Asymptotics

STA 941. Surya Tokdar

1 Prelude

Quantifying asymptotic guarantees of nonparametric Bayesian methods is an useful exercise for several reasons. Asymptotic guarantees provide frequentist justification of these methods in large samples, which could be attractive to non-Bayesian practitioners who use these methods for their flexibility and convenience. Second, asymptotic guarantees, particularly guarantees of adaptive rates of convergence across model complexity classes, are an indirect validation that the spread of the underlying prior distribution is appropriately balanced across its infinite dimensional support, maintaining a good trade-off between flexibility and complexity.

For Bayesian methods, asymptotic guarantees are usually characterized by convergence properties of the entire posterior distribution, rather than a single estimate. Consider a sequence of samples D^n, $n = 1, 2, \ldots$, modeled as

$$D^n \sim p_n(d^n|\theta), \theta \in \Theta; \quad \theta \sim \Pi_n,$$

and $\Pi_n(\cdot|D^n)$ denoting the posterior distribution on Θ based on the n-th sample:

$$\Pi_n(B|D^n) = \frac{\int_B p_n(D^n|\theta)d\Pi_n(\theta)}{\int_\Theta p_n(D^n|\theta)d\Pi_n(\theta)}, \quad B \subset \Theta.$$

Assume that with increasing n, data D^n provide more information on the parameter θ. This is usually the case in IID settings, where, $D^n = (X_1, \ldots, X_n) \in \mathcal{X}^n$ with $p_n(d^n|\theta) = \prod_{i=1}^n f(x_i|\theta)$, $\theta \in \Theta$. In such scenarios, if the true value of θ equaled a $\theta_0 \in \Theta$, we would expect $\Pi_n(\cdot|D^n)$ to converge *weakly* to the Dirac measure δ_{θ_0} on Θ. If such convergence happens, we say the posterior is consistent at θ_0. When consistency holds, it is often possible to quantify the rate at which the posterior contracts to δ_{θ_0}. Estimators derived from the posterior distribution (e.g., posterior mean) usually converge to θ_0 at leas as fast – making it possible to judge the efficiency of such estimators in a purely frequentist sense.

2 Posterior Consistency

2.1 Definition

The notion of weak convergence of $\Pi_n(\cdot|D^n)$ to δ_{θ_0} depends on the topology of Θ. To formalize ideas, let Θ be a metric space with metric d. Then weak convergence of probability measures on Θ could be metrized by a metric d_W (Levy-Prokhorov or Wasserstein). Define $\rho_n(\theta) := d_W(\Pi_n(\cdot|D^n), \delta_\theta)$, which is a function of D^n and θ.

Definition 1. We say the posterior (sequence) is consistent at θ_0 if $\rho_n(\theta_0) \to 0$ a.s. or in probability when $D^n \sim p_n(\cdot|\theta_0)$, $n = 1, 2, \ldots$.

Working with ρ_n directly is technically challenging. Fortunately simpler but equivalent definitions of consistency are available with a little extra assumption on Θ.

Lemma 1. *When the metric space (Θ, d) is separable (which holds for most models we have dealt with), $\rho_n(\theta_0) \to 0$ almost surely/in probability if and only if for every open neighborhood U of θ_0, $\Pi_n(U^c|D^n) \to 0$ almost surely/in probability (all convergences to be evaluated when true $\theta = \theta_0$).*

This is a particularly convenient characterization of posterior consistency and could be taken as the definition. Notice that $\Pi_n(U^c|D^n) \to 0$ in probability under $\theta = \theta_0$ if and only $\mathbb{E}\{\Pi_n(U^c|D^n)|\theta = \theta_0\} \to 0$.

Posterior consistency at any θ_0 automatically implies some rate of convergence of $\Pi_n(\cdot|D^n)$ to δ_{θ_0}. This is easiest to see when consistency holds almost surely. For any $\epsilon > 0$, the sequence $\delta_n(\epsilon) := \Pi_n(d(\theta, \theta_0) > \epsilon|D^n)$ converges to zero almost surely by definition of consistency. So, with probability one, there exist natural numbers $m_1 < m_2 < \cdots$ such that $\delta_n(1/k) \le 1/k$ for all $n \ge m_k$. Take $\epsilon_n = 1/k$ for $n = m_k, m_k + 1, \ldots, m_{k+1} - 1$. Then, $\delta_n(\epsilon_n) \to 0$ almost surely, i.e., the posterior contracts to the truth at least as fast as ϵ_n. The posterior contract rate is defined as the fastest rate ϵ_n at which contraction takes place, i.e., $\Pi_n(d(\theta, \theta_0) > \epsilon_n|D^n) \to 0$ but $\limsup_n \Pi_n(d(\theta, \theta_0) > \epsilon'_n|D^n) > 0$ for any $\epsilon'_n = o(\epsilon_n)$. For in probability convergence, the same arguments hold, but on the sequence $\delta_n(\epsilon) = \mathbb{E}\{\Pi_n(d(\theta, \theta_0) > \epsilon|D^n)|\theta_0\}$.

