Bayesian asymptotics for nonparametric or high dimensional problems

Antik Chakraborty, Bani K. Mallick

1 Introduction

Suppose we observe n iid samples $X_1, X_2, \ldots, X_n$ on a sample space $(\mathcal{X}^{(n)}, \mathbb{X}^{(n)})$. Assume the data is generated from a distribution P_θ, where θ is the indexing parameter belonging to some metric space $(\mathbf{\Theta}, d)$ with $d(\cdot, \cdot)$ being the underlying metric. As Bayesians, we let Π to be a prior distribution over the Borel sets of $\mathbf{\Theta}$. Now consider $\theta_0 \in \mathbf{\Theta}$ to be the true data generating parameter. Write $\Pi(\cdot \mid X^{(n)})$ as the posterior distribution. In what follows, we will be interested broadly in questions of the form,

1. Is the posterior distribution consistent? In other words, how much mass does $\Pi(\cdot \mid X^{(n)})$ place on 'small' neighborhoods of θ_0? (Need to formalize)
2. If consistent, then how fast does it converge? Or, can we say something about the size of those neighborhoods? (Again, need to formalize)
3. What is the shape of the posterior distribution around θ_0? (We won't be discussing this.)

First, we try to collect the notions that we would need to define posterior consistency. An estimator $\hat{\theta}$ is said to be consistent for $\theta_)$ in frequentist sense if it gets arbitrarily close to θ_0 as the sample size increases with probability tending to one. Note that, we have two requirements in the definition of consistency, namely, $\hat{\theta}$ gets arbitrarily close to θ and since $\hat{\theta}$ is random we want this closeness to happen with high probability. In the Bayesian paradigm, we instead have the posterior distribution and we would naturally want the posterior distribution to place its maximum mass to arbitrarily small neighborhoods around θ_0. These neighborhoods will play a central role in our development. A δ-neighborhood/ball around θ_0 in Θ is the set $B = \{\theta \in \Theta : d(\theta.\theta_0) < \delta\}$. Moreover, because the posterior distribution is a random measure (function of the random data) we want this to happen very frequently.

Definition 1.1. *The posterior distribution* $\Pi(\cdot \mid X^{(n)})$ *is said to be (weakly) consistent if* $\Pi(d(\theta, \theta_0) > \epsilon \mid X^{(n)}) \to 0$ *in* $P_{\theta_0}^{(n)}$*-probability as* $n \to \infty$ *and for every* $\epsilon > 0$. $\Pi(\cdot \mid X^{(n)})$ *is said to be (strongly) consistent if* $\Pi(d(\theta, \theta_0) > \epsilon \mid X^{(n)}) \to 0$ *in* $P_{\theta_0}^{(n)}$*-a.s. as* $n \to \infty$ *and for every* $\epsilon > 0$.

Typically, strong consistency needs more assumptions. Now suppose for a given data and a prior Π, the posterior $\Pi(\cdot \mid X^{(n)})$ is consistent. Also assume d is convex. Then by Jensen's inequality the posterior mean $\int \theta d\Pi(\theta \mid X^{(n)})$ is a consistent estimator of θ in the frequentist sense; $P_{\theta_0}^{(n)}[d\{\int \theta d\Pi(\cdot \mid X^{(n)}), \theta_0\} > \epsilon] \leq P_{\theta_0}^{(n)} \int d(\theta, \theta_0) d\Pi(\cdot \mid X^{(n)}) \to 0$ due to consistency.

Theorem 1.2. *(Doob) For any prior* Π *on* $\mathbf{\Theta}_d$*, the posterior is consistent, except possibly on a set of* Π*-measure 0.*

Doob's theorem says as long as a Bayesian is certain about the prior Π, she does not need to worry about consistency. Add to this the fact that under fairly standard regularity conditions on a finite dimensional parametric statistical model, we have by Bernstein-von-Mises theorem that $\Pi(\cdot \mid X^{(n)}) \to \mathrm{N}(\hat{\theta}, \mathrm{I}_{\theta_0}^{-1})$ where $\hat{\theta}$ is the maximum likelihood estimator (mle) and $I_{\theta_0}^{-1}$ is the inverse Fisher information matrix computed at θ_0. Since the mle is known to be consistent, we get posterior consistency. However, unless care is exercised posterior consistency might not obtain in non-parametric problems for seemingly innocuous priors. David Freedman provided examples of such peculiar behavior of the posterior. He also proved that while Doob's set of parameter values with inconsistent posterior might be small relative to Π but that set is not be 'topologically' small in many statistical problems.

2 Posterior consistency in non-parametric problems

As discussed earlier, our main focus would be on non-parametric problems. To that end, let us suppose that we have iid data $X_1, X_2, \ldots, X_n \sim P \in \mathcal{P}$ where P is some measure in the class of probability measures $\mathcal{P}$. Let us assume the true data generating distribution is $P_0 \in \mathcal{P}$. In the absence of any indexing parameter, P itself is our parameter. Consider a prior Π over $\mathcal{P}$. The Dirichlet process prior is one popular example of a prior over the set of probability measures. In order to talk about consistency, we would need some means of quantifying distances between probability measures. In the following we assume p and q to be probability densities with respect to some common dominating measure μ corresponding to probability measures $P, Q \in \mathcal{P}$. Below we define some of the common ways of measuring distance between probability densities.

1. Hellinger distance: $h^2(p,q) = \frac{1}{2}\int(\sqrt{p} - \sqrt{q})^2 d\mu$.
2. Total variation distance: $\| p - q \|_{\mathrm{TV}} = \frac{1}{2}\int \mid p - q \mid d\mu$.
3. Kullback-Liebler (KL) divergence: $D(p \,\|\, q) = \int p \log \frac{p}{q} d\mu$.

