polya) urn with $\alpha(\Omega)$ balls. For any color $i \in \Omega$, there are $\alpha(i)$ balls. We draw balls randomly and we replace each drawn ball by two balls of the same color, namely the color of the drawn ball. We denote $X_i = j$ if the ith drawn ball is of color j. Then, we have

$$P(X_1 = j) = \frac{\alpha(j)}{\alpha(\Omega)},$$

$$P(X_2 = j | X_1) = \frac{\alpha(j) + \delta_{X_1}(j)}{\alpha(\Omega) + 1}$$

and so on, and at step $n+1$, we have:

$$P(X_{n+1} = j | X_1, \dots, X_n) = \frac{\alpha(j) + \sum_{i=1}^n \delta_{X_i}(j)}{\alpha(\Omega) + n},$$

which corresponds to the predictive distribution of Remark 2.6. We then have for all $(x_1, \dots, x_n) \in \Omega^n$,

$$\begin{aligned} m(x_1, \dots, x_n) &= P(X_1 = x_1, \dots, X_n = x_n) \\ &= \frac{\alpha(x_1)}{\alpha(\Omega)} \prod_{i=1}^{n-1} \left(\frac{\alpha(x_{i+1}) + \sum_{j=1}^i \delta_{x_j}(x_{i+1})}{\alpha(\Omega) + i} \right). \end{aligned}$$

If we denote for any $i \in \Omega$, $n_i = \sum_{j=1}^n 1_{\{x_j = i\}}$, and for any $\beta > 0$ and $n \in \mathbb{N}$,

$$[\beta]^{[n]} = \begin{cases} \beta \times (\beta + 1) \times \cdots \times (\beta + n - 1) & \text{if} \quad n \geq 1, \\ 1 & \text{if} \quad n = 0, \end{cases}$$

then

$$\begin{aligned} m(x_1, \dots, x_n) &= \frac{[\alpha(1)]^{[n_1]} \times \cdots \times [\alpha(k)]^{[n_k]}}{\alpha(\Omega) \times (\alpha(\Omega) + 1) \times \cdots \times (\alpha(\Omega) + n - 1)} \\ &= \frac{[\alpha(1)]^{[n_1]} \times \cdots \times [\alpha(k)]^{[n_k]}}{[\alpha(\Omega)]^{[n]}}. \end{aligned}$$

We have established that $(X_j)_{j \in \mathbb{N}^*}$ is exchangeable. We now state de Finetti's theorem.

Theorem 2.1. *A sequence of Ω-valued random variables is exchangeable if and only if there is a unique measure Π on $\mathcal{M}(\Omega)$ such that for all n, for all $(x_1, \dots, x_n) \in \Omega^n$,*

$$\int_{\mathcal{M}(\Omega)} \prod_{i=1}^n p(x_i) \, \mathrm{d}\Pi(p) = P(X_1 = x_1, \dots, X_n = x_n).$$

The direct application of the previous result shows that there exists Π such that for any n and any $x_1, \dots, x_n$,

$$m(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n) = \int_{\mathcal{M}(\Omega)} \prod_{i=1}^{n} p(x_i) \, \mathrm{d}\Pi(p).$$

This result establishes that the Polya urn scheme corresponds to a Bayesian procedure and using Remark 2.6, we have $\Pi = \mathcal{D}(\alpha)$.

We can generalize the Polya urn scheme to non-integers, by considering the following (slightly) more sophisticated scheme. At the beginning, we have an empty urn and a k-set of colors.

- At step 1, we pick a new color with probability distribution $\alpha(\cdot)/\alpha(\Omega)$ from the set of colors. We paint a new ball that color and add it to the urn.
- At step $n+1$, with probability $\alpha(\Omega)/(n+\alpha(\Omega))$ we apply step 1, with probability $n/(n+\alpha(\Omega))$ we pick a ball out of the urn and put it back with another ball of the same color.

In any case, after step $n+1$, the number of balls in the urn is $n+1$ and the color drawn has one more representative in the urn. Note that when $n \to +\infty$, the urn has more importance. The larger n, the higher the probability that it will grow. This is a "*rich-gets-richer phenomenon*". Note that, if for any $n \in \mathbb{N}^*$, X_n is the color of the ball put into the urn at step n, for any $j \in \Omega$,

$$\begin{aligned} P(X_{n+1} = j | X_1, \dots, X_n) &= \frac{\alpha(\Omega)}{n+\alpha(\Omega)} \times \frac{\alpha(j)}{\alpha(\Omega)} + \frac{n}{n+\alpha(\Omega)} \times \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(j) \\ &= \frac{\alpha(j) + \sum_{i=1}^{n} \delta_{X_i}(j)}{n+\alpha(\Omega)}. \end{aligned}$$

We can interpret the previous setting in terms of **Chinese restaurants**. Indeed, let us consider the meal $j \in \Omega = \{1, \dots, k\}$ of customers, in a Chinese restaurant, who sit around tables of infinite capacity.

- The first customer enters the restaurant and sits at the first table.
- Let us assume that the first n clients are seated at tables $1, \dots, K_n$, where K_n is the number of occupied tables. The customer $n+1$ has the choice to have a seat between two already present customers or to have a seat at table $K_n + 1$. We decide each customer has the weight 1 and the new table $\alpha(\Omega)$. Since there are n available places between all seated customers, the probability the customer $n+1$ sits at one of these places is $(\alpha(\Omega)+n)^{-1}$, the probability the customer $n+1$ sits at a new table is $\alpha(\Omega) \times (\alpha(\Omega)+n)^{-1}$.
- If he sits at a new table, he chooses his meal $j \in \Omega$ with probability $\alpha(j)/\alpha(\Omega)$. If he sits at an already occupied table, he eats the same meal as the other customers of the table.

- With X_{n+1} the meal of customer $n+1$, we have for any $j \in \Omega$,

$$\begin{aligned} P(X_{n+1} = j | X_1, \ldots, X_n) &= \frac{\alpha(\Omega)}{n + \alpha(\Omega)} \times \frac{\alpha(j)}{\alpha(\Omega)} + \frac{1}{n + \alpha(\Omega)} \times \sum_{i=1}^{n} \delta_{X_i}(j) \\ &= \frac{\alpha(j) + \sum_{i=1}^{n} \delta_{X_i}(j)}{n + \alpha(\Omega)}. \end{aligned}$$

The last expression is the expression we obtained for Polya urn schemes.

2.3 Probability measures on $\mathcal{M}(\mathbb{R})$

This section is devoted to the extension of the prior models built on $\mathcal{M}(\Omega)$ for Ω finite to the case where $\Omega = \mathbb{R}$. We consider the statistical model $(\mathbb{R}, \mathcal{B}, P \in \mathcal{M}(\mathbb{R}))$ with $\mathcal{B}$ the Borelian σ-algebra. It is quite easy to extend the Polya-tree construction (see Ghosh and Ramamoorthi (2003)). The construction of Dirichlet processes is much more involved.

2.3.1 Tail free priors

We extend to $\mathbb{R}$ the construction built for a finite set Ω. We denote

$$E_k = \{0,1\}^k, \ k \in \mathbb{N}^*, \quad E_0 = \emptyset, \quad E^* = \bigcup_{k \in \mathbb{N}} E_k.$$

We start with a partition of $\mathbb{R}$, $I_0 = \{B_0, B_1\}$. Then, we set

$$I_1 = \{B_{00}, B_{01}, B_{10}, B_{11}\}, \text{ with } B_0 = B_{00} \cup B_{01} \text{ and } B_1 = B_{10} \cup B_{11}$$

and then for any $n \geq 2$,

$$I_n = \{B_{\varepsilon 0}, B_{\varepsilon 1} \ \varepsilon \in E_n\}, \text{ with } B_{\varepsilon 0} \cup B_{\varepsilon 1} \text{ partition of } B_\varepsilon.$$

We assume that $\sigma(B_\varepsilon, \varepsilon \in E^*) = \mathcal{B}$ and we introduce tail free prior distributions.

Definition 2.6. *We say that $\Pi \in \mathcal{M}(\mathbb{R})$ is tail free with respect to $I = (I_n)_{n \in \mathbb{N}^*}$ if the following rows are independent:*
- $P(B_0)$
- $P(B_{00}|B_0)$, $P(B_{10}|B_1)$
- $P(B_{000}|B_{00})$, $P(B_{010}|B_{01})$ $P(B_{100}|B_{10})$, $P(B_{110}|B_{11})$
- $\cdots$
- $P(B_{\varepsilon 0}|B_\varepsilon)$, $\varepsilon \in E_n$
- $\cdots$

We can now construct a tail free prior on $\mathcal{M}(\mathbb{R})$. Let $Q = \{q_\varepsilon,\ \varepsilon \in E^*\}$ a dense subset of $\mathbb{R}$ such that for any ε, $q_{\varepsilon 0} < q_\varepsilon < q_{\varepsilon 1}$. We set $B_0 =]-\infty, q_0]$, $B_1 =]q_0, +\infty[$. Then, we set $B_{00} =]-\infty, q_{00}]$, $B_{01} =]q_{00}, q_0]$, $B_{10} =]q_0, q_{01}]$, $B_{11} =]q_{01}, +\infty[$. Proceeding this way, we build successive partitions of $\mathbb{R}$, consisting in sets of the form $B_{\varepsilon_1 \cdots \varepsilon_n}$, with $(\varepsilon_1 \cdots \varepsilon_n) \in E_n$. Note that $\{B_{\varepsilon 0}, B_{\varepsilon 1}\}$ is a partition of B_ε. Now, since Q is dense, it can be proved that $\sigma(B_\varepsilon, \varepsilon \in E^*) = \mathcal{B}$ and we obtain the following result.

Theorem 2.2. *Let $Y = P(B_0)$ and $Y_\varepsilon = P(B_{\varepsilon 0}|B_\varepsilon)$ for any $\varepsilon \in E^*$. We assume that the random variables $(Y_\varepsilon)_{\varepsilon \in E^*}$ satisfy*

1. *$Y \perp \{Y_0, Y_1\} \perp \{Y_{00}, Y_{01}, Y_{10}, Y_{11}\} \perp \cdots$, where $\perp$ means "independent"*

2. *$Y_{\varepsilon 0} \times Y_{\varepsilon 00} \times Y_{\varepsilon 000} \times \cdots = 0$ and $Y_1 \times Y_{11} \times Y_{111} \times \cdots = 0$*

then there exists a tail free prior Π on $\mathcal{M}(\mathbb{R})$ such that under Π,

$$Y_\varepsilon = P(B_{\varepsilon 0}|B_\varepsilon).$$

Observe that given the variables $(Y_\varepsilon)_{\varepsilon \in E^*}$, we obtain a joint distribution for $(P(B_\varepsilon))_{\varepsilon \in E^*}$. The next theorem gives conjugacy properties.

Theorem 2.3. *Suppose Π is a tail free prior on $\mathcal{M}(\mathbb{R})$ with respect to the sequence of partitions $(I_n)_{n \geq 1}$. Given P, let $X_, \ldots, X_n \overset{i.i.d.}{\sim} P$. Then, the posterior is also tail free with respect to $(I_n)_{n \geq 1}$.*

2.3.2 Dirichlet processes

Dirichlet process priors are a natural generalization to $\mathcal{M}(\mathbb{R})$ of the finite-dimensional Dirichlet distributions.

Theorem 2.4. *Let α a finite measure on $(\mathbb{R}, \mathcal{B})$. There exists a unique probability measure $\mathcal{D}_\alpha$ on $\mathcal{M}(\mathbb{R})$ called the Dirichlet process with parameter α satisfying, if $P \sim \mathcal{D}_\alpha$, for any partition $(B_1, B_2, \ldots, B_k)$ of $\mathbb{R}$,*

$$(P(B_1), P(B_2), \ldots, P(B_k)) \sim \mathcal{D}(\alpha(B_1), \alpha(B_2), \ldots, \alpha(B_k)),$$

where the latter is the Dirichlet distribution with parameter $(\alpha(B_1), \alpha(B_2), \ldots, \alpha(B_k))$.