2.2 Consequences of consistency

Here I highlight two important consequences of posterior consistency, one of strong frequentist interest and one of a very Bayesian interpretation. As before (Θ, d) is assumed to be a metric space

Proposition 2. *Let $\Theta^* \subset \Theta$ be a subset such that posterior consistency holds at every $\theta_0 \in \Theta^*$. Then*

1. *there exists an estimator $\hat{\theta}_n = T(D^n)$ that is consistent for every $\theta_0 \in \Theta^*$, i.e., for every $\theta_0 \in \Theta^*$, $d(\hat{\theta}_n, \theta_0) \to 0$ almost surely/in probability when $D^n \sim p_n(\cdot|\theta_0)$. If the posterior contracts to a $\theta_0 \in \Theta^*$ at a rate ϵ_n then $\epsilon_n^{-1} \cdot d(\hat{\theta}_n, \theta_0)$ is bounded almost surely/in probability.*

2. *If Θ is convex and d is bounded and convex then one could take $\hat{\theta}_n$ to be the posterior mean $\bar{\theta}_n = \int \theta d\Pi_n(\theta|D^n)$.*

For many nonparametric estimation problems, the parameter space Θ is indeed convex. This happens for density estimation with IID data $X_i \overset{\text{IID}}{\sim} f(\cdot|\theta)$ where $\theta \in \Theta = \mathcal{P}(\mathcal{X}) \cap L_1(\lambda)$ for a given σ-finite measure λ on $\mathcal{X}$ and $f(\cdot|\theta) = d\theta/d\lambda$ is the density of θ wrt λ. Convexity also holds for nonparametric regression $Y_i = \theta(X_i) + \epsilon_i$, $\theta \in \Theta = C(\mathcal{X})$ or $L_2(\lambda)$.

Our second result is on "merging" of posterior inferences of two Bayesians who start out with different prior specifications. We state this in the IID context only, with a single prior specification $\Pi_n \equiv \Pi$ on Θ that is used across all n. Let Γ be a different prior on Θ chosen by another statistician. Let P_Γ^∞ denote the marginal probability distribution of $(X_1, X_2, \ldots)$ under the model: $X_i \overset{\text{IID}}{\sim} f(\cdot|\theta)$, $\theta \sim \Gamma$.

Theorem 3. *$d_W(\Pi(\cdot|D^n), \Gamma(\cdot|D^n)) \to 0$ almost surely $[P_\Gamma^\infty]$ if and only if $\Pi(\cdot|D^n)$ is consistent at every $\theta \in \text{supp}(\Gamma)$ in the almost sure sense, i.e., $\rho_n(\theta) \to 0$ a.s. P_θ^∞.*

A proof may be found in Diaconis and Freedman (1986). The theorem implies that posterior inference drawn by the statistician using Π will merge with the inference drawn by the other statistician on almost every data that latter expects to see. When merging happens, the two statisticians also agree on the predictions they make. Let $Q_\Pi^\infty(\cdot|D^n)$ and $Q_\Gamma^\infty(\cdot|D^n)$ denote the posterior predictive distributions of $X_{n+1}, X_{n+2}, \ldots$ obtained by these two statisticians given D^n, then $d_W(Q_\Pi^\infty(\cdot|D^n), Q_\Gamma^\infty(\cdot|D^n)) \to 0$ almost surely $[P_\Gamma^\infty]$ whenever Π attains posterior consistency at every $\theta \in \text{supp}(\Gamma)$ in the almost sure sense.

2.3 When does posterior consistency hold?

Given the sequence of statistical models $p_n(\cdot|\theta)$, $\theta \in \Theta$, what properties of the prior sequence Π_n will guarantee posterior consistency at a given θ_0? Clearly a minimum requirement should be that Π_n should not *a priori* rule out θ_0 as a possibility, i.e., θ_0 should be included in the support of each Π_n for all large n. This is often all that is needed, e.g., in regular parametric models. The same holds even for some nonparametric models that are almost like parametric models.

Example 1. Consider the IID case with $\mathcal{X} = \mathbb{N}$, $\Theta = \Delta_\infty$ and $f(x|\theta) = \theta_x$, $x \in \mathbb{N}$, $\theta \in \Theta$. Also let $d(\theta, \theta') = \|\theta - \theta'\|_1 = \sum_{l=1}^\infty |\theta_l - \theta'_l|$. Suppose $\Pi_n \equiv \Pi$, a fixed probability measure on Θ. If $\theta_0 \in \Theta$ has only finitely many non-zero elements and

$$\Pi(\{\theta \in \Theta : \|\theta - \theta_0\|_1 < \epsilon\}) > 0, \text{ for every } \epsilon > 0, \tag{1}$$

then the posterior is consistent at θ_0. □

In the situation above, if θ_0 has infinitely many non-zero elements, then posterior consistency may not hold at θ_0 even if (1) holds. Freedman (1963) provides a host of counterexamples, including the following one where data comes from the *Geo*(1/4) distribution but the posterior concentrates at the *Geo*(3/4) distribution.