The KL divergence is not a metric because it is not symmetric in its arguments but it has a very important role in Bayesian asymptotics as will be seen later. The Hellinger and total variation distance are related by the following inequalities,

$$h^2(p,q) \lesssim \| p - q \|_{\mathrm{TV}} \lesssim h(p,q), \tag{1}$$

where $a \lesssim b$ means $a \leq Cb$ for some constant C. The above inequality clearly indicates that densities which are close in Hellinger sense are also close in total variation. There is also a famous inequality due to Pinsker which basically says that the KL divergence is stronger than Hellinger and total variation,

$$\| p - q \|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D(p \,\|\, q)} \qquad \text{(Pinsker)}. \tag{2}$$

Next, we will state a theorem due to Schwatrz. For that we first define what it means for a $p_0 \in \mathcal{P}$ to be in the KL support of the prior Π and move on to define a test function. These two will play a crucial role in proving Schwartz's result.

Definition 2.1. *Fix $p_0 \in \mathcal{P}$. We say that p_0 is in the KL support of Π, written as $p_0 \in \mathrm{KL}(\Pi)$ if for every $\epsilon > 0$*

$$\Pi\{p \in \mathcal{P} : D(p_0, p) < \epsilon\} > 0$$

.

Definition 2.2. *A test function ϕ_n is a measurable function of the data defined as $\phi_n : \mathcal{X} \to \{0, 1\}$, where $\mathcal{X}$ is the σ-field corresponding to the sample space.*

Usually, test functions are of the form $\phi_n = 1$ if $X^{(n)} \in R \subset \mathcal{X}$ and 0 otherwise. Now we are ready to state Schwartz's theorem. The notation $P\,X$ for a random variable X and probability measure P means $E_P(X)$. In particular, for a test function ϕ, since it is either 1 or 0, $P\,\phi = E_P(\phi) = P(R)$, where R is the rejection region of ϕ.

Theorem 2.3. *(Schwartz) Assume that we have observed iid $X_1, X_2, \ldots, X_n \mid p \sim p$ and $p \sim \Pi$. Let $p_0 \in \mathcal{P}$ be the true data generating distribution. Suppose for every neighborhood $\mathcal{U}$ of p_0 there exists a constant $C > 0$, measurable sets $\mathcal{P}_n \subset \mathcal{P}$ and tests ϕ_n such that,*

1. *$p_0 \in \mathrm{KL}(\Pi)$ (prior does not miss p_0)*
2. *$\Pi(\mathcal{P} - \mathcal{P}_n) < e^{-Cn}$ ($\mathcal{P}_n$ covers most of $\mathcal{P}$ relative to Π),*
3. *$P_0^n \phi_n = E_{P_0}(\phi_n) \leq e^{-Cn}$ and $\sup_{p \in \mathcal{P}_n \cap \mathcal{U}^c} P^n(1 - \phi_n) = \sup_{p \in \mathcal{P}_n \cap \mathcal{U}^c} E_{P^n}(1 - \phi_n) \leq e^{-Cn}$ (likelihood can separate p_0 from other p's in $\mathcal{P}$),*

then the posterior distribution $\Pi(\cdot \mid X^n)$ is strongly consistent at p_0. Here P_0 and P are probability measures corresponding to densities p_0 and p. P_0^n and P^n are the relative product measures.

Before providing a proof of this theorem we discuss the assumptions of Schwartz's theorem. To define a neighborhood we first need to specify a metric on $\mathcal{P}$. As nothing is mentioned about the choice of the metric, we can, for example, use the distances discussed above. The assumption of existence of sets $\mathcal{P}_n$ is made since it is not always possible to construct tests for every $p \in \mathcal{P}$. The way around is that we construct tests for $\mathcal{P}_n$ and can 'ignore' $\mathcal{P} - \mathcal{P}_n$ since it is small due to assumption 2. Furthermore, these tests must have small error probabilities; $P_0^n(\phi_n)$ is the type-I error and $P^n(1 - \phi_n)$ is the type-II error. Finally, note that the only purely Bayesian condition is assumption 1 where we are basically saying that p_0 has some positive mass assigned by the prior. In the proof, when we write $\frac{p}{q}(X)$ for two densities p and q, we mean $\frac{p(X)}{q(X)}$.

Proof. Let us first choose a metric ρ on $\mathcal{P}$ and define $\mathcal{U} = \{p \in \mathcal{P} : \rho(p, p_0) < \epsilon\}$ for a fixed $\epsilon > 0$. The posterior probability of $\mathcal{U}^c$ is then given by,

$$\Pi(\mathcal{U}^c \mid X^{(n)}) = \frac{\int_{\mathcal{U}^c} \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi(p)}{\int \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi(p)} := \frac{N_n}{D_n}. \tag{3}$$

In view of the above display, we want to show $\Pi(\mathcal{U}^c \mid X^{(n)}) \to 0$, P_0^∞ almost surely.