See Ghosh and Ramamoorthi (2003) for the proof. We now establish properties associated with Dirichlet processes. For α a finite measure on $(\mathbb{R}, \mathcal{B})$, we denote $\bar{\alpha}$ the probability measure defined by

$$\bar{\alpha}(\cdot) = \frac{\alpha(\cdot)}{\alpha(\mathbb{R})}.$$

Proposition 2.13. *If $P \sim \mathcal{D}_\alpha$, then for any $A \in \mathcal{B}$,*

$$\mathbb{E}[P(A)] = \bar{\alpha}(A), \quad var(P(A)) = \frac{\bar{\alpha}(A)(1-\bar{\alpha}(A))}{\alpha(\mathbb{R})+1}.$$

Proof. We have $(P(A), P(A^c)) \sim \mathcal{D}(\alpha(A), \alpha(A^c)) = Beta(\alpha(A), \alpha(A^c))$. Then,

$$\mathbb{E}[P(A)] = \frac{\alpha(A)}{\alpha(A)+\alpha(A^c)} = \bar{\alpha}(A),$$

$$\operatorname{var}(P(A)) = \frac{\alpha(A)(\alpha(\mathbb{R})-\alpha(A))}{(\alpha(A)+\alpha(A^c))^2(\alpha(A)+\alpha(A^c)+1)} = \frac{\bar{\alpha}(A)(1-\bar{\alpha}(A))}{\alpha(\mathbb{R})+1}.$$

□

The following theorem shows that the Dirichlet process has conjugacy properties.

Theorem 2.5. *For each $P \in \mathcal{M}(\mathbb{R})$, conditionally on P, let $X_1, \dots, X_n \overset{i.i.d}{\sim} P$. Let us assume that the prior Π on P is $\mathcal{D}_\alpha$, where α is a finite measure on $(\mathbb{R}, \mathcal{B})$. The posterior distribution of P given $X_1, \dots, X_n$ is $\mathcal{D}_{\alpha+\sum_{i=1}^n \delta_{X_i}}$.*

Proof. Let $(B_1, B_2, \dots, B_k)$ a partition of $\mathbb{R}$. Under Π,

$$(P(B_1), P(B_2), \dots, P(B_k)) \sim \mathcal{D}(\alpha(B_1), \alpha(B_2), \dots, \alpha(B_k)).$$

Therefore, introducing the discrete random variables $\tilde{X}_i = j \iff X_i \in B_j$, for any $i \in \{1, \dots, n\}$ and any $j \in \{1, \dots, k\}$,

$$\begin{aligned} ((P|X_1, \dots, X_n)(B_1), \dots, (P|X_1, \dots, X_n)(B_k)) &= (P(B_1), \dots, P(B_k)|X_1, \dots, X_n) \\ &= (P(B_1), \dots, P(B_k)|\tilde{X}_1, \dots, \tilde{X}_n) \\ &\sim \mathcal{D}\left(\Big(\alpha + \sum_{i=1}^n \delta_{X_i}\Big)(B_1), \dots, \Big(\alpha + \sum_{i=1}^n \delta_{X_i}\Big)(B_k)\right), \end{aligned}$$

where we have used the "discretization" of the set $\mathbb{R}$. These computations show that

$$P|X_1, \dots, X_n \sim \mathcal{D}_{\alpha+\sum_{i=1}^n \delta_{X_i}}.$$

□

Theorem 2.6. *For each $P \in \mathcal{M}(\mathbb{R})$, conditionally on P, let $X_1, \dots, X_n \overset{i.i.d}{\sim} P$. Let us assume that the prior Π on P is $\mathcal{D}_\alpha$, where α is a finite measure on $(\mathbb{R}, \mathcal{B})$. We have the following results.*

1. *The marginal distribution of X_1 is $\bar{\alpha}$.*

2. *The predictive distribution is as follows: for any $A \in \mathcal{B}$,*

$$P(X_{n+1} \in A | X_1, \dots, X_n) = \frac{\alpha(A) + \sum_{i=1}^n \delta_{X_i}(A)}{\alpha(\mathbb{R}) + n}.$$

3. *If α is absolutely continuous with respect to the Lebesgue measure, then*

$$P(X_2 = X_1) = \frac{1}{\alpha(\mathbb{R}) + 1}.$$

Proof. The first point is an easy consequence of Proposition 2.13 combined with the following result. For any $A \in \mathcal{B}$, with $\Pi = \mathcal{D}_\alpha$,

$$m(A) := \int P(A) \, \mathrm{d}\Pi(P) = \mathbb{E}[P(A)] = \bar{\alpha}(A).$$

For the second point, since $\Pi(\cdot | X_1, \dots, X_n) = \mathcal{D}_{\alpha + \sum_{i=1}^n \delta_{X_i}}$, we have:

$$\begin{aligned} P(X_{n+1} \in A | X_1, \dots, X_n) &= \mathbb{E}_{\Pi | X_1, \dots, X_n}[P(A)] \\ &= \frac{\alpha(A) + \sum_{i=1}^n \delta_{X_i}(A)}{\alpha(\mathbb{R}) + n}. \end{aligned}$$

For the third point, we have:

$$\begin{aligned} P(X_2 = X_1) &:= \mathbb{E}[P(X_2 = X_1 | X_1)] \\ &= \int_{\mathbb{R}} P(X_2 = x | X_1 = x) \, \mathrm{d}m(x) \\ &= \int \frac{\alpha(\{x\}) + \delta_x(x)}{\alpha(\mathbb{R}) + 1} \, \mathrm{d}\bar{\alpha}(x) \\ &= \frac{1}{\alpha(\mathbb{R}) + 1}. \end{aligned}$$

□

2.3.3 The stick-breaking representation

In this section, we provide a constructive way to build a Dirichlet process. As before, let α be a finite measure on $\mathbb{R}$ and $\bar{\alpha}(\cdot) = \alpha(\cdot)/\alpha(\mathbb{R})$. For this purpose, we define on $\mathbb{R}$,

$$\theta_1, \theta_2, \dots, \theta_n, \dots \overset{i.i.d}{\sim} Beta(1, \alpha(\mathbb{R}))$$

and, independently of the θ_i's,

$$Y_1, Y_2, \dots, Y_n, \dots \overset{i.i.d}{\sim} \bar{\alpha}.$$

Then, we set $p_1 = \theta_1$ and for any $n \geq 2$,

$$p_n = \theta_n \prod_{i=1}^{n-1} (1 - \theta_i).$$

We first prove the following lemma.

Lemma 2.1. *For any $n \in \mathbb{N}^*$,*

$$\sum_{j=1}^{n} p_j = 1 - \prod_{j=1}^{n} (1 - \theta_j).$$

We deduce

$$\sum_{j=1}^{+\infty} p_j = 1 \textit{ almost everywhere.}$$

Proof. The first point of the lemma is proved by induction.
- For $n = 1$, since $p_1 = 1 - (1 - \theta_1)$, the result if obvious.
- Let us assume that

$$\sum_{j=1}^{n} p_j = 1 - \prod_{j=1}^{n} (1 - \theta_j).$$

Then,

$$\begin{aligned}\sum_{j=1}^{n+1} p_j &= \sum_{j=1}^{n} p_j + p_{n+1} \\ &= 1 - \prod_{j=1}^{n} (1 - \theta_j) + \theta_{n+1} \prod_{j=1}^{n} (1 - \theta_j) \\ &= 1 - \prod_{j=1}^{n} (1 - \theta_j)(1 - \theta_{n+1}) \\ &= 1 - \prod_{j=1}^{n+1} (1 - \theta_j),\end{aligned}$$

which proves the first result for all $n \in \mathbb{N}^*$. In particular, we deduce that $n \mapsto \sum_{j=1}^{n} p_j$ is increasing and bounded by 1 almost everywhere. Then, this sequence converges when $n \to +\infty$ and

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{j=1}^{n} \log(1 - \theta_j) = \mathbb{E}[\log(1 - \theta_1)]] \text{ almost everywhere}$$

and since $\alpha(\mathbb{R}) > 0$,

$$\mathbb{E}[\log(1-\theta_1)]] = \frac{\Gamma(\alpha(\mathbb{R})+1)}{\Gamma(\alpha(\mathbb{R}))\Gamma(1)} \int_0^1 \log(1-x)(1-x)^{\alpha(\mathbb{R})-1}\,\mathrm{d}x < 0.$$

This yields

$$\lim_{n\to+\infty} \sum_{j=1}^{n} \log(1-\theta_j) = -\infty \text{ almost everywhere}$$

and

$$\lim_{n\to+\infty} \prod_{j=1}^{n} (1-\theta_j) = 0 \text{ almost everywhere.}$$

□

Finally, for any $A \in \mathcal{B}$, we set for any $w \in \mathbb{R}$,

$$(P(A))(w) = \sum_{n=1}^{+\infty} p_n(w)\delta_{Y_n(w)}(A).$$

The previous lemma shows that P is, almost surely, a (random) probability measure. It puts the weight p_n to the variable Y_n. We then obtain the following result due to Sethuraman (1994).

Theorem 2.7. *The process $w \longmapsto P(\cdot)(w)$ is distributed according to $\mathcal{D}_\alpha$. Its marginal distribution is $\bar{\alpha}$.*

Proof. The second point holds since the Y_n's are i.i.d. and distributed according to α and we have:

$$\begin{aligned}\mathbb{E}[P(A)] &= \mathbb{E}[\mathbb{E}[P(A)|p_1,\ldots,p_n,\ldots]] \\ &= \mathbb{E}\left[\sum_{n=1}^{+\infty} p_n \mathbb{E}[\delta_{Y_n}(A)|p_1,\ldots,p_n,\ldots]\right] \\ &= \mathbb{E}\left[\sum_{n=1}^{+\infty} p_n \bar{\alpha}(A)\right] = \bar{\alpha}(A),\end{aligned}$$

by using Lemma 2.1. For the first point, let $(B_1, B_2, \ldots, B_k)$ a partition of $\mathbb{R}$. We show that

$$(P(B_1), P(B_2), \ldots, P(B_k)) \sim \mathcal{D}(\alpha(B_1), \alpha(B_2), \ldots, \alpha(B_k)).$$

Let, for any $i \in \mathbb{N}^*$

$$U_{Y_i}^k := (\delta_{Y_i}(B_1), \ldots, \delta_{Y_i}(B_k)),$$

where, in the last vector, all coordinates are 0 except 1, the jth coordinate equal to 1 (where j is the only set of the partition such that $Y_i \in B_j$). We define

$$P_1 = p_1 U^k_{Y_1} + (1 - p_1)Q,$$

where Q is taken as follows: Q is independent from the Y_i's and the θ_i's and

$$Q \sim \mathcal{D}(\alpha(B_1), \ldots, \alpha(B_k)).$$

Using Proposition 2.10, conditionally on $Y_1 \in B_j$,

$$P_1 \sim \mathcal{D}(\alpha(B_1), \ldots, \alpha(B_j) + 1, \ldots, \alpha(B_k)),$$

where j is defined as before. Since $m(B_j) = P(Y_1 \in B_j) = \bar{\alpha}(B_j)$, by definition, for any Borelian set $C \in \mathcal{A}$, we have

$$\begin{aligned}
P(P_1 \in C) &= \sum_{j=1}^{k} P(P_1 \in C | Y_1 \in B_j) P(Y_1 \in B_j) \\
&= \sum_{j=1}^{k} \mathcal{D}(\alpha(B_1), \ldots, \alpha(B_j) + 1, \ldots, \alpha(B_k))(C)\bar{\alpha}(B_j) \\
&= \mathcal{D}(\alpha(B_1), \ldots, \alpha(B_k))(C),
\end{aligned}$$

using Equation (2.1). It means that

$$P_1 \sim \mathcal{D}(\alpha(B_1), \ldots, \alpha(B_k)). \tag{2.2}$$

Now, for some $N \in \mathbb{N}^*$, we assume that

$$\sum_{n=1}^{N} p_n U^k_{Y_n} + \left(1 - \sum_{n=1}^{N} p_n\right) Q \sim \mathcal{D}(\alpha(B_1), \ldots, \alpha(B_k)). \tag{2.3}$$