Example 2 (Freedman (1963))**.** Consider the function $S : [1/8, 7/8] \to \mathbb{R} \cup \{\infty\}$ given by

$$
\begin{aligned}
&S(1/4) = S(3/4) = \infty,\\
&S(\eta) = \log_4 \log_{10} \frac{1}{|\eta - 1/4|}, \quad \eta \in \frac{1}{4} \mp \frac{1}{10^4}\\
&S(\eta) = \frac{1}{16}\frac{1}{|\eta - 3/4|}, \quad \eta \in \frac{3}{4} \mp \frac{1}{16}\\
&S(\eta) = 1, \quad \text{otherwise.}
\end{aligned}
$$

Then $S(\eta)$ is continuous, and has only two local (and global) modes at $1/4$ and $3/4$. Consider a map ϕ that maps an $\eta \in [1/8, 7/8]$ to a $\theta \in \Delta_\infty$ given as follows. Let $k = \lfloor S(\eta) \rfloor$, the largest integer smaller than or equal to $S(\eta)$. Take,

$$
\theta_i = \eta(1-\eta)^{i-1}, \; i = 1, 2, \ldots, k, \text{ and } \theta_i = 0, \; i = k+3, k+4, \ldots.
$$

If $S(\eta) = k$ then take $\theta_{k+1} = (1-\eta)^k$, $\theta_{k+2} = 0$. Otherwise, find the nearest $\underline{\eta}$, $\bar{\eta}$ such that $S(\underline{\eta}) = k$, $S(\bar{\eta}) = k+1$ and define

$$
\begin{aligned}
\theta_{k+1} &= (1-\eta)^k \frac{\bar{\eta} - \eta}{\bar{\eta} - \underline{\eta}} + \eta(1-\eta)^k \frac{\eta - \underline{\eta}}{\bar{\eta} - \underline{\eta}}\\
\theta_{k+2} &= (1-\eta)^{k+1} \frac{\eta - \underline{\eta}}{\bar{\eta} - \underline{\eta}}.
\end{aligned}
$$

This indeed defines a θ that is a an element of Δ_∞. Importantly, the map $\phi : \eta \mapsto \theta$ is continuous with respect to the L_1 metric on Δ_∞ (which is equivalent to coordinatewise continuity, which follows directly from the construction above). Let $\Theta \subset \Delta_\infty$ be the image of $[1/8, 7/8]$ under ϕ and Π be the probability measure of $\phi(\eta)$ when $\eta \sim$ *Unif*$(1/8, 7/8)$. By continuity of ϕ, Π satisfies (1) at $\theta_0 = \phi(1/4)$, which is same as the *Geo*$(1/4)$ distribution. However when data comes from *Geo*$(1/4)$ [under the IID setting as in the previous example], the posterior $\Pi(\cdot|D^n)$ concentrates around $\theta^* = \phi(3/4)$.

To see why this happens, it is easier to work on the η parametrization. The posterior on η concentrates on the subinterval of $[1/8, 7/8]$ in which $S(\eta) \geq U_n - 1$ where $U_n = \max(X_1, \ldots, X_n)$, which equals roughly $\log_4 n$ when $X_i \overset{\text{IID}}{\sim}$ *Geo*$(1/4)$. This interval consists of two sub-intervals, one centered at $\eta = 1/4$ with width $\approx 10^{-n}$, and the other centered around $\eta = 3/4$ with width $\approx (\log_4 n)^{-1}$. The likelihood ratio between an η in the first sub-interval to one in the second is $\approx 9^n$. Hence the ratio of the posterior masses assigned to these two sub-intervals is $\log_4 n \cdot (0.9)^n \to 0$, i.e., the posterior attaches almost all its mass to the sub-interval centered at $3/4$. $\square$

Several remarks are due. The example is not really about a nonparametric model, but rather a parametric model that is highly irregular[1]. The choice of a uniform prior over $[1/8, 7/8]$ in presence of two singularity points is debatable – similar to the situations involving point null testing.