Step 1. We first analyze D_n. Let $\mathcal{P}_0 = \{p \in \mathcal{P} : D(p_0 \,||\, p) < \epsilon\}$. Note that we always have the following lower bound for D_n,

$$\begin{aligned} D_n &\geq \int_{\mathcal{P}_0} \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi(p) \\ &= \Pi(\mathcal{P}_0) \int_{\mathcal{P}_0} \prod_{i=1}^n \frac{p}{p_0}(X_i) \frac{d\Pi(p)}{\Pi(\mathcal{P}_0)} \\ &= \Pi(\mathcal{P}_0) \int_{\mathcal{P}_0} \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi_0(p), \end{aligned}$$

where Π_0 is the restriction of Π to $\mathcal{P}_0$. Then by Jensen's inequality applied to concave logarithm function we have the following,

$$\begin{aligned}\log D_n &\geq \log \Pi(\mathcal{P}_0) + \log \int_{\mathcal{P}_0} \prod_{i=1}^{n} \frac{p}{p_0}(X_i) d\Pi_0(p) \\ &\geq \log \Pi(\mathcal{P}_0) + \int_{\mathcal{P}_0} \log \prod_{i=1}^{n} \frac{p}{p_0}(X_i) d\Pi_0(p) \\ &= \log \Pi(\mathcal{P}_0) - n \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{P}_0} \log \frac{p_0}{p}(X_i)\end{aligned}$$

Observe that, by Fubini's theorem and strong law of large numbers

$$\frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{P}_0} \log \frac{p_0}{p}(X_i) d\Pi_0(p) \underset{a.s.}{\to} E_{P_0} \int_{\mathcal{P}_0} \log \frac{p_0}{p}(X) d\Pi_0(p).$$

The right hand side of the above display is $\int_{\mathcal{P}_0} D(p_0 \,||\, p) d\Pi_0(p)$ which is strictly greater than ϵ by definition of $\mathcal{P}_0$. Hence $\int_{\mathcal{P}_0} \log \prod_{i=1}^{n} \frac{p}{p_0}(X_i) d\Pi_0(p) > e^{-n\epsilon}$ eventually P_0^∞. This gives the following lower bound for D_n for sufficiently large n,

$$D_n \geq \Pi(\mathcal{P}_0) e^{-n\epsilon} \tag{4}$$

Step 2. Now we turn our attention to the numerator N_n. The steps we will be following are very standard in any posterior consistency proof. The idea is to make use of the tests ϕ_n which have exponentially small error probabilities. We have,

$$\begin{aligned}\Pi(\mathcal{U}^c \mid X^{(n)}) &= \Pi(\mathcal{U}^c \cap \mathcal{P}_n \mid X^{(n)}) + \Pi(\mathcal{U}^c \cap \mathcal{P}_n^c \mid X^{(n)}) \\ &\leq \Pi(\mathcal{U}^c \cap \mathcal{P}_n \mid X^{(n)}) + \Pi(\mathcal{P}_n^c \mid X^{(n)}) \\ &= T_1 + T_2,\end{aligned}$$

where the first inequality holds since $P(A \cap B) \leq P(B)$. We will now separately analyze T_1 and T_2. Note that,

$$\begin{aligned}T_1 &= \Pi(\mathcal{U}^c \cap \mathcal{P}_n \mid X^{(n)}) \\ &= \phi_n \Pi(\mathcal{U}^c \cap \mathcal{P}_n \mid X^{(n)}) + (1 - \phi_n) \Pi(\mathcal{U}^c \cap \mathcal{P}_n \mid X^{(n)}) \\ &\leq \phi_n + (1 - \phi_n) \Pi(\mathcal{U}^c \cap \mathcal{P}_n \mid X^{(n)}) \\ &= T_{11} + T_{12}.\end{aligned}$$

The expectation of T_{11} is $P_0^n \phi_n \leq e^{-Cn}$ by assumption 3. By Markov's inequality we have for every positive δ, $\sum_{n=1}^{\infty} P_0^n(\phi_n > \delta) \leq \sum_{n=1}^{\infty} \delta^{-1} e^{-Cn} < \infty$. Hence by the Borel-Cantelli lemma $\phi_n \to 0$ almost surely. (if $\sum_{n=1}^{\infty} P(\mid X_n - X \mid > \delta) < \infty$ then $X_n \overset{a.s.}{\to} X$.)

For the second term, we have the following expression by Bayes theorem,

$$T_{12} = (1 - \phi_n) \Pi(\mathcal{U}^c \cap \mathcal{P}_n \mid X^{(n)}) = (1 - \phi_n) \frac{\int_{\mathcal{U}^c \cap \mathcal{P}_n} \prod_{i=1}^{n} \frac{p}{p_0}(X_i) d\Pi(p)}{\int \prod_{i=1}^{n} \frac{p}{p_0}(X_i) d\Pi(p)}.$$

The denominator is D_n and we proved $D_n \geq \Pi(\mathcal{P}_0)e^{-n\epsilon}$ eventually almost surely. By assumption 1, $\Pi(\mathcal{P}_0)$ is positive. Hence for T_2 we need to prove that the numerator times $e^{-n\epsilon}$ goes to zero almost surely. For that we first take expectation and apply Fubini's theorem,

$$\begin{aligned} E_{P_0^n}\left\{(1-\phi_n)\int_{\mathcal{U}^c\cap\mathcal{P}_n}\prod_{i=1}^n\frac{p}{p_0}(X_i)d\Pi(p)\right\} &= \int\int_{\mathcal{U}^c\cap\mathcal{P}_n}(1-\phi_n)\prod_{i=1}^n\frac{p}{p_0}(X_i)p_0(X_i)d\Pi(p) \\ &= \int_{\mathcal{U}^c\cap\mathcal{P}_n}E_{P^n}(1-\phi_n)d\Pi(p) \leq \sup_{p\in\mathcal{U}^c\cap\mathcal{P}_n}P_n(1-\phi_n)\int_{\mathcal{U}^c\cup\mathcal{P}_n}d\Pi(p) \\ &\leq e^{-Cn}\,\Pi(\mathcal{U}^c\cup\mathcal{P}_n) \leq e^{-Cn}, \end{aligned}$$

where we used assumption 3 to upper bound $\sup_{p\in\mathcal{U}^c\cap\mathcal{P}_n}P_n(1-\phi_n)$ by e^{-Cn}. Then for every positive δ, by Markov's inequality, $\sum_{n=1}^\infty P_0^n(T_{12}>\delta) \leq \sum_{n=1}^\infty \frac{e^{n\epsilon}e^{-Cn}}{\Pi(\mathcal{P}_0)} < \infty$ if $\epsilon < C$ using the fact that $D_n \geq \Pi(\mathcal{P}_0)e^{-n\epsilon}$ eventually almost surely. Thus by Borel-Cantelli, $T_{12}\to 0$ almost surely.