Since by Lemma 2.1,

$$p_{N+1} = \theta_{N+1} \prod_{n=1}^{N} (1 - \theta_n) = \theta_{N+1}\left(1 - \sum_{n=1}^{N} p_n\right),$$

we have

$$\begin{aligned}
\sum_{n=1}^{N+1} p_n U^k_{Y_n} + \left(1 - \sum_{n=1}^{N+1} p_n\right) Q &= \sum_{n=1}^{N} p_n U^k_{Y_n} + \left(1 - \sum_{n=1}^{N} p_n\right) Q + p_{N+1} U^k_{Y_{N+1}} - p_{N+1} Q \\
&= \sum_{n=1}^{N} p_n U^k_{Y_n} + \left(1 - \sum_{n=1}^{N} p_n\right) \big(Q + \theta_{N+1} U^k_{Y_{N+1}} - \theta_{N+1} Q\big) \\
&= \sum_{n=1}^{N} p_n U^k_{Y_n} + \left(1 - \sum_{n=1}^{N} p_n\right) \big(\theta_{N+1} U^k_{Y_{N+1}} + (1 - \theta_{N+1}) Q\big) \\
&\sim \sum_{n=1}^{N} p_n U^k_{Y_n} + \left(1 - \sum_{n=1}^{N} p_n\right) Q \\
&\sim \mathcal{D}(\alpha(B_1), \dots, \alpha(B_k)).
\end{aligned}$$

where we have used (2.2) and (2.3). Consequently, (2.3) is true for any $N \in \mathbb{N}^*$. By letting $N \to +\infty$, we conclude

$$\sum_{n=1}^{+\infty} p_n U^k_{Y_n} \sim \mathcal{D}(\alpha(B_1), \dots, \alpha(B_k)).$$

Since this is true for any partition $(B_1, B_2, \dots, B_k)$ of $\mathbb{R}$, we obtain

$$\sum_{n=1}^{+\infty} p_n \delta_{Y_n} \sim \mathcal{D}_\alpha.$$

□

The construction justifies the stick-breaking terminology. Starting with a stick of length 1, we break it at θ_1, p_1 is the length of the stick we just broke off. What remains has length $1 - \theta_1$. We break a θ_2-fraction of the remaining stick, that is $p_2 = \theta_2(1 - \theta_1)$. What is left after this step is $(1 - \theta_1) - \theta_2(1 - \theta_1) = (1 - \theta_1)(1 - \theta_2)$. At the kth step, we have a stick of length $\prod_{i=1}^{k-1}(1 - \theta_i)$ remaining and to produce p_k, we break off a θ_k-portion of it, so $p_k = \theta_k \prod_{i=1}^{k-1}(1 - \theta_i)$. The result is a sequence $(p_i)_{i=1,\dots,k}$.

The stick-breaking representation exhibits the Dirichlet distribution as a random discrete measure and we obtain the following corollary.

Corollary 2.1. *Almost every realization from $\mathcal{D}_\alpha$ is a discrete measure: If $\Pi = \mathcal{D}_\alpha$,*

$$\Pi(P : \ P \ \textit{is discrete}) = 1.$$

A realization from the Dirichlet process is discrete with probability one, also when the base measure α is absolute continuous. This is perhaps disappointing, especially if the intention is to model absolutely continuous probabilty measures.

2.3.4 Estimation of the cumulative distributive function by using Dirichlet processes

We wish to provide a Bayes estimate to infer the cumulative distributive function of an n-sample $X = (X_1, \dots, X_n)$ of real variables, with the loss-function defined, for any F and G two cumulative distributive functions, by

$$L(F, G) := \int_{\mathbb{R}} (F(t) - G(t)^2 \, \mathrm{d}t.$$

We denote F the cumulative distributive function of X and we have for any estimate $\hat{F}$ of F,

$$R(F, \hat{F}) := \int L(F, \hat{F}) \, \mathrm{d}P_F = \mathbb{E}_F[L(F, \hat{F})],$$

where P_F is the probability distribution associated with X and for any prior Π on F,

$$\begin{aligned} r(\pi, \hat{F}) &:= \int R(F, \hat{F}) \, \mathrm{d}\Pi(F) \\ &= \iiint (F(t) - \hat{F}(t))^2 \, \mathrm{d}t \, \mathrm{d}P_F \, \mathrm{d}\Pi(F) \\ &= \int \left(\iint (F(t) - \hat{F}(t))^2 \, \mathrm{d}m(x) \, \mathrm{d}\Pi(F|x) \right) \mathrm{d}t, \end{aligned}$$

where m denotes the marginal distribution of X. We can perform estimation at fixed t, which boils down to estimate the real $F(t)$. Therefore, using Theorem 1.2, the Bayes rule is for any $t \in \mathbb{R}$

$$\hat{F}_\Pi(t) = \int F(t) \, \mathrm{d}\Pi(F|X).$$

If $\Pi = \mathcal{D}_\alpha$, then

$$\Pi(\cdot|X) = \mathcal{D}_{\alpha + \sum_{i=1}^n \delta_{X_i}}$$

and since for any $t \in \mathbb{R}$,

$$F(t) = P((-\infty; t]),$$

we have

$$\begin{aligned} \hat{F}_\Pi(t) &= \int F(t) \, \mathrm{d}\Pi(F|X) \\ &= \mathbb{E}_{\mathcal{D}_{\alpha + \sum_{i=1}^n \delta_{X_i}}} [P((-\infty; t])] \\ &= \frac{\left(\alpha + \sum_{i=1}^n \delta_{X_i}\right)((-\infty; t])}{\alpha(\mathbb{R}) + n}. \end{aligned}$$

Observe that the Bayes rule can be written as a convex combination between the classical frequentist rule $F_n(t) := \sum_{i=1}^n \delta_{X_i}((-\infty;t]) = \frac{1}{n}\sum_{i=1}^n 1_{X_i \leq t\}}$ and the Bayes rule $\bar{\alpha}((-\infty;t])$:

$$\hat{F}_\Pi(t) = \frac{\alpha(\mathbb{R})}{\alpha(\mathbb{R})+n}\bar{\alpha}((-\infty;t]) + \frac{n}{\alpha(\mathbb{R})+n}F_n(t).$$

The weight $\frac{\alpha(\mathbb{R})}{\alpha(\mathbb{R})+n}$ goes to 0 when $n \to +\infty$ whereas the weight $\frac{n}{\alpha(\mathbb{R})+n}$ goes to 1. Dirichlet processes lead to natural interpretations.

Chapter 3

Asymptotic properties of nonparametric posterior distributions: consistency and convergence rates

Our goal is to extend results of Chapter 1 to the nonparametric setting. Before studying asymptotic properties of general nonparametric posterior distributions, let us first focus on the case of posterior Dirichlet processes. We rely on notions introduced and results proved in Section 1.2.1.

3.1 Consistency of posterior Dirichlet processes on $\mathbb{R}$

We first recall the definition of the weak convergence on $\mathcal{M}(\mathbb{R})$. The σ-algebra $\mathcal{A}$ is such that $P \in \mathcal{M}(\mathbb{R}) \mapsto P(B)$ is $\mathcal{A}$-measurable for any $B \in \mathcal{B}$.

Definition 3.1. *A sequence of probability measures $(\Pi_n)_n$ on $\mathcal{M}(\mathbb{R})$* ***converges weakly*** *to a probability measure Π if and only if for any $\mathbb{R}$-valued bounded continuous function ϕ on $\mathcal{M}(\mathbb{R})$*

$$\int_{\mathcal{M}(\mathbb{R})} \phi(P)\,\mathrm{d}\Pi_n(P) \overset{n\to+\infty}{\longrightarrow} \int_{\mathcal{M}(\mathbb{R})} \phi(P)\,\mathrm{d}\Pi(P).$$

Remark 3.1. *The continuity on $\mathcal{M}(\mathbb{R})$ is defined through the weak convergence on $\mathbb{R}$. For instance, if f is bounded and continuous, the function ϕ_f defined by $\phi_f : P \mapsto \int f\,\mathrm{d}P$ is continuous on $\mathcal{M}(\mathbb{R})$. Unfortunately, we cannot explicit all bounded and continuous functions of $\mathcal{M}(\mathbb{R})$.*

We now characterize the tightness on $\mathcal{M}(\mathbb{R})$ by using the Bolzano-Weierstrass theorem.

Theorem 3.1 (Bolzano-Weierstrass)**.** *A metrizable space X is compact if and only if each sequence in X has a convergent subsequence in X.*

For any probability measure Π on $\mathcal{M}(\mathbb{R})$, we set

$$\mathbb{E}_\Pi(B) := \int_{\mathcal{M}(\mathbb{R})} P(B)\,\mathrm{d}\Pi(P), \quad \forall B \in \mathcal{B}. \tag{3.1}$$

We then have the following result.

Theorem 3.2. *A sequence of probability measures $(\Pi_n)_n$ on $\mathcal{M}(\mathbb{R})$ is tight for the weak convergence of $\mathcal{M}(\mathbb{R})$ if and only if the sequence $(\mathbb{E}_{\Pi_n})_n$ is tight on $\mathbb{R}$.*

Proof. We first prove the sufficient condition, which is the crucial point for this course. For any n, we set $\mu_n := \mathbb{E}_{\Pi_n}$. We assume that $(\mu_n)_n$ is tight on $\mathbb{R}$. Let $\delta > 0$. We prove that there exists a compact set $M \subset \mathcal{M}(\mathbb{R})$ such that for any n, $\Pi_n(M) \geq 1 - \delta$. By tightness of $(\mu_n)_n$, for any $d \in \mathbb{N}^*$, there exists a compact set K_d such that for any n,

$$\mu_n(K_d) \geq 1 - \frac{6\delta}{d^3\pi^2}.$$

Now, we set

$$M_d := \left\{ P \in \mathcal{M}(\mathbb{R}) : \; P(K_d^c) \leq \frac{1}{d} \right\}$$

and

$$M = \cap_{d\in\mathbb{N}^*} M_d.$$

We show now that M is compact (for the weak convergence). Let $(P_n)_n \in M$. We show that $(P_n)_n$ has a convergent subsequence in M. Let $\varepsilon > 0$ and $d \in \mathbb{N}^*$ such that $\frac{1}{d} \leq \varepsilon$. Let n be fixed. Since $P_n \in M$, $P_n \in M_d$. Then

$$P_n(K_d^c) \leq \frac{1}{d} \leq \varepsilon \Rightarrow P_n(K_d) \geq 1 - \varepsilon.$$

This shows that $(P_n)_n$ is tight. Prohorov's theorem implies that there exists $(P_{n'})_{n'}$ a subsequence of $(P_n)_n$ such that $(P_{n'})_{n'}$ converges (since $\mathbb{R}$ is a Polish space). Therefore, there exists $P \in \mathcal{M}(\mathbb{R})$ such that $P_{n'} \overset{n'\to+\infty}{\rightsquigarrow} P$. Now, observe that for any $d \in \mathbb{N}^*$, K_d^c is an open set of $\mathbb{R}$. The Portmanteau theorem implies

$$P(K_d^c) \leq \liminf_{n'\to+\infty} P_{n'}(K_d^c) \leq \frac{1}{d}$$

and $P \in M$. We have proved that $(P_n)_n$ has a convergent subsequence in M. The Bolzano-Weierstrass theorem implies that M is compact. We now prove that for any n, $\Pi_n(M^c) \leq \delta$. For this purpose, we use

$$\Pi_n(M^c) \leq \sum_{d\in\mathbb{N}^*} \Pi_n(M_d^c)$$

and the Markov inequality:

$$\begin{aligned}
\Pi_n(M_d^c) &= \Pi_n\left(P \in \mathcal{M}(\mathbb{R}) : \ P(K_d^c) > \frac{1}{d}\right) \\
&= \int 1_{\{P\in\mathcal{M}(\mathbb{R}): \ P(K_d^c)>\frac{1}{d}\}} \, \mathrm{d}\Pi_n(P) \\
&\leq d \int_{\mathcal{M}(\mathbb{R})} P(K_d^c) \, \mathrm{d}\Pi_n(P) = d\mathbb{E}_{\Pi_n}[K_d^c] = d\mu_n(K_d^c) \\
&\leq \frac{6\delta}{d^2\pi^2}.
\end{aligned}$$

This gives $\Pi_n(M^c) \leq \delta$ and then $\Pi_n(M) \geq 1-\delta$. This yields the tightness of $(\Pi_n)_n$.