Step 3. We have, $E_{P_0^n}(T_2) = \int\int_{\mathcal{P}_n^c}\prod_{i=1}^n\frac{p}{p_0}(X_i)p_0(X_i)d\Pi(p)dX_i = \int_{\mathcal{P}_n^c}d\Pi(p) = \Pi(\mathcal{P}_n^c) < e^{-Cn}$. Thus $\sum_{n=1}^\infty P_0^n(T_2>\delta) < \sum_{n=1}^\infty e^{-Cn} < \infty$ by Markov. Therefore, $T_2\to 0$ almost surely.

□

To see how consistency is proved in special cases see [1]. The authors follow the exact same steps as discussed here.

3 Tests and metric entropy

In the proof of Schwatrz's theorem we assumed existence of tests of exponentially small error probabilities. Question is when do these tests exist. The pioneering theory of minimax testing in general statistical models was developed back in 1970's and 80's by Lucien Le Cam and Lucien Brigé. Here we just state few results without proof on existence of tests.

Definition 3.1. *Let P be a probability measure and $\mathcal{Q}$ be a set of probability measures on a measure space $(\mathcal{X},\mathbb{X})$. The minimax risk for testing P versus $\mathcal{Q}$ is defined by,*

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

where ϕ is a measurable test function $\phi:\mathcal{X}\to[0,1]$ and the infimum is taken over all such ϕ's.

Let $\text{conv}(\mathcal{Q})$ denote the convex hull of $\mathcal{Q}$, i.e. $\text{conv}(\mathcal{Q})$ contains all probability measures of the form $\sum_{i=1}^k\omega_1 Q_i$, $\sum_{i=1}^k\omega_i = 1$, $k\in\mathbb{N}$.

Theorem 3.2.

$$\pi(P,\mathcal{Q}) = 1 - ||\ P - \text{conv}(\mathcal{Q})\ ||_{\text{TV}} \leq \sup_{Q\in\text{conv}(\mathcal{Q})}\rho_{1/2}(p,q),$$

where $\rho_{1/2}(p,q) = 1-\frac{1}{2}h^2(p,q)$ is the Hellinger affinity between p and q. Here p and q are the densities of P and Q relative some common dominating measure.

In view of the above theorem, when computing minimax testing error probabilities it is enough to consider the convex hull $\text{conv}(\mathcal{Q})$.

Fact: Let $\mathcal{Q}$ be a convex set of probability measures. Let P^n and Q^n be n-fold product measures for every $Q\in\mathcal{Q}$. Then we have the following,

$$\pi(P^n,\mathcal{Q}^n) \leq \rho_{1/2}(P,\mathcal{Q})^n$$

.

Theorem 3.3. *For any probability measure P and convex set of probability (dominated) measures $\mathcal{Q}$ with $h(p,q) > \epsilon$ for every $q \in \mathcal{Q}$ and any $n \in \mathbb{N}$, there exists a test ϕ such that,*

$$P^n\phi \leq e^{-n\epsilon^2/2}, \qquad \sup_{Q\in\mathcal{Q}} Q^n(1-\phi) \leq e^{-n\epsilon^2/2}.$$

So the above theorem basically says that it is always possible to construct tests between a probability measure P and a convex set of probability measures with exponentially small error probability given that elements of the convex set are strictly ϵ distance away from P in the Hellinger sense.

Recall the proof of Schwartz's theorem. We need to construct tests for $\mathcal{U}^c$ which is not convex in general. However, we can get past this obstacle by following a simple rule. Suppose $\mathcal{P}$ is a set of probability measures. Fix $P_0, P_1 \in \mathcal{P}$. Consider the problem of testing $H_0 : P = P_0$ vs $H_1 : P \in \{P : \rho(P, P_1) < \rho(P_0, P_1)/2\}$. The ball in H_1 is in general convex and by the previous theorem we know that a test exists with small error probability. Now, we can cover $\mathcal{P}$ by balls centered at different points $P_2, \ldots, P_N$ and have a test for each of those balls. We can then combine all these tests to form a single test for the entire space. The power of this combined test then will depend on the number of balls needed to cover $\mathcal{P}$. This number has a special name which we are going to define next. It should also be clear that some control on this number is needed, otherwise the test will loose power.

In the following we assume $(\mathbb{T}, \rho)$ to be a metric space.

Definition 3.4. *A δ-cover of a set $\mathbb{T}$ with respect to a metric ρ is a set $\{\theta_1, \ldots, \theta_N\} \subset \mathbb{T}$ such that for every $\theta \in \mathbb{T}$ there exists some $i \in 1, 2, \ldots, N$ such that $\rho(\theta, \theta_i) \leq \delta$. The δ-covering number $N(\delta, \mathbb{T}, \rho)$ is the cardinality of the smallest δ-cover.*

The number $\log N(\delta, \mathbb{T}, \rho)$ is known as the metric entropy of $\mathbb{T}$ with respect to ρ. Essentially, this measures the size of the set $\mathbb{T}$ relative to ρ. Below we include two examples of the behavior of covering numbers.

Example 3.5. *Consider a norm $||\cdot||$ on $\mathbb{R}^d$. Define the unit ball $\mathbb{B}$ as $\mathbb{B} = \{x \in \mathbb{R}^d : || x ||\leq 1\}$. Note that $\rho(x, y) =|| x - y ||$ defines a metric on $\mathbb{R}^d$. Hence $(\mathbb{B}, ||\cdot||)$ is a valid metric space. It can be shown that,*

$$\log N(\delta, \mathbb{B}, ||\cdot||) \asymp d \log(1/\delta).$$

Example 3.6. *Let $\mathcal{F} = \{f(x) : f : [0,1] \to \mathbb{R}, \forall x, y \in [0,1], \mid f(x) - f(y) \mid\lesssim \mid x - y \mid\}$. The function class $\mathcal{F}$ has a special name, its members are known as 1-Lipschitz functions. It should be observed that $\mathcal{F} \subset C[0,1]$, the set of all continuous functions on $[0,1]$. The couplet $(\mathbb{B}, ||\cdot||_\infty)$ is a metric space where $||\cdot||_\infty = \sup_x \mid f(x) \mid$. For this set it is known that,*

$$\log N(\delta, \mathcal{F}, ||\cdot||_\infty) \asymp \frac{1}{\delta}.$$

If instead we consider the d-dimensional analogue of $\mathcal{F}$ defined as $\mathcal{F}_d = \{f(x) : f : [0,1]^d \to \mathbb{R}, \forall x, y \in [0,1]^d, \mid f(x) - f(y) \mid\lesssim || x - y ||_\infty\}$, the corresponding covering number is,

$$\log N(\delta, \mathcal{F}, ||\cdot||_\infty) \asymp \left(\frac{1}{\delta}\right)^d$$

Comparing the covering numbers for $\mathbb{B}$ and $\mathcal{F}_d$ we can immediately see how the dimension d appears linearly for $\mathbb{B}$ and in the exponent of $\mathcal{F}_d$ even in the log scale. The other major difference

is that the presence of the large number $1/\delta$ (assuming δ to be small) in $\log N(\delta, \log \mathcal{F}_d, ||\cdot||_\infty)$, whereas for $\log N(\delta, \mathbb{B}, ||\cdot||)$, $1/\delta$ only appears through $\log(1/\delta)$. An excellent reading material on covering numbers and related topics can be found here.

The next result we state relates the covering number of a set of statistical models $\mathcal{P}$ (a set of probability measures) with metric ρ to existence of certain tests with small error probabilities. The proof of the theorem can be found in Section 7 of [2]. Suppose we observe the iid data $X_1, \ldots, X_n \mid P \sim P$, and let $(\mathcal{X}^{(n)}, \mathbb{X}^{(n)})$ be the sample space. Here P is some probability measure in $\mathcal{P}$, the set of all probability measures on $(\mathcal{X}^{(n)}, \mathbb{X}^{(n)})$. Suppose for a fixed $P \in \mathcal{P}$, P^n denote the product probability measure, $P^n = \otimes_{i=1}^n P_i = \otimes_{i=1}^n P = P^n$.

Theorem 3.7. *(Theorem 7.1 of [2]) Consider the metric space (P, ρ). Assume ρ is dominated by the Hellinger metric. Suppose that for some non-increasing function $N(\epsilon)$, some $\epsilon_n \geq 0$ and every $\epsilon > \epsilon_n$,*

$$\mathcal{M}(\epsilon/2, \{P : \rho(P, P_0) \leq 2\epsilon\}, \rho) \leq N(\epsilon).$$

Then for every $\epsilon > \epsilon_n$ there exist tests $\phi_n = \phi_n(\epsilon)$ such that, for a universal constant K and every $j \in \mathbb{N}$,

$$P_0^n \phi_n \leq N(\epsilon) \frac{e^{-Kn\epsilon^2}}{1 - e^{-Kn\epsilon^2}}, \tag{5}$$

and

$$\sup_{\rho(P,P_0) > j\epsilon} P^n(1 - \phi_n) \leq e^{-Kn\epsilon^2 j^2}. \tag{6}$$

Proof. For any $j \in \mathbb{N}$ define $S_j = \{P : j\epsilon < \rho(P, P_0) < (j+1)\epsilon\}$. Suppose S_j^* is a $j\epsilon/2$-cover of S_j, i.e. $S_j^* = \{P_{j1}, P_{j2}, \ldots, P_{jN_j}\}$ such that for any $P \in S_j$ there exists $l \in 1, \ldots, N_j$ such that $\rho(P, P_{jl}) < j\epsilon/2$. Note that N_j is the $j\epsilon/2$ covering number of S_j with respect to the metric ρ which by assumption of the theorem is bounded above by $N(j\epsilon)$. Now fix $l \in \{1, \ldots, N_j\}$ and consider the ball $B_{jl} = \{P : \rho(P, P_{jl}) < j\epsilon/2\}$ centered at P_{jl}. Then by theorem 3.3 there exists a test ϕ_{jl} for testing $H_0 : P = P_0$ vs $H_1 : P \in B_{jl}$ such that,

$$P_0^n \phi_{jl} \leq e^{-nj^2\epsilon^2/2},$$
$$\sup_{P \in B_{jl}} P^n(1 - \phi_{jl}) \leq e^{-nj^2\epsilon^2/2}.$$