Now, assume that $(\Pi_n)_n$ is tight for the weak convergence of $\mathcal{M}(\mathbb{R})$. It means that for any $\varepsilon > 0$, there exists a compact set M for the weak convergence such that for any n, $\Pi_n(M) \geq 1 - \varepsilon/2$. So, M is a precompact set. Then, by Prohorov's theorem (see Remark 1.7), M is tight and there exists K, a compact set of $\mathbb{R}$, such that for any $P \in M$, $P(K) \geq 1 - \varepsilon/2$. Then, for all n,

$$\mathbb{E}_{\Pi_n}[K^c] = \int_M P(K^c)d\Pi_n(P) + \int_{M^c} P(K^c)d\Pi_n(P) \leq \sup_{P\in M} P(K^c) + \Pi_n(M^c) < \varepsilon.$$

We have proved that the sequence $(\mathbb{E}_{\Pi_n})_n$ is tight on $\mathbb{R}$. □

We now show the weak convergence of the posterior distribution.

Theorem 3.3. *Let $(\alpha_n)_n$ a sequence of finite measures on $\mathbb{R}$, such that $\alpha_n(\mathbb{R}) \overset{n\to+\infty}{\longrightarrow} +\infty$ and assume there exists a probability measure $\bar{\alpha}$ such that $\bar{\alpha}_n(\cdot) := \frac{\alpha_n(\cdot)}{\alpha_n(\mathbb{R})} \overset{n\to+\infty}{\rightsquigarrow} \bar{\alpha}$. If $\Pi_n = \mathcal{D}_{\alpha_n}$, then*

$$\Pi_n \overset{n\to+\infty}{\rightsquigarrow} \delta_{\bar{\alpha}}.$$

Proof. We show that $(\mathbb{E}_{\Pi_n})_n$ is tight on $(\mathbb{R}, \mathcal{B})$. For any $B \in \mathcal{B}$, $\mathbb{E}_{\Pi_n}(B) = \bar{\alpha}_n(B)$. Since $\bar{\alpha}_n$ converges weakly and since $\mathbb{R}$ is a Polish space, $(\mathbb{E}_{\Pi_n})_n$ is tight. Using Theorem 3.2, we have that $(\Pi_n)_n$ is tight. Now, if for any subsequence $(\Pi_{n'})_{n'}$ of $(\Pi_n)_n$, $(\Pi_{n'})_{n'} \overset{n\to+\infty}{\rightsquigarrow} \delta_{\bar{\alpha}}$, unicity of the limit will provide the result of the theorem. We essentially admit the result but observe that for any $B \in \mathcal{B}$ such that $\bar{\alpha}(\delta B) = 0$ and any $\varepsilon > 0$,

$$\begin{aligned}
\Pi_n(|P(B) - \bar{\alpha}(B)| > \varepsilon) &\leq \Pi_n(|P(B) - \mathbb{E}_{\Pi_n}(B)| > \varepsilon/2) + \Pi_n(|\mathbb{E}_{\Pi_n}(B) - \bar{\alpha}(B)| > \varepsilon/2) \\
&\leq \frac{4}{\varepsilon^2}\mathrm{var}_{\Pi_n}(P(B)) + \Pi_n(|\bar{\alpha}_n(B) - \bar{\alpha}(B)| > \varepsilon/2) \\
&\leq \frac{4}{\varepsilon^2}\frac{\bar{\alpha}_n(B)(1-\bar{\alpha}_n(B))}{\alpha_n(\mathbb{R})+1} + \Pi_n(|\bar{\alpha}_n(B) - \bar{\alpha}(B)| > \varepsilon/2).
\end{aligned}$$

Using assumptions of the theorem, since $\bar{\alpha}(\delta B) = 0$ $\bar{\alpha}_n(B) \overset{n\to+\infty}{\longrightarrow} \bar{\alpha}(B)$ and the second term vanishes for n large enough. Therefore, under Π_n, $P(B) \overset{n\to+\infty}{\longrightarrow} \bar{\alpha}(B)$ in probability and then in distribution. This explains why the limit is unique. See Theorem 2.5.2 of Ghosh and R.V. Ramamoorthi (2003) for more details. □

Corollary 3.1. *Let α a finite measure on $\mathbb{R}$. Then, for any $P_0 \in \mathcal{M}(\mathbb{R})$, the posterior distribution studied in the previous chapter is consistent at P_0:*

$$\mathcal{D}_{\alpha+\sum_{i=1}^n \delta_{X_i}} \overset{n\to+\infty}{\rightsquigarrow} \delta_{P_0}, \quad P_0 - a.e.$$

Proof. We apply Theorem 3.3 with

$$\alpha_n = \alpha + \sum_{i=1}^n \delta_{X_i}.$$

We have

$$\alpha_n(\mathbb{R}) = \alpha(\mathbb{R}) + n \overset{n\to+\infty}{\longrightarrow} +\infty$$

and

$$\bar{\alpha}_n = \frac{\alpha + \sum_{i=1}^n \delta_{X_i}}{\alpha(\mathbb{R}) + n}.$$

Let t such that F_0, the cumulative distributive function of P_0, is continuous at t, we have

$$\begin{aligned}\bar{\alpha}_n((-\infty;t]) &= \frac{1}{\alpha(\mathbb{R}) + n} \times \Big(\alpha((-\infty;t]) + \sum_{i=1}^n \delta_{X_i}((-\infty;t])\\ &= \frac{1}{\alpha(\mathbb{R}) + n} \times \alpha((-\infty;t]) + \frac{n}{\alpha(\mathbb{R}) + n} \times \frac{1}{n}\sum_{i=1}^n 1_{\{X_i \le t\}}\end{aligned}$$

and

$$\bar{\alpha}_n((-\infty;t]) \overset{n\to+\infty}{\longrightarrow} P_0((-\infty;t]), \quad P_0 - \text{a.e.}$$

Therefore,

$$\bar{\alpha}_n \overset{n\to+\infty}{\rightsquigarrow} P_0, \quad P_0 - \text{a.e.}$$

and, by Theorem 3.3,

$$\mathcal{D}_{\alpha+\sum_{i=1}^n \delta_{X_i}} \overset{n\to+\infty}{\rightsquigarrow} \delta_{P_0}, \quad P_0 - \text{a.e.}$$

□

3.2 Consistency for general nonparametric posterior distributions

Now, we consider a general statistical model denoted $(\Omega, \mathcal{B}, P \in \mathcal{T})$ and conditionally on P, $X_1, \ldots, X_n \overset{i.i.d}{\sim} P$. We denote Π a prior distribution on the general measured space $(\mathcal{T}, \mathcal{A})$. We first study strong consistency of the posterior distribution.

3.2.1 Strong consistency

Definition 3.2. *The posterior distribution $\Pi(\cdot|X_1,\dots,X_n)$ is said to be **strongly consistent (i.e. $\mathbb{L}_1$-consistent) at** P_0 if there exists Ω_0 with $P_0(\Omega_0) = 1$ such that $\forall w \in \Omega_0$, $\forall U$ an open set of $\mathcal{T}$ for the total variation norm containing P_0,*

$$\Pi(U|X_1(w),\dots,X_n(w)) \overset{n\to+\infty}{\longrightarrow} 1.$$

Remark 3.2. *Since $(\mathcal{T}, \|\cdot\|_{TV})$ is a metric space, for strong consistency, we can take U of the form $U = \{P: \ \|P - P_0\|_{TV} < \varepsilon\}$ for $\varepsilon > 0$.*

The following result shows that if the posterior distribution $\Pi(\cdot|X_1,\dots,X_n)$ is strongly consistent, then we can derive an estimate with nice properties.

Proposition 3.1. *Let $\hat{P}_n$ defined as*

$$\hat{P}_n(A) := \int P(A)\,\mathrm{d}\Pi(P|X_1,\dots,X_n), \quad \forall A \in \mathcal{B}.$$

If $\Pi(\cdot|X_1,\dots,X_n)$ is strongly consistent at P_0, then $\hat{P}_n$ satisfies

$$\|\hat{P}_n - P_0\|_{TV} \overset{n\to+\infty}{\longrightarrow} 0 \quad P_0 - a.e.$$

Proof. By assumption, there exists Ω_0 with $P_0(\Omega_0) = 1$ such that $\forall w \in \Omega_0$, $\forall \varepsilon > 0$, with $U_\varepsilon = \{P: \ \|P - P_0\|_{TV} < \varepsilon\}$,

$$\Pi(U_\varepsilon|X_1(w),\dots,X_n(w)) \overset{n\to+\infty}{\longrightarrow} 1.$$

Now, on Ω_0,

$$\begin{aligned}
\|\hat{P}_n - P_0\|_{TV} &= \Big\| \int (P - P_0)\,\mathrm{d}\Pi(P|X_1,\dots,X_n) \Big\|_{TV} \\
&\le \int \|P - P_0\|_{TV}\,\mathrm{d}\Pi(P|X_1,\dots,X_n) \\
&\le \int_{U_\varepsilon} \|P - P_0\|_{TV}\,\mathrm{d}\Pi(P|X_1,\dots,X_n) + \int_{U_\varepsilon^c} \|P - P_0\|_{TV}\,\mathrm{d}\Pi(P|X_1,\dots,X_n) \\
&\le \varepsilon + \Pi(U_\varepsilon^c|X_1,\dots,X_n),
\end{aligned}$$

since the TV-norm is bounded by 1. Finally, on Ω_0, $\forall \varepsilon > 0$,

$$\limsup_{\varepsilon\to 0} \|\hat{P}_n - P_0\|_{TV} \le \varepsilon$$

and

$$\|\hat{P}_n - P_0\|_{TV} \overset{n\to+\infty}{\longrightarrow} 0 \quad P_0 - \text{a.e.}$$

□

3.2.2 Weak consistency and Schwartz theorem

Most of the time, strong consistency is too demanding and we can only prove weak consistency.

Definition 3.3. *The posterior distribution* $\Pi(\cdot|X_1,\dots,X_n)$ *is said to be* ***weakly consistent at*** P_0 *if there exists* Ω_0 *with* $P_0(\Omega_0)=1$ *such that* $\forall w\in\Omega_0$, $\forall U$ *an open set of* $\mathcal{T}$ *for the weak convergence containing* P_0,

$$\Pi(U|X_1(w),\dots,X_n(w)) \overset{n\to+\infty}{\longrightarrow} 1.$$

Remember that it is not a good idea to consider consistency under the total variation norm. For this reason, we consider posterior consistency on densities and for μ a σ-finite measure on Ω, we set

$$\mathcal{L}_\mu := \left\{P\in\mathcal{T}:\ P\ll\mu\right\}.$$

If for $(P,Q)\in\mathcal{L}_\mu^2$, we set

$$f=\frac{\mathrm{d}P}{\mathrm{d}\mu},\quad g=\frac{\mathrm{d}Q}{\mathrm{d}\mu},$$

we have

$$\|f-g\|_1 = 2\|P-Q\|_{TV}.$$

In the sequel, without loss of generality, we assume that μ is the Lebesgue measure. Let U be a set containing $f_0=\frac{\mathrm{d}P_0}{\mathrm{d}\mu}$. To obtain the convergence of the posterior probability of U given $X_1,\dots,X_n$ to 1, f_0 and U^c need to be separated. This idea of separation is conveniently formalized through the existence of appropriate tests for testing $H_0: f=f_0$ versus $H_1: f\in U^c$, where f is the density of the X_i's, expressed by following results. We first introduce test functions with specific properties.

Definition 3.4. *Let* f_0 *be a density and* U *a set containing* f_0. *Assume we are given* $X^n=(X_1,\dots,X_n)$ *an n-sample. We denote* f *the density of the* X_i*'s.*
The test $\phi(X^n)$ *is* **strictly unbiased** *for testing* H_0*:* $f=f_0$ *versus* $H_1: f\in U^c$ *if*

$$\mathbb{E}_{f_0}[\phi(X^n)] < \inf_{f\in U^c}\mathbb{E}_f[\phi(X^n)].$$

Now let $(\phi_n(X^n))_n$ *be a sequence of test functions.*

- *The sequence* $(\phi_n(X^n))_n$ *is* **uniformly consistent** *for testing* H_0*:* $f=f_0$ *versus* $H_1: f\in U^c$ *if, as* $n\to+\infty$,

$$\mathbb{E}_{f_0}[\phi_n(X^n)]\to 0,\quad \inf_{f\in U^c}\mathbb{E}_f[\phi_n(X^n)]\to 1.$$

- *The sequence* $(\phi_n(X^n))_n$ *is* **uniformly exponentially consistent** *for testing* H_0*:* $f=f_0$ *versus* $H_1: f\in U^c$ *if there exist* C *and* β *positive constants such that for any* n,

$$\mathbb{E}_{f_0}[\phi_n(X^n)]\le Ce^{-n\beta},\quad \inf_{f\in U^c}\mathbb{E}_f[\phi_n(X^n)]\ge 1-Ce^{-n\beta}.$$

Remark 3.3. *In the previous definition, we consider randomized tests, namely test taking values in* $[0;1]$.

Proposition 3.2. *Let* f_0 *be a density and* U *a set containing* f_0. *Let* $X = (X_n)_{n\in\mathbb{N}^*}$ *a sequence of iid variables. We denote* f *the density of the* X_i*'s. The following facts are equivalent.*

(i) There exists a uniformly exponentially consistent sequence of tests for testing H_0*:* $f = f_0$ *versus* $H_1 : f \in U^c$.

(ii) There exists a uniformly consistent sequence of tests for testing H_0*:* $f = f_0$ *versus* $H_1 : f \in U^c$.

(iii) There exist $n \in \mathbb{N}^*$ *and a strictly unbiased test* $\phi(X^n)$ *for testing* H_0*:* $f = f_0$ *versus* $H_1 : f \in U^c$.

Proof. Of course (i)⇒(ii)⇒(iii). We now assume that there exist ϕ and $m \in \mathbb{N}^*$ such that

$$\alpha := \mathbb{E}_{f_0}[\phi(X^m)] < \beta := \inf_{f\in U^c} \mathbb{E}_f[\phi(X^m)].$$

First assume $m = 1$: Let

$$A_k = \left\{ \frac{1}{k}\sum_{i=1}^k \phi(X_i) > \frac{1}{2}(\alpha+\beta) \right\}.$$

Lemma 3.1 (Hoeffding)**.** *Let* $(Y_1,\dots,Y_n)$ *be a sequence of independent variables such that* $\mathbb{E}[Y_i] = 0$ *and for all* i, $a_i \le Y_i \le b_i$ *almost everywhere, then*

$$\forall\, \lambda > 0, \quad \mathbb{P}\left(\sum_{i=1}^n Y_i \ge \lambda\right) \le \exp\left(-\frac{2\lambda^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

By using the Hoeffding lemma, we have:

$$\begin{aligned}
\mathbb{P}_{f_0}(A_k) &= \mathbb{P}_{f_0}\left(\sum_{i=1}^k (\phi(X_i) - \mathbb{E}_{f_0}[\phi(X_i)]) > \frac{k}{2}(\beta-\alpha)\right) \\
&\le \exp\left(-\frac{2k^2(\beta-\alpha)^2}{4k}\right) \\
&\le \exp\left(-\frac{k(\beta-\alpha)^2}{2}\right)
\end{aligned}$$

and for $f \in U^c$

$$\begin{aligned}
\mathbb{P}_f(A_k) &\geq \mathbb{P}_f\left(\sum_{i=1}^{k}(\phi(X_i) - \mathbb{E}_f[\phi(X_i)]) > \frac{k}{2}(\alpha - \beta)\right) \\
&= 1 - \mathbb{P}_f\left(\sum_{i=1}^{k}(\phi(X_i) - \mathbb{E}_f[\phi(X_i)]) \leq \frac{k}{2}(\alpha - \beta)\right) \\
&= 1 - \mathbb{P}_f\left(\sum_{i=1}^{k}(\mathbb{E}_f[\phi(X_i)] - \phi(X_i)) \geq \frac{k}{2}(\beta - \alpha)\right) \\
&\geq 1 - \exp\left(-\frac{k(\beta-\alpha)^2}{2}\right).
\end{aligned}$$

So, $\phi_k(X^k) = 1_{A_k}$ provides a uniformly exponentially consistent sequence of tests for testing H_0: $f = f_0$ versus $H_1 : f \in U^c$.

Now, let $m \in \mathbb{N}^*$. We apply the previous construction with $\ell = km$ and we use $X^\ell = (X_1, \ldots, X_\ell)$. We have:

$$\begin{aligned}
I &:= \mathbb{P}_{f_0}\left(\sum_{i=1}^{k}\phi(X_{1+(i-1)m}, \ldots, X_{m+(i-1)m}) - \mathbb{E}_{f_0}[\phi(X_{1+(i-1)m}, \ldots, X_{m+(i-1)m})] > \frac{k}{2}(\beta - \alpha)\right) \\
&\leq \exp\left(-\frac{k(\beta-\alpha)^2}{2}\right) \\
&= \exp\left(-\frac{\ell(\beta-\alpha)^2}{2m}\right).
\end{aligned}$$

So, for n such that $km \leq n < (k+1)m$, we set

$$\phi_n(X^n) := 1_{A_\ell}$$

with

$$A_\ell = \left\{\sum_{i=1}^{k}\phi(X_{1+(i-1)m}, \ldots, X_{m+(i-1)m}) - \mathbb{E}_{f_0}[\phi(X_{1+(i-1)m}, \ldots, X_{m+(i-1)m}) > \frac{k}{2}(\beta - \alpha)\right\}.$$

With $C = \exp((\beta - \alpha)^2/2)$, we have:

$$\begin{aligned}
\mathbb{E}_{f_0}[\phi_n(X^n)] &\leq C\exp\left(-(\ell + m)(\beta - \alpha)^2/(2m)\right) \\
&\leq C\exp\left(-n(\beta - \alpha)^2/(2m)\right).
\end{aligned}$$

Similarly, we obtain for $f \in U^c$

$$\mathbb{P}_f(A_\ell) \geq 1 - C\exp\left(-n(\beta - \alpha)^2/(2m)\right).$$

So, $\phi_n(X^n) = 1_{A_\ell}$ provides a uniformly exponentially consistent sequence of tests for testing H_0: $f = f_0$ versus $H_1 : f \in U^c$. □

Weak consistency results are based on the Schwartz theorem that has an interest in its own right. We need following definitions.

Definition 3.5. *We set*

$$\mathbb{L}_1(\mu) := \left\{ f : \quad f \geq 0, \ \textit{measurable and} \ \int f(x)\,\mathrm{d}\mu(x) = 1 \right\}$$

and for any $f_0 \in \mathbb{L}_1(\mu)$, for any $\varepsilon > 0$,

$$K_\varepsilon(f_0) := \{ g \in \mathbb{L}_1(\mu) : \quad K(f_0, g) < \varepsilon \},$$

with

$$K(f_0, g) = \begin{cases} \int \log\left(\frac{f_0(x)}{g(x)}\right) f_0(x)\,\mathrm{d}\mu(x) & \textit{if } f_0\,\mathrm{d}\mu \ll g\,\mathrm{d}\mu, \\ +\infty & \textit{otherwise.} \end{cases}$$

Let $f_0 \in \mathbb{L}_1(\mu)$. If Π is a prior on $\mathbb{L}_1(\mu)$, f_0 ***is said to be in the support of*** *Π if for any $\varepsilon > 0$,*

$$\Pi(K_\varepsilon(f_0)) > 0.$$

Theorem 3.4 (Schwartz theorem (1965))**.** *Let Π a prior on $\mathbb{L}_1(\mu)$. Let $f_0 \in \mathbb{L}_1(\mu)$ such that f_0 is in the support of Π. We take $U \subset \mathbb{L}_1(\mu)$ such that there exists a strictly unbiased test for testing*

$$H_0 : f = f_0 \quad \textit{versus} \quad H_1 : f \in U^c.$$

Then,

$$\Pi(U | X_1, \ldots, X_n) \overset{n\to+\infty}{\longrightarrow} 1 \quad P_{f_0} - a.e.$$

Proof. Denoting $X^n = (X_1, \ldots, X_n)$, we write

$$\begin{aligned} \Pi(U^c | X^n) &= \frac{\int_{U^c} \prod_{i=1}^n f(X_i)\,\mathrm{d}\Pi(f)}{\int_{\mathbb{L}_1(\mu)} \prod_{i=1}^n f(X_i)\,\mathrm{d}\Pi(f)} \\ &= \frac{\int_{U^c} \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)}\,\mathrm{d}\Pi(f)}{\int_{\mathbb{L}_1(\mu)} \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)}\,\mathrm{d}\Pi(f)} =: \frac{N_n(X^n)}{D_n(X^n)}. \end{aligned}$$

We prove the following useful lemma.

Lemma 3.2. *There exists a constant $\rho > 0$ such that*

$$A_n := \{ N_n(X^n) \leq \exp(-\rho n) \}$$

satisfies $\sum_n \mathbb{P}_{f_0}(A_n^c) < \infty$.

Proof. Using Proposition 3.2, we know that for any n, there exist $\phi_n(X^n)$ a test and C and β two positive constants such that

$$\mathbb{E}_{f_0}[\phi_n(X^n)] \leq Ce^{-n\beta}, \quad \inf_{f\in U^c} \mathbb{E}_f[\phi_n(X^n)] \geq 1 - Ce^{-n\beta}.$$

In particular, for any $f \in U^c$,

$$\mathbb{E}_f[1-\phi_n(X^n)] \leq Ce^{-n\beta}.$$

We have

$$\begin{aligned}\mathbb{P}_{f_0}(A_n^c) &= \mathbb{E}_{f_0}[1_{A_n^c}\phi_n(X^n)] + \mathbb{E}_{f_0}[1_{A_n^c}(1-\phi_n(X^n))]\\ &\leq Ce^{-n\beta} + \mathbb{E}_{f_0}[1_{\{N_n(X^n)>\exp(-\rho n)\}}(1-\phi_n(X^n))].\end{aligned}$$

We bound the second term of the right hand side as follows.

$$\begin{aligned}\mathbb{E}_{f_0}[1_{\{N_n(X^n)>\exp(-\rho n)\}}(1-\phi_n(X^n))] &= \int (1-\phi_n(x^n))1_{\{N_n(x^n)>\exp(-\rho n)\}}\prod_{i=1}^n [f_0(x_i)\,\mathrm{d}x_i]\\ &= \int (1-\phi_n(x^n))N_n(x^n)\exp(\rho n)\prod_{i=1}^n [f_0(x_i)\,\mathrm{d}x_i]\\ &= \int (1-\phi_n(x^n))\int_{U^c}\mathrm{d}\Pi(f)\exp(\rho n)\prod_{i=1}^n [f(x_i)\,\mathrm{d}x_i]\\ &= \exp(\rho n)\int_{U^c}\mathbb{E}_f[1-\phi_n(X^n)]\,\mathrm{d}\Pi(f)\\ &\leq C\exp((\rho-\beta)n).\end{aligned}$$

By taking $0<\rho<\beta$, we have $\sum_n \mathbb{P}_{f_0}(A_n^c) < \infty$ and the lemma is proved. □

Combining the result of the previous lemma and Borel Cantelli Lemma, we obtain

$$\mathbb{P}_{f_0}\Big(\bigcap_{n_0\in\mathbb{N}^*}\bigcup_{n\geq n_0} A_n^c\Big) = 0$$

and with $\Omega_0 = \bigcup_{n_0\in\mathbb{N}^*}\bigcap_{n\geq n_0} A_n$, we have $\mathbb{P}_{f_0}(\Omega_0)=1$. Therefore, for any $w\in\Omega_0$, there exists $n_0(w)$ such that for any $n\geq n_0(w)$, $N_n(X^n(w))\leq \exp(-\rho n)$. Before dealing with the denominator $D_n(X^n)$, we prove the following lemma.