Let us define $\phi_j = \max \phi_{j1}, \ldots, \phi_{jN_j}$. Finally, let $\phi_n = \max \phi_j, \phi_{(j+1)}, \ldots,$ for any fixed $j \in \mathbb{N}$. Now we will compute the error probabilities of the test ϕ_n. Recall that since ϕ_{jl}'s are test functions they are of the form $\phi_{jl} = 1$ if $X^{(n)} \in R_{jl}$ and 0 otherwise, where R_{jl} is a subset of $\mathbb{X}^{(n)}$. Hence for any $P^n \in \mathcal{P}$, $P^n \phi_{jl} = E_{P^n}(\phi_{jl}) = P^n(X^{(n)} \in R_{jl})$. And $P^n \phi_j$ is nothing but $P^n(X^{(n)} \in \cup_l R_{jl})$ and similarly $P^n \phi_n = E_{P^n}(\phi_n) = P^n(X^{(n)} \in \cup_j \cup_l R_{jl})$. Thus, by the union probability bound $P(\cup A_i) \leq \sum P(A_i)$ we get,

$$P_0^n \phi_n = P_0^n(\cup_j \cup_l \phi_{jl}) \leq \sum_j \sum_l P_0^n(R_{jl}) \leq \sum_j N(j\epsilon) e^{-Kn\epsilon^2} \leq N(\epsilon) \sum_j e^{-Kn\epsilon^2} = N(\epsilon) \frac{e^{-Kn\epsilon^2}}{1 - e^{-Kn\epsilon^2}},$$

where $K = 1/2$ and $N(j\epsilon) \leq N(\epsilon)$ for any $j \in \mathbb{N}$ since $N(\cdot)$ is non-increasing by assumption. For the type-II error, we have for any $P \in S_j$, $P^n(1 - \phi_j) = P^n(X^{(n)} \in \cap_l R_{jl}^c) \leq \sup_{P^n \in \cup B_{jl}} P^n(R_{jl}^c) \leq e^{-nj^2\epsilon^2/2}$. Thus,

$$\sup_{\rho(P,P_0) > j\epsilon} P^n(1 - \phi_n) = \sup_{P \in \cup_{l>j} S_l} P^n(1 - \phi_n) = \sup_{P \in \cup_{l>j} S_l} P^n(1 - \phi_l) \leq \sup_{l>j} e^{-nl^2\epsilon^2/2} \leq e^{-Knj^2\epsilon^2}.$$

□

In the previous theorem, one possible choice of $N(\cdot)$ can be $\mathcal{N}(\epsilon, \mathcal{P}, \rho)$, the covering number of the entire space or a large subset $\mathcal{P}_n \subset \mathcal{P}$ where we will see later that 'small' is measured with respect to the prior measure. See section 7 of [2] for more on tests and metric entropy.

4 Posterior contraction rates

As in section 2 we will restrict our attention to nonparametric problems. Schwartz's theorem, while important to understand certain operating characteristics of Bayes procedures in nonparametric problems, is only asymptotic in nature. In statistical terms, to have a better gauge of the performance of a method we look at its convergence rate, i.e. how fast with respect to the sample size does it converge to the true distribution. For example, in a fixed dimensional parametric model it is well known that the maximum likelihood estimator has a convergence rate of $O_P(n^{-1/2})$. In a Bayesian framework however, we are interested in the convergence rate of the entire posterior distribution. More precisely, if P_0 is the true distribution and ρ is our chosen metric on $\mathcal{P}$ can we create a shrinking neighborhood around P_0 of radius ϵ_n with $\epsilon_n \to 0$ (note the dependence of the radius on the sample size n) so that the posterior distribution places most of its mass in that neighborhood. Let us now define formally what is meant by the convergence rate of posterior distribution. We will restrict our attention to data which are iid according to some distribution.
Setup. Suppose we observe the iid data $X_1, \ldots, X_n \mid P \sim P$, where P is some probability measure in $\mathcal{P}$. Let $(\mathcal{X}^{(n)}, \mathbb{X}^{(n)})$ be the sample space and $P_0 \in \mathcal{P}$ be the true distribution. Consider the prior Π on P. Let $\Pi_n(\cdot \mid X^{(n)})$ be the posterior distribution obtained by using Bayes theorem. Assume the metric ρ is dominated by the Hellinger metric (eg. Total variation). Also assume that the probability measures $P \in \mathcal{P}$ have a density with respect respect to some common dominating measure λ.

Definition 4.1. *The posterior distribution $\Pi_n(\cdot \mid X^{(n)})$ is said to contract at rate $\epsilon_n \to 0$ at $P_0 \in \mathcal{P}$ if $\Pi_n(P : \rho(P, P_0) > M\epsilon_n \mid X^{(n)}) \to 0$-in P_0^n-probability for some positive constant M.*

Note that, if ϵ_n is the rate of contraction then every $\tilde{\epsilon}_n$ such that $\tilde{\epsilon}_n > \epsilon_n \to 0$ is a contraction rate. Naturally, we are interested in the fastest rate possible which is attained for every $P_0 \in \tilde{\mathcal{P}}$, some subclass of $\mathcal{P}$. If ρ is convex then by a similar argument provided after definition 1.1, we can see that the posterior mean also converges to P_0 at rate ϵ_n. Next we state a famous result from [2] (theorem 2.1) which provides sufficient conditions for the posterior distribution to contract at rate ϵ_n. Consider the setup mentioned previously.

Theorem 4.2. *Suppose $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$. If for a constant $C > 0$ and sets $\mathcal{P}_n \subset \mathcal{P}$, we have,*

1. $\log N(\epsilon_n, \mathcal{P}_n, \rho) \leq n\epsilon_n^2$.
2. $\Pi(\mathcal{P}_n^c) \leq e^{-(C+4)n\epsilon_n^2}$.
3. $\Pi\{p : \int p_0 \log \frac{p_0}{p} \leq \epsilon_n^2, \int p_0 \log^2 \frac{p_0}{p} \leq \epsilon_n^2\} \geq e^{-Cn\epsilon_n^2}$,

then for sufficiently large M, $\Pi_n(P : \rho(P, P_0) > M\epsilon_n \mid X^{(n)}) \to 0$ in P_0^n-probability, i.e. the posterior distribution contracts at P_0 at rate ϵ_n.