Lemma 3.3. *There exists $B\in\mathcal{B}$ such that $\mathbb{P}_{f_0}(B)=1$ and for any $w\in B$, there exists $G_w\in\mathcal{A}$ such that $\Pi(G_w)=1$ and for any $f\in G_w$,*

$$\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^n \log\left(\frac{f_0(X_i(w))}{f(X_i(w))}\right) = K(f_0,f).$$

Proof. We denote

$$k_n(w,f) := \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f_0(X_i(w))}{f(X_i(w))}\right)$$

and

$$G := \left\{(w,f):\ \lim_{n\to+\infty} k_n(w,f) = K(f_0,f)\right\}$$

and its sections

$$G_f = \{w:\ (w,f)\in G\}, \quad G_w = \{f:\ (w,f)\in G\}.$$

We admit that all these spaces are measurable. The strong law of large numbers implies that $\mathbb{P}_{f_0}(G_f) = 1$ for all $f \in \mathbb{L}_1(\mu)$. Then, Fubini's theorem gives

$$\begin{aligned} 1 :=& \int_{\mathbb{L}_1(\mu)} \mathbb{P}_{f_0}(G_f)\,\mathrm{d}\Pi(f) \\ =& \int_{\mathbb{L}_1(\mu)}\int_{\Omega} 1_G(w,f)\,\mathrm{d}\mathbb{P}_{f_0}(w)\,\mathrm{d}\Pi(f) \\ =& \int_{\Omega}\left[\int_{\mathbb{L}_1(\mu)} 1_G(w,f)\,\mathrm{d}\Pi(f)\right]\mathrm{d}\mathbb{P}_{f_0}(w) \\ =& \int_{\Omega}\Pi(G_w)\,\mathrm{d}\mathbb{P}_{f_0}(w). \end{aligned}$$

The previous equality implies that $\Pi(G_w) = 1$ $\mathbb{P}_{f_0}$-a..e. There exists $B \in \mathcal{B}$ such that $\mathbb{P}_{f_0}(B) = 1$ and for any $w \in B$, $\Pi(G_w) = 1$. □

Now, we deal with $D_n(X^n)$.

Lemma 3.4. *For any $\varepsilon > 0$, we have on B:*

$$\lim_{n\to+\infty} \exp(n\varepsilon)D_n(X^n) = +\infty.$$

Proof. For any $w \in B$,

$$\begin{aligned} D_n(X^n(w)) =& \int_{\mathbb{L}_1(\mu)}\prod_{i=1}^{n}\frac{f(X_i(w))}{f_0(X_i(w))}\,\mathrm{d}\Pi(f) \\ \geq& \int_{K_\varepsilon(f_0)\cap G_w}\exp\left(-\sum_{i=1}^{n}\log\left(\frac{f_0(X_i(w))}{f(X_i(w))}\right)\right)\mathrm{d}\Pi(f), \end{aligned}$$

with

$$K_\varepsilon(f_0) = \{g\in\mathbb{L}_1(\mu):\quad K(f_0,g) < \varepsilon\}.$$

Therefore, by Fatou's lemma,

$$\begin{aligned}\liminf_{n\to+\infty}\exp(2n\varepsilon)D_n(X^n(w)) &\geq \int_{K_\varepsilon(f_0)\cap G_w}\liminf_{n\to+\infty}\exp\left(2n\varepsilon-\sum_{i=1}^n\log\left(\frac{f_0(X_i(w))}{f(X_i(w))}\right)\right)\mathrm{d}\Pi(f)\\ &= +\infty,\end{aligned}$$

since on $K_\varepsilon(f_0)\cap G_w$,

$$\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^n\log\left(\frac{f_0(X_i(w))}{f(X_i(w))}\right)=K(f_0,f)\leq\varepsilon$$

and $\Pi(K_\varepsilon(f_0)\cap G_w)>0$. □

The previous lemma gives that for any $w\in B$, there exists $n_1(w)$ such that for any $n\geq n_1(w)$,

$$D_n(X^n(w))\geq\exp(-n\varepsilon).$$

This shows that for any $n\geq\max(n_0(w),n_1(w))$,

$$\Pi(U^c|X^n(w))=\frac{N_n(X^n(w))}{D_n(X^n(w))}\leq\exp((\varepsilon-\rho)n).$$

Taking $\varepsilon<\rho$, we obtain for $w\in B$,

$$\lim_{n\to+\infty}\Pi(U^c|X^n(w))=0.$$

Theorem 3.4 is proved. □

Theorem 3.5. *Let* Π *a prior on* $\mathbb{L}_1(\mu)$*. If* f_0 *is in the support of* Π*, then the posterior distribution is weakly consistent at* f_0*. It means that for any open set* U *for the topology of weak convergence containing* f_0*, we have:*

$$\Pi(U|X_1,\ldots,X_n)\overset{n\to+\infty}{\longrightarrow}1\quad P_{f_0}-a.e.$$

Proof. We can show that it is enough to take U as follows:

$$U=\bigcap_{i\in I}\left\{f:\ \left|\int h_i(x)f(x)\,\mathrm{d}x-\int h_i(x)f_0(x)\,\mathrm{d}x\right|<\varepsilon\right\},$$

where I is a finite set, $(h_i)_{i\in I}$ are continuous bounded functions and $\varepsilon>0$. Therefore, it is enough to show the result for

$$U(h,\varepsilon)=\left\{f:\ \left|\int h(x)f(x)\,\mathrm{d}x-\int h(x)f_0(x)\,\mathrm{d}x\right|<\varepsilon\right\},$$

with h bounded continuous and $\varepsilon > 0$. We set

$$\phi = \frac{h + \|h\|_\infty}{2\|h\|_\infty} \in [0;1].$$

We have

$$f \in U(h,\varepsilon) \iff \int h(x)(f(x) - f_0(x))\,\mathrm{d}x < \varepsilon \text{ and } \int h(x)(f(x) - f_0(x))\,\mathrm{d}x > -\varepsilon.$$

and we observe that

$$\int \phi(x)(f(x) - f_0(x))\,\mathrm{d}x = \int \frac{h(x)}{2\|h\|_\infty}(f(x) - f_0(x))\,\mathrm{d}x$$

$$\int (1-\phi(x))(f(x) - f_0(x))\,\mathrm{d}x = \int \frac{h(x)}{2\|h\|_\infty}(f_0(x) - f(x))\,\mathrm{d}x.$$

Then,

$$\begin{aligned} f \in U(h,\varepsilon) \iff & \int \phi(x)(f(x) - f_0(x))\,\mathrm{d}x < \frac{\varepsilon}{2\|h\|_\infty} \\ & \text{and } \int (1-\phi(x))(f(x) - f_0(x))\,\mathrm{d}x < \frac{\varepsilon}{2\|h\|_\infty}. \end{aligned}$$

So, it is enough to show that for any ϕ bounded, continuous and taking values in $[0;1]$ and any $\varepsilon > 0$,

$$\Pi(V(\phi,\varepsilon)|X_1,\dots,X_n) \overset{n\to+\infty}{\longrightarrow} 1 \quad \mathbb{P}_{f_0} - \text{a.e.}$$

with

$$V(\phi,\varepsilon) = \left\{ f : \int \phi(x)(f(x) - f_0(x))\,\mathrm{d}x < \varepsilon \right\}.$$

Let ϕ a bounded and continuous function such that ϕ takes values in $[0;1]$. We build a strictly unbiased test for testing H_0: $f = f_0$ versus $H_1 : f \in V(\phi,\varepsilon)^c$. For this purpose, we consider $\phi(X_1)$. If $f \in V(\phi,\varepsilon)^c$, we have

$$\begin{aligned} \mathbb{E}_f[\phi(X_1)] &= \int \phi(x) f(x)\,\mathrm{d}x \\ &\geq \int \phi(x) f_0(x)\,\mathrm{d}x + \varepsilon \\ &\geq \mathbb{E}_{f_0}[\phi(X_1)] + \varepsilon. \end{aligned}$$

This shows that

$$\mathbb{E}_{f_0}[\phi(X_1)] < \inf_{f \in V(\phi,\varepsilon)^c} \mathbb{E}_f[\phi(X_1)]$$

and $\phi(X_1)$ is strictly unbiased for testing H_0: $f = f_0$ versus $H_1 : f \in V(\phi,\varepsilon)^c$. Theorem 3.4 provides the conclusion of the proof. □

3.3 Posterior contraction rates of convergence

3.3.1 General result

Let $X^n = (X_1, \dots, X^n) \overset{i.i.d.}{\sim} P_0$. We consider $(\Pi_n)_n$ a sequence of prior probability measures supported on a set $\mathcal{P} \subset \mathbb{L}_1(\mu)$, where, here, $\mathbb{L}_1(\mu)$ is the set of all probability measures absolutely continuous with respect to μ. We denote d either the Hellinger distance or the total variation metric on $\mathcal{P}$. We denote

$$f_0 = \frac{\mathrm{d}P_0}{\mathrm{d}\mu}, \quad f = \frac{\mathrm{d}P}{\mathrm{d}\mu},$$

for any $P \in \mathbb{L}_1(\mu)$. At some places, we refer to P (respectively P_0) and at others to f (respectively f_0), which is equivalent. We introduce $D(\varepsilon, \mathcal{P}, d)$ the ε-packing number of $\mathcal{P}$:

Definition 3.6. *For any $\varepsilon > 0$, $D(\varepsilon, \mathcal{P}, d)$ is the maximal number of points in $\mathcal{P}$ such that the distance between every pair of points is at least ε.*

We have the following result due to Ghosal, Ghosh and van der Vaart (2000) and Ghosal and van der Vaart (2007).

Theorem 3.6. *Suppose that for a sequence $(\varepsilon_n)_n$ with $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to +\infty$, for a constant $C > 0$ and sets $\mathcal{P}_n \subset \mathcal{P}$, we have for n large enough:*

(a) $\log D(\varepsilon_n, \mathcal{P}_n, d) \leq n\varepsilon_n^2$,

(b) $\Pi_n(\mathcal{P} \setminus \mathcal{P}_n) \leq \exp(-(C+3)n\varepsilon_n^2)$,

(c) $\Pi_n(P : \; K(f_0, f) \leq \varepsilon_n^2, \; V(f_0, f) \leq \varepsilon_n^2) \geq \exp(-Cn\varepsilon_n^2)$,

where

$$K(f_0, f) = \begin{cases} \int \log\left(\frac{f_0(x)}{f(x)}\right) f_0(x)\,\mathrm{d}\mu(x) & \textit{if } f_0\,\mathrm{d}\mu \ll f\,\mathrm{d}\mu, \\ +\infty & \textit{otherwise,} \end{cases}$$

$$V(f_0, f) = \begin{cases} \int \log^2\left(\frac{f_0(x)}{f(x)}\right) f_0(x)\,\mathrm{d}\mu(x) & \textit{if } f_0\,\mathrm{d}\mu \ll f\,\mathrm{d}\mu, \\ +\infty & \textit{otherwise.} \end{cases}$$

Then, for a constant M large enough, we have:

$$\Pi_n(P : \; d(P, P_0) \geq M\varepsilon_n | X^n) \overset{n \to +\infty}{\longrightarrow} 0 \quad \textit{in } P_0 - \textit{probability.}$$

We say that $(\varepsilon_n)_n$ ***is the posterior contraction rate of convergence (or posterior concentration rate of convergence) of*** *$\Pi_n(\cdot | X^n)$* ***on*** *$\mathcal{P}$* ***associated with*** *d.*

Proof. We prove the following lemma.

Lemma 3.5. *Assume that d is either the Hellinger distance or the total variation metric. If for any n, $\log D(\varepsilon_n, \mathcal{P}_n, d) \leq n\varepsilon_n^2$, then there exists an absolute constant $K > 0$ such that for any M a positive constant larger than 2, there exists $(\phi_n(X^n))_n$ a sequence of tests such that for any n,*

$$\mathbb{E}_{f_0}[\phi_n(X^n)] \leq \exp((1 - KM^2)n\varepsilon_n^2)$$

$$\sup_{P \in \mathcal{P}_n:\ d(P,P_0) > M\varepsilon_n} \mathbb{E}_f[1 - \phi_n(X^n)] \leq \exp(-KnM^2\varepsilon_n^2).$$

Proof. We admit the following lemma (see Birgé (1983)).

Lemma 3.6. *We assume that d is either the Hellinger distance or the total variation distance. Then, for any probability measures P_0 and P_1, there exists a test ψ_n such that, for K a universal constant,*

$$\mathbb{E}_{f_0}[\psi_n(X^n)] \leq \exp(-Knd^2(P_0, P_1)), \tag{3.2}$$

$$\sup_{P:\ d(P,P_1) < d(P_0,P_1)/2} \mathbb{E}_f[1 - \psi_n(X^n)] \leq \exp(-Knd^2(P_0, P_1)). \tag{3.3}$$

Let $M \geq 2$. We set $\mathcal{S}_M = \{P \in \mathcal{P}_n :\ d(P_0, P) > M\varepsilon_n\}$. We build a maximal net $\mathcal{N}_M \subset \mathcal{S}_M$ such that the distance between any probability distributions of $\mathcal{N}_M$, Q_M and Q'_M with $Q_M \neq Q'_M$, satisfies $d(Q_M, Q'_M) \geq M\varepsilon_n/2$. Since $M \geq 2$, $d(Q_M, Q'_M) \geq \varepsilon_n$ and then $|\mathcal{N}_M|$, the cardinal of $\mathcal{N}_M$, is smaller than $D(\varepsilon_n, \mathcal{P}_n, d)$ and then smaller than $\exp(n\varepsilon_n^2)$. For any $P \in \mathcal{S}_M$, there exists $Q_M \in \mathcal{N}_M$ such that $d(P, Q_M) \leq M\varepsilon_n/2$ (otherwise $\mathcal{S}_M$ is not a maximal net) and

$$d(P, Q_M) \leq M\varepsilon_n/2 < d(P_0, Q_M)/2.$$

We apply Lemma 3.6 with P_0 and all $Q_M \in \mathcal{N}_M$ and we denote ψ_{n,Q_M} the associated tests satisfying (3.2) and (3.3). We set

$$\phi_n(X^n) = \max_{Q_M \in \mathcal{N}_M} \psi_{n,Q_M}(X^n).$$

We have

$$\begin{aligned}
\mathbb{E}_{f_0}[\phi_n(X^n)] &\leq \sum_{Q_M \in \mathcal{N}_M} \mathbb{E}_{f_0}[\psi_{n,Q_M}(X^n)] \\
&\leq \exp(n\varepsilon_n^2)\exp(-Knd^2(P_0, Q_M)) \\
&\leq \exp((1 - KM^2)n\varepsilon_n^2)
\end{aligned}$$

and

$$\begin{aligned}\sup_{P\in\mathcal{P}_n:\ d(P,P_0)>M\varepsilon_n} \mathbb{E}_f[1-\phi_n(X^n)] &= \sup_{P\in\mathcal{S}_M} \mathbb{E}_f[1-\phi_n(X^n)]\\ &\le \sup_{P\in\mathcal{S}_M}\min_{Q_M\in\mathcal{N}_M} \mathbb{E}_f[1-\psi_{n,Q_M}(X^n)]\\ &\le \min_{Q_M\in\mathcal{N}_M} \exp(-Knd^2(P_0,Q_M))\\ &\le \exp(-KnM^2\varepsilon_n^2).\end{aligned}$$

□

Let $\alpha>0$. We have

$$P_0(\Pi_n(P:\ d(P,P_0)\ge M\varepsilon_n|X^n)\ge\alpha)\le\alpha^{-1}\mathbb{E}_{f_0}[\Pi_n(P:\ d(P,P_0)\ge M\varepsilon_n|X^n)].$$

We denote

$$B_n:=\left\{f:\ K(f_0,f)\le\varepsilon_n^2, V(f_0,f)\le\varepsilon_n^2\right\}$$

and

$$A_n:=\left\{\int\prod_{i=1}^n\frac{f(X_i)}{f_0(X_i)}\,\mathrm{d}\Pi_{B_n}(f)>\exp(-2n\varepsilon_n^2)\right\}$$

with

$$\mathrm{d}\Pi_{B_n}(f)=\frac{1_{B_n}(f)\,\mathrm{d}\Pi_n(f)}{\Pi_n(B_n)}.$$

We prove the following lemma.

Lemma 3.7. *We have*

$$\lim_{n\to+\infty}P_0(A_n)=1.$$

Proof. We consider A_n^c. Observe that

$$\int\prod_{i=1}^n\frac{f(X_i)}{f_0(X_i)}\,\mathrm{d}\Pi_{B_n}(f)\le\exp(-2n\varepsilon_n^2)\iff\log\left(\int\prod_{i=1}^n\frac{f(X_i)}{f_0(X_i)}\,\mathrm{d}\Pi_{B_n}(f)\right)\le-2n\varepsilon_n^2.$$

But,

$$\begin{aligned}\log\left(\int\prod_{i=1}^n\frac{f(X_i)}{f_0(X_i)}\,\mathrm{d}\Pi_{B_n}(f)\right)&\ge\int\log\left(\prod_{i=1}^n\frac{f(X_i)}{f_0(X_i)}\right)\mathrm{d}\Pi_{B_n}(f)\\ &=\sum_{i=1}^n\int\log\left(\frac{f(X_i)}{f_0(X_i)}\right)\mathrm{d}\Pi_{B_n}(f)\end{aligned}$$

and the right hand side is less that $-2n\varepsilon_n^2$ if and only if

$$\begin{aligned}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\int\log\left(\frac{f(X_i)}{f_0(X_i)}\right)\mathrm{d}\Pi_{B_n}(f)-\mathbb{E}_{f_0}\left[\int\log\left(\frac{f(X_i)}{f_0(X_i)}\right)\mathrm{d}\Pi_{B_n}(f)\right]\right)&\\ =\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\int\log\left(\frac{f(X_i)}{f_0(X_i)}\right)\mathrm{d}\Pi_{B_n}(f)+\int K(f_0,f)\,\mathrm{d}\Pi_{B_n}(f)\right)&\\ \leq -2\sqrt{n}\varepsilon_n^2+\sqrt{n}\int K(f_0,f)\,\mathrm{d}\Pi_{B_n}(f)&\end{aligned}$$

and the last term is smaller than $-\sqrt{n}\varepsilon_n^2$. We obtain that, by denoting

$$U_n=\frac{1}{n}\sum_{i=1}^{n}\int\log\left(\frac{f(X_i)}{f_0(X_i)}\right)\mathrm{d}\Pi_{B_n}(f)$$

and by using the Bienaymé-Chebyshev inequality,

$$\begin{aligned}P_0(A_n^c)&\leq P_0\left(\sqrt{n}\,|U_n-\mathbb{E}_{f_0}[U_n]|\geq\sqrt{n}\varepsilon_n^2\right)\\ &\leq\frac{\mathrm{var}_{f_0}(\sqrt{n}U_n)}{n\varepsilon_n^4}\\ &\leq\frac{\mathrm{var}_{f_0}\left(\int\log\left(\frac{f(X_1)}{f_0(X_1)}\right)\mathrm{d}\Pi_{B_n}(f)\right)}{n\varepsilon_n^4}\\ &\leq\frac{\mathbb{E}_{f_0}\left[\left(\int\log\left(\frac{f(X_1)}{f_0(X_1)}\right)\mathrm{d}\Pi_{B_n}(f)\right)^2\right]}{n\varepsilon_n^4}\\ &\leq\frac{\mathbb{E}_{f_0}\left[\int\log^2\left(\frac{f(X_1)}{f_0(X_1)}\right)\mathrm{d}\Pi_{B_n}(f)\right]}{n\varepsilon_n^4}=\frac{\int V(f_0,f)1_{B_n}(f)\,\mathrm{d}\Pi_n(f)}{n\varepsilon_n^4\Pi_n(B_n)}\leq\frac{1}{n\varepsilon_n^2}\end{aligned}$$

and we obtain that

$$\lim_{n\to+\infty}P_0(A_n^c)=0,$$

which proves the lemma. □

We denote

$$V_n=\{P:\ d(P,P_0)\geq M\varepsilon_n\}$$

and we prove that

$$\mathbb{E}_{f_0}[\Pi_n(P:\ d(P,P_0)\geq M\varepsilon_n|X^n)]=\mathbb{E}_{f_0}[\Pi_n(V_n|X^n)]\to 0.$$

We have

$$\begin{aligned}\mathbb{E}_{f_0}[\Pi_n(V_n|X^n)] &= \mathbb{E}_{f_0}[1_{A_n}\Pi_n(V_n|X^n)] + \mathbb{E}_{f_0}[1_{A_n^c}\Pi_n(V_n|X^n)] \\ &\leq \mathbb{E}_{f_0}[1_{A_n}\Pi_n(V_n|X^n)\phi_n(X^n)] + \mathbb{E}_{f_0}[1_{A_n}\Pi_n(V_n|X^n)(1-\phi_n(X^n))] + P_0(A_n^c) \\ &\leq \mathbb{E}_{f_0}[\phi_n(X^n)] + \mathbb{E}_{f_0}[1_{A_n}\Pi_n(V_n|X^n)(1-\phi_n(X^n))] + P_0(A_n^c). \qquad (3.4)\end{aligned}$$

By using Lemma 3.5, for n large enough,

$$\begin{aligned}\mathbb{E}_{f_0}[\phi_n(X^n)] &\leq \exp(n\varepsilon_n^2) \times \frac{\exp(-KnM^2\varepsilon_n^2)}{1-\exp(-Kn\varepsilon_n^2)} \\ &\leq 2\exp(-(KM^2-1)n\varepsilon_n^2) \\ &\leq 2\exp(-Kn\varepsilon_n^2),\end{aligned}$$

for $M > \sqrt{\frac{K+1}{K}}$. Now,

$$\begin{aligned}&\mathbb{E}_{f_0}[1_{A_n}\Pi_n(V_n|X^n)(1-\phi_n(X^n))] \\ &\quad= \mathbb{E}_{f_0}\left[1_{A_n}(1-\phi_n(X^n))\frac{\int 1_{\{P:\ d(P,P_0)>M\varepsilon_n\}}\prod_{i=1}^n f(X_i)\,\mathrm{d}\Pi_n(f)}{\int\prod_{i=1}^n f(X_i)\,\mathrm{d}\Pi_n(f)}\right] \\ &\quad= \mathbb{E}_{f_0}\left[1_{A_n}(1-\phi_n(X^n))\frac{\int 1_{\{P:\ d(P,P_0)>M\varepsilon_n\}}\prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)}\,\mathrm{d}\Pi_n(f)}{\int\prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)}\,\mathrm{d}\Pi_n(f)}\right] \\ &\quad= \mathbb{E}_{f_0}\left[1_{A_n}(1-\phi_n(X^n))\frac{N_n(X^n)}{D^n(X^n)}\right],\end{aligned}$$

with

$$N_n(X^n) = \int_{\mathcal{P}} 1_{\{P:\ d(P,P_0)>M\varepsilon_n\}}\prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)}\,\mathrm{d}\Pi_n(f), \quad D_n(X^n) = \int_{\mathcal{P}}\prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)}\,\mathrm{d}\Pi_n(f).$$

We lower bound $D_n(X^n)$ on A_n:

$$\begin{aligned}D_n(X^n) &\geq \Pi_n(B_n)\int\prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)}\frac{1_{B_n}(f)\,\mathrm{d}\Pi_n(f)}{\Pi_n(B_n)} \\ &= \Pi_n(B_n)\int\prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)}\,\mathrm{d}\Pi_{B_n}(f) \\ &\geq \Pi_n(B_n)\exp(-2n\varepsilon_n^2).\end{aligned}$$

We finally upper bound

$$\tilde{N}_n := \mathbb{E}_{f_0}\left[1_{A_n}(1-\phi_n(X^n))N_n(X^n)\right].$$

We have

$$\begin{aligned}
\tilde{N}_n &= \mathbb{E}_{f_0}\left[1_{A_n}(1-\phi_n(X^n))\int 1_{\{P:\ d(P,P_0)>M\varepsilon_n\}}\prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)}\,\mathrm{d}\Pi_n(f)\right]\\
&\leq \mathbb{E}_{f_0}\left[\int_{\mathcal{P}\backslash\mathcal{P}_n}\prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)}\,\mathrm{d}\Pi_n(f)\right]\\
&\qquad\qquad + \mathbb{E}_{f_0}\left[1_{A_n}(1-\phi_n(X^n))\int_{\mathcal{P}_n} 1_{\{P:\ d(P,P_0)>M\varepsilon_n\}}\prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)}\,\mathrm{d}\Pi_n(f)\right]\\
&\leq \int_{\mathcal{P}\backslash\mathcal{P}_n}\mathrm{d}\Pi_n(f) + \int_{\mathcal{P}_n}\mathbb{E}_f\left[(1-\phi_n(X^n))\right]1_{\{P:\ d(P,P_0)>M\varepsilon_n\}}\,\mathrm{d}\Pi_n(f)\\
&\leq \Pi_n(\mathcal{P}\setminus\mathcal{P}_n) + \exp(-KnM^2\varepsilon_n^2)\\
&\leq 2\exp(-n\varepsilon_n^2(C+3)),
\end{aligned}$$

if $M > \sqrt{\frac{C+3}{K}}$. Finally,

$$\begin{aligned}
\mathbb{E}_{f_0}[1_{A_n}\Pi_n(V_n|X^n)(1-\phi_n(X^n))] &\leq 2\exp(-n\varepsilon_n^2(C+3)) \times \Big(\Pi_n(B_n)\exp(-2n\varepsilon_n^2)\Big)^{-1}\\
&\leq 2\exp(-n\varepsilon_n^2).
\end{aligned}$$