Let us first discuss the conditions of the theorem before going into the proof which we hope would provide a brief insight into the ideas guiding the proof. It should be noted that the conditions

are very similar to Schwartz's theorem. Our main goal here is to show the posterior probability of the set $U = \{P : \rho(P, P_0) > M\epsilon_n\}$ is small i.e.

$$\Pi_n(U \mid X^{(n)}) = \frac{\int_U \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi(p)}{\int \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi(p)} := \frac{N_n}{D_n},$$

is small. For that to happen, naturally we would want an upper bound on N_n and lower bound on D_n. Let $B = \{p : \int p_0 \log \frac{p_0}{p} \leq \epsilon_n^2, \int p_0 \log^2 \frac{p_0}{p} \leq \epsilon_n^2\} \geq e^{-n\epsilon_n^2}\}$. This set should be recognized as a Kullback-Liebler type neighborhood around P_0 with an additional condition on the second moment of the log-likelihood ratio. The denominator D_n as we have seen in the proof of Schwartz's theorem, can be lower bounded easily in terms of the prior probability assigned to B and with the help of the second moment condition we would be able to attach a probability statement to this lower bound, i.e. we would be able to say that $D_n \geq \Pi(B)e^{-n\epsilon_n^2}$ with high probability which is a more precise statement than just saying $D_n \geq D$ ultimately for some $n \geq N$. However, the expected value of the numerator N_n with respect to P_0^n is, $E_{P_0^n} \int_U \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi(p) = \int \int_U \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi(p) p_0(X_i) dX_i = \Pi(U)$, the prior probability of the set U by Fubini's theorem. This means *apriori* we need to have a very strong idea as to what P_0 is, which in practice we rarely have. To make the numerator small, we would use test functions to test $H_0 : P = P_0$ vs $H_1 : P \in U$. Since by assumption ρ is dominated by Hellinger metric, by theorem 3.7 we have a test with small error probabilities for testing $H_0 : P = P_0$ vs $H_1 : P \in \{P : \rho(P, P_0) > j\epsilon_n\}$ for every $j \in \mathbb{N}$. For every $\epsilon > \epsilon_n$, the type- I error of this test is bounded above by $N(\epsilon)e^{-Kn\epsilon^2}/(1 - e^{-Kn\epsilon_n^2})$. We shall choose $N(\epsilon) = N(\epsilon_n, \mathcal{P}_n, \rho)$, the covering number of $\mathcal{P}_n$ with respect to ρ. The denominator of the error bound is close to 1 since $n\epsilon_n^2 \to \infty$ and if $N(\epsilon_n) \leq e^{-n\epsilon_n^2}$ then the error is of the order $e^{-n\epsilon_n^2}$. This is guaranteed by assumption 1. Finally, like Schwartz's theorem, we would show that the posterior probability of $\mathcal{P}_n^c$ is small since its prior probability is small by assumption 2. We now prove theorem 4.2.

Proof. We will prove the theorem in three steps. In the first step we analyze D_n. N_n is analyzed in the second step.

Step 1. Let $B = \{p : \int p_0 \log \frac{p_0}{p} \leq \epsilon_n^2, \int p_0 \log^2 \frac{p_0}{p} \leq \epsilon_n^2\} \geq e^{-n\epsilon_n^2}\}$. Similar to step 1 in the proof of Schwartz's theorem, we can lower bound D_n as, $D_n \geq \Pi(B) \int_B \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi_0(p)$ where Π_0 is the restriction of Π on B (replacing $\mathcal{P}_0$ by B). Then again using a similar argument as in the proof of Schwartz's theorem, $\log D_n \geq \log \Pi(B) + \int_B \log \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi_0(p)$. Let $Z = \int_B \log \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi_0(p) = \sum_{i=1}^n \int_B \log \frac{p}{p_0}(X_i) d\Pi_0(p)$. The expected value of Z under P_0^n is $-n \int_B D(p_0 \,||\, p) d\Pi_0(p) \geq -n\epsilon_n^2$ by definition of B. For the variance of Z we have the following due to independence:

$$\begin{aligned}
\operatorname{var}_{P_0^n}(Z) \leq n \operatorname{var}_{P_0^n} \int_B \log \frac{p}{p_0}(X_1) d\Pi_0(p) &\leq n E_{P_0^n} \left(\int_B \log \frac{p}{p_0}(X_1) d\Pi_0(p) \right)^2 \\
&\leq n \int_B \left(\log \frac{p}{p_0}(X_1) d\Pi_0(p) \right)^2 d\Pi_0(p) \\
&= n \int_B \left(-\log \frac{p_0}{p}(X_1) \right)^2 d\Pi_0(p) \\
&= n \int_B \left(\log \frac{p_0}{p}(X_1) \right)^2 d\Pi_0(p) \leq n\epsilon_n^2,
\end{aligned}$$

again by definition of B. Hence,

$$P_0^n(Z < -2n\epsilon_n^2) = P_0^n(Z - (-n\epsilon_n^2) < -n\epsilon_n^2) \leq P_0^n(Z - E(Z) < -n\epsilon_n^2) \leq \frac{n\epsilon_n^2}{(n\epsilon_n^2)^2}.$$

Thus we have $D_n \geq \Pi(B)e^{-2n\epsilon_n^2} = e^{-(C+2)n\epsilon_n^2}$ with P_0^n-probability at least $1 - 1/(n\epsilon_n^2)$.