Each term of (3.4) goes to 0 when n goes to $+\infty$. This ends the proof of Theorem 3.6 □

Remark 3.4. *Conditions (a) and (c) are the essential ones. Condition (a) requires that the model $\mathcal{P}_n$ is not too big. It ensures the existence of certain tests and can be replaced by a testing condition (see Lemma 3.5). Condition (c) requires that the prior measures put a sufficient amount of mass near the true measure P_0. Here "near" means that the closeness is measured through a combination of the Kullback-Leibler divergence of f_0 and f and the $\mathbb{L}_2(f_0)$-norm of* $\log\left(\frac{f_0(\cdot)}{f(\cdot)}\right)$. *Condition (b) says that $\mathcal{P}_n$ is almost the support of the prior.*

Remark 3.5. *The proof of Theorem 3.6 shows that for condition (b), we can replace the term $C+3$ by any constant larger than $C+2$.*

Example 3.1. *Suppose that $\mathcal{P}$ consists of all measures with densities whose square-root $\sqrt{f}$ belongs to a ball of the Hölder class $\mathcal{C}^\alpha[0;1]$ for some $\alpha>0$. It is well-known that for all $\varepsilon>0$, $\log D(\varepsilon,\mathcal{P},d)\approx\varepsilon^{-1/\alpha}$. We take ε_n proportional to $n^{-\alpha/(1+2\alpha)}$. Therefore*

$$n\varepsilon_n^2 \approx n^{1/(1+2\alpha)} \approx \varepsilon_n^{-1/\alpha}.$$

Then (a) is satisfied since

$$\log D(\varepsilon_n,\mathcal{P},d) \leq n\varepsilon_n^2.$$

By taking $\mathcal{P}_n=\mathcal{P}$, (b) is satisfied. It remains to choose Π_n so that (c) is satisfied. Observe that $\varepsilon_n = n^{-\alpha/(1+2\alpha)}$ is the optimal rate of convergence on $\mathcal{C}^\alpha[0;1]$.

3.3.2 Applications: rates of convergence for frequentist estimators

We now consider $\hat{P}_n$, the posterior mean estimator of Proposition 3.1, defined as

$$\hat{P}_n(A) := \int P(A)\, \mathrm{d}\Pi_n(P|X_1,\dots,X_n), \quad \forall A \in \mathcal{B},$$

that achieves the following rate of convergence.

Theorem 3.7. *Under assumptions of Theorem 3.6,*

$$H^2(\hat{P}_n, P_0) \leq M^2\varepsilon_n^2 + \Pi_n(P:\ H(P,P_0) \geq M\varepsilon_n|X^n)$$

and

$$\|\hat{P}_n - P_0\|_{TV} \leq M\varepsilon_n + \Pi_n(P:\ H(P,P_0) \geq M\varepsilon_n|X^n).$$

Proof. We recall that Hellinger and total variation metrics are bounded metrics with bound 1. Since for any probability measures P and Q,

$$H^2(P,Q) = 2 - 2\int \sqrt{\frac{\mathrm{d}P}{\mathrm{d}\mu}\frac{\mathrm{d}Q}{\mathrm{d}\mu}}\,\mathrm{d}\mu$$

and the function $x \longmapsto \sqrt{x}$ is concave on $\mathbb{R}_+$, the function $P \longmapsto H^2(P,Q)$ is convex and, with

$$B_{M\varepsilon_n} = \{P:\ H(P,P_0) \leq M\varepsilon_n\},$$

we obtain

$$\begin{aligned}
H^2(\hat{P}_n, P_0) &= H^2\left(\int P(\cdot)\,\mathrm{d}\Pi_n(P|X_1,\dots,X_n), P_0\right)\\
&\leq \int H^2(P,P_0)\,\mathrm{d}\Pi_n(P|X_1,\dots,X_n)\\
&\leq \int_{B_{M\varepsilon_n}} H^2(P,P_0)\,\mathrm{d}\Pi_n(P|X_1,\dots,X_n) + \int_{B^c_{M\varepsilon_n}} H^2(P,P_0)\,\mathrm{d}\Pi_n(P|X_1,\dots,X_n)\\
&\leq M^2\varepsilon_n^2 + \Pi_n(P:\ H(P,P_0) \geq M\varepsilon_n|X^n).
\end{aligned}$$

We use similar arguments for the total variation metric combined with the convexity of $P \longmapsto \|P - Q\|_{TV}$. □

Thus the rate of posterior contraction is transferred to the posterior mean provided the posterior probability of the complement of the $M\varepsilon_n$-ball around P_0 converges to zero sufficiently fast. This is usually the case; in fact, the contraction of the posterior is typically exponentially fast (see the proof of Theorem 3.6).

However, we can construct an estimate which achieves the rate ε_n without previous or further assumptions, still by using d being either the Hellinger or the total variation metric.

Theorem 3.8. *We consider assumptions of Theorem 3.6. Let $(\delta_n)_{n\in\mathbb{N}^*}$ a sequence of positive real numbers such that $\delta_n \to 0$ when $n \to +\infty$. We set for any $P \in \mathcal{P}$ and any $r > 0$,*

$$B(P, r) = \{Q \in \mathcal{P} : \ d(P, Q) \leq r\}.$$

Then, for any $n \in \mathbb{N}^*$, we set $\hat{P}_n$ such that

$$\hat{r}_n(\hat{P}_n) \leq \inf_{P\in\mathcal{P}} \hat{r}_n(P) + \delta_n$$

with

$$\hat{r}_n(P) := \inf\left\{r > 0 : \quad \Pi_n(B(P, r)|X_1, \ldots, X_n) \geq \frac{1}{2}\right\}.$$

Then,

$$P_0\Big(d(\hat{P}_n, P_0) \leq 2M\varepsilon_n + \delta_n\Big) \overset{n\to+\infty}{\longrightarrow} 1.$$

Remark 3.6. *We introduce $(\delta_n)_{n\in\mathbb{N}^*}$ since $\inf_{P\in\mathcal{P}} \hat{r}_n(P)$ may be non achieved. Of course, we can take $\delta_n = M\varepsilon_n$.*

Proof. Let n and P be fixed. Using the definition of $\hat{r}_n(P)$, there exists a decreasing sequence $(r_q)_{q\in\mathbb{N}^*}$ such that $\lim_{q\to+\infty} r_q = \hat{r}_n(P)$ and $\Pi_n(B(P, r_q)|X_1, \ldots, X_n) \geq \frac{1}{2}$. Since

$$B(P, \hat{r}_n(P)) = \bigcap_{q\in\mathbb{N}^*} B(P, r_q),$$

then

$$\Pi_n(B(P, \hat{r}_n(P))|X_1, \ldots, X_n) = \lim_{q\to+\infty} \Pi_n(B(P, r_q)|X_1, \ldots, X_n) \geq \frac{1}{2}. \tag{3.5}$$

We now consider $\hat{P}_n$. We have

$$\hat{r}_n(\hat{P}_n) \leq \inf_{P\in\mathcal{P}} \hat{r}_n(P) + \delta_n \leq \hat{r}_n(P_0) + \delta_n.$$

But Theorem 3.6 gives

$$P_0\Big(\Pi_n(B(P_0, M\varepsilon_n)|X_1, \ldots, X_n) \overset{n\to+\infty}{\longrightarrow} 1) \overset{n\to+\infty}{\longrightarrow} 1, \tag{3.6}$$

which yields

$$P_0\Big(\Pi_n(B(P_0, M\varepsilon_n)|X_1, \ldots, X_n) \geq 1/2) \overset{n\to+\infty}{\longrightarrow} 1$$

and

$$P_0\Big(\hat{r}_n(P_0) \leq M\varepsilon_n\Big) \overset{n\to+\infty}{\longrightarrow} 1.$$

Finally

$$P_0\Big(\hat{r}_n(\hat{P}_n) \leq M\varepsilon_n + \delta_n\Big) \overset{n\to+\infty}{\longrightarrow} 1.$$

By using (3.5) and (3.6), we have

$$P_0\Bigg(\Big\{\Pi_n(B(\hat{P}_n,\hat{r}_n(\hat{P}_n))|X_1,\ldots,X_n)\geq\frac{1}{2}\Big\}\cap\Big\{\Pi_n(B(P_0,M\varepsilon_n)|X_1,\ldots,X_n)\overset{n\to+\infty}{\longrightarrow}1\Big\}\Bigg)\overset{n\to+\infty}{\longrightarrow}1.$$

It implies that

$$P_0\Bigg(\exists P\in B(\hat{P}_n,\hat{r}_n(\hat{P}_n))\cap B(P_0,M\varepsilon_n)\Bigg)\overset{n\to+\infty}{\longrightarrow}1.$$

Since

$$d(\hat{P}_n,P_0)\leq d(\hat{P}_n,P)+d(P,P_0),$$

we finally obtain

$$P_0\Bigg(d(\hat{P}_n,P_0)\leq\hat{r}_n(\hat{P}_n)+M\varepsilon_n\Bigg)\overset{n\to+\infty}{\longrightarrow}1$$

and

$$P_0\Bigg(d(\hat{P}_n,P_0)\leq 2M\varepsilon_n+\delta_n\Bigg)\overset{n\to+\infty}{\longrightarrow}1.$$

□

Bibliography

[1] A. Barron, M.J. Schervish and L. Wasserman, *The consistency of posterior distributions in nonparametric problems*, The Annals of Statistics, **27**, 536–561, 1999.

[2] P. Billingsley, Probability and measure, *John Wiley & Sons, Inc., New York*, 1995.

[3] L. Birgé, *Approximation dans les espaces mt́riques et théorie de l'estimation*, Z. Wahrsch. Verw. Gebiete **65**, 181–238, 2003

[4] P. Billingsley, Convergence of Probability Measures, *John Wiley & Sons, Inc., New York*, 1999.

[5] P. Diaconis and D. Freedman, *On the consistency of Bayes estimates*, The Annals of Statistics, **27**, 1–67, 1986.

[6] R.M. Dudley, Real analysis and probability. *The Wadsworth & Brooks/Cole Mathematics Series.*, 1989.

[7] T.S. Ferguson, *A Bayesian analysis of some nonparametric problems*, The Annals of Statistics, **1**, 209–230, 1973.

[8] B.A. Frigyok, A. Kapila and M. Gupta, *Introduction to the Dirichlet distribution and related processes*, Tutorial, 2010

[9] S. Ghosal, J.K. Ghosh and R.V. Ramamoorthi, *Posterior consistency of Dirichlet mixtures in density estimation*, The Annals of Statistics, **27**, 143–158, 1999.

[10] S. Ghosal, J.K. Ghosh and A.W. van der Vaart, *Convergence rates of posterior distributions*, The Annals of Statistics, **28**(2), 500–531, 2000.

[11] S. Ghosal and A.W. van der Vaart, Fundamental of Nonparametric Bayesian Inference, *Cambridge University Press*, 2017.

[12] S. Ghosal and A.W. van der Vaart, *Convergence rates of posterior distributions for non i.i.d. observations*, The Annals of Statistics, **35**(1), 192–223, 2007.

[13] J.K. Ghosh and R.V. Ramamoorthi, Bayesian Nonparametrics, *Springer*, 2003.

[14] P. Müller and F.A. Quintana, *Nonparametric Bayesian data Analysis*, Statistical Science, 95–110, 2004.

[15] C.P. Robert, The Bayesian choice, *Springer*, 2006.

[16] J. Sethuraman, *A constructive definition of Dirichlet priors*, The Statistica Sinica, **4**, 639–650, 1994.