Step 2. Since $N(\epsilon, \mathcal{P}_n, \rho)$ is decreasing in ϵ, we have for very $\epsilon > 2\epsilon_n$,

$$\log N(\epsilon/2, \mathcal{P}_n, \rho) \leq \log N(\epsilon_n, \mathcal{P}_n, \rho) \leq n\epsilon_n^2,$$

where the last inequality follows from assumption 1. Now we will apply theorem 3.7 with $N(\epsilon) = e^{n\epsilon_n^2}$ and $\epsilon = M\epsilon_n$ and set $j = 1$ for some positive constant M to be chosen later. By theorem 3.7 there exists a test ϕ_n such that,

$$P_0^n\phi_n \leq e^{n\epsilon_n^2}\frac{e^{-KnM^2\epsilon_n^2}}{1 - e^{-KnM^2\epsilon_n^2}} \tag{7}$$

$$\sup_{P \in \mathcal{P}_n : \rho(P,P_0) > M\epsilon_n} P^n(1 - \phi_n) \leq e^{-KnM^2\epsilon_n^2} \tag{8}$$

Then consider the posterior probability assigned to the set $U = \{P : \rho(P, P_0) > M\epsilon_n\}$,

$$\Pi_n(U \mid X^{(n)}) = \frac{\int_U \prod_{i=1}^n \frac{p}{p_0}(X_i)d\Pi(p)}{\int \prod_{i=1}^n \frac{p}{p_0}(X_i)d\Pi(p)} = \frac{N_n}{D_n},$$

Suppose A_n is the event where $D_n \geq e^{-(C+2)n\epsilon_n^2}$. Write $\Pi_n(U \mid X^{(n)}) = \Pi_n(U \mid X^{(n)})\mathbb{I}_{A_n} + \Pi_n(U \mid X^{(n)})\mathbb{I}_{A_n^c}$. Then $E_{P_0^n}\Pi_n(U \mid X^{(n)}) \leq E_{P_0^n}(\Pi_n(U \mid X^{(n)})\mathbb{I}_{A_n}) + E_{P_0^n}(\mathbb{I}_{A_n^c})$ since $\Pi_n(U \mid X^{(n)})$ is a number between 0 and 1. Thus $E_{P_0^n}\Pi_n(U \mid X^{(n)}) \leq E_{P_0^n}(\Pi_n(U \mid X^{(n)}))\mathbb{I}_{A_n} + P_0^n(A_n^c) \leq \Pi_n(U \mid X^{(n)})\mathbb{I}_{A_n} + 1/(n\epsilon_n^2)$. Since $n\epsilon_n^2 \to \infty$ by assumption, we only need to consider the expected value of $\Pi_n(U \mid X^{(n)})$ inside A_n. Hence for every $X^{(n)} \in A_n$,

$$\begin{aligned}
E_{P_0^n}(\Pi_n(U \mid X^{(n)})) &= E_{P_0^n}(\Pi_n(U \cap \mathcal{P}_n \mid X^{(n)})) + E_{P_0^n}(\Pi_n(U \cap \mathcal{P}_n^c \mid X^{(n)})) \\
&\leq E_{P_0^n}(\Pi_n(U \cap \mathcal{P}_n \mid X^{(n)})) + E_{P_0^n}(\Pi_n(\mathcal{P}_n^c \mid X^{(n)})) \\
&= E_{P_0^n}(\phi_n\Pi_n(U \cap \mathcal{P}_n \mid X^{(n)})) + E_{P_0^n}((1 - \phi_n)\Pi_n(U \cap \mathcal{P}_n \mid X^{(n)})) + E_{P_0^n}(\Pi_n(\mathcal{P}_n^c \mid X^{(n)})) \\
&\leq E_{P_0^n}(\phi_n) + E_{P_0^n}((1 - \phi_n)\Pi_n(U \cap \mathcal{P}_n \mid X^{(n)})) + E_{P_0^n}(\Pi_n(\mathcal{P}_n^c \mid X^{(n)})) \\
&= T_{11} + T_{12} + T_2
\end{aligned}$$

It is easy to observe that $E_{P_0^n}(T_2) \leq e^{-(C+4)n\epsilon_n^2}$ (see step3 of Schwartz's proof) and $E_{P_0^n}(T_2) \to 0$. If M is such that $KM^2 - 1 > K$ then by equation (7),

$$E_{P_0^n}(T_{11}) \leq \frac{e^{(1-KM^2)n\epsilon_n^2}}{1 - e^{-KnM^2\epsilon_n^2}} \leq \frac{e^{-Kn\epsilon_n^2}}{1 - e^{-KnM^2\epsilon_n^2}} \leq 2e^{-Kn\epsilon_n^2}.$$

Finally, $E_{P_0^n}(T_{12}) \leq e^{-KnM^2\epsilon_n^2}e^{(C+2)n\epsilon_n^2}$. To see this, first use the lower bound on D_n inside A_n and then take expectation of the numerator. The expected value of the numerator is bounded by $\sup_{P \in U \cap \mathcal{P}_n} P^n(1 - \phi_n) \leq e^{-KnM\epsilon_n^2}$ which follows by equation (8). Then if $M \geq \sqrt{(C+4)/K}$, we see that the posterior distribution contracts at the rate ϵ_n. □

References

[1] Artin Armagan, David B Dunson, Jaeyong Lee, Waheed U Bajwa, and Nate Strawn. Posterior consistency in linear models under shrinkage priors. *Biometrika*, 100(4):1011–1018, 2013.

[2] Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, 2000.