[1] e.g., the support of $f(x|\theta)$ changes with θ

Example 3. For a more standard nonparametric situation, consider the IID model with $X_i \overset{\text{IID}}{\sim} P$, $P \in \mathcal{P}(\mathcal{X})$, where we have assigned $P \sim \text{DP}(\alpha, G)$ for some fixed $\alpha > 0$ and $G \in \mathcal{P}(\mathcal{X})$ with $\text{supp}(G) = \mathcal{X}$. We know the posterior distribution of P given D^n is $\text{DP}(\alpha + n, (1 - w_n)G + w_n\mathbb{P}_n)$ where $w_n = n/(\alpha + n)$ and $\mathbb{P}_n$ denotes the empirical measure $(1/n)\sum_{i=1}^n \delta_{X_i}$ induced by the n observations. When the true data generating measure is $P = P_0$, we know that $d_W(\mathbb{P}_n, P_0) \to 0$ almost surely, from which it follows that $d_W(\Pi(\cdot|D^n), \delta_{P_0}) \to 0$ almost surely as well. □

In the above example, it is critical that $\text{supp}(G) = \mathcal{X}$, otherwise the conjugacy property of DP breaks down. Since $\text{supp}(P_0) \subset \mathcal{X}$ we have $\text{supp}(P_0) \subset \text{supp}(G)$. A fundamental result about Dirichlet process distributions is that every P_0 with $\text{supp}(P_0) \subset \text{supp}(G)$ belongs to the weak support of $\text{DP}(\alpha, G)$, for any $\alpha > 0$, i.e., any weak neighborhood U of P_0 receives positive mass from $\text{DP}(\alpha, G)$.

Example 4 (Diaconis and Freedman (1986))**.** In the above DP model, posterior consistency came rather cheap. But this is rather fragile. Consider data $D^n = (X_1, \ldots, X_n)$ modeled as $X_i = \theta + \epsilon_i$, $\epsilon_i \overset{\text{IID}}{\sim} P$, where $\theta \in \mathbb{R}$ is an unknown location parameter of interest, and P is an unknown, symmetric probability measure on $\mathbb{R}$. Consider the product prior $(\theta, P) \sim N(0, 1) \times \text{SDP}(\alpha, G_0)$, where $\text{SDP}(\alpha, G_0)$ denotes the symmetrized version of $\text{DP}(\alpha, G_0)$, i.e., a $P \sim \text{SDP}(\alpha, G_0)$ can be written as $P(A) = G(A) + G(-A)$ where $G \sim \text{DP}(\alpha, G_0)$. Assume G_0 is taken to be the Cauchy distribution, with density $g_0(y) = 1/\{\pi(1 + y^2)\}$.

Suppose the truth is given by $\theta = 0$ and $P = P_0$, where P_0 admits a density function p_0 that is compactly supported, symmetric about 0, infinitely differentiable and has a unique strict maximum at 0 but very small mass near 0, nearly 1/2 mass near each of the two other minor modes $\pm a$, $a > 1$. Then the posterior on θ given D^n, as $n \to \infty$, keeps switching between concentrating near $\pm\gamma$ where $\gamma = \sqrt{a^2 - 1}$, i.e., for any $\epsilon > 0$

$$\limsup_{n\to\infty} \Pi(|\theta - \gamma| < \epsilon \; |D^n) = \limsup_{n\to\infty} \Pi(|\theta + \gamma| < \epsilon \; |D^n) = 1,$$

with probability 1.

To see why this happens, notice that the model implies $\epsilon_i = \eta_i v_i$ with η_i being IID ± 1 with probability 1/2 and $v_i \overset{\text{IID}}{\sim} G$, $G \sim \text{DP}(\alpha, G_0)$. By the assumption on P_0, there are no ties in the data, and so, by symmetry of g_0,

$$p(D^n|\theta) \propto \prod_{i=1}^n g_0(X_i - \theta) \propto \exp[-\sum_{i=1}^n \log\{1 + (X_i - \theta)^2\}],$$

which leads to the result because the posterior on θ concentrates around the minimizer of $\int \log\{1 + (y - \theta)^2\} p_0(y)dy$, which keeps switching between $-\gamma$ and γ by the assumptions on p_0. □

In this case posterior inconsistency manifests even though $\text{SDP}(\alpha, G_0)$ includes P_0 in its weak support, i.e., gives positive mass to every weak neighborhood of P_0[2]. From

[2] Follows from the fact that the weak support of $\text{DP}(\alpha, G)$ consists of all probability measures G with $\text{supp}(G) \subset \text{supp}(G_0)$.

the "proof" above, it should be clear that the main problem is a bad choice of G_0 – because the inference on θ essentially takes place under the (misspecified) parametric model: $X_i = \theta + \epsilon_i$, $\epsilon_i \stackrel{\text{IID}}{\sim} G_0$. What is important is that even though we seemingly took care of the misspecification by specifying a SDP prior on the unknown P, it made absolutely no difference to estimating θ. One should use a DP prior, or any prior that induces discrete random measures, with extreme caution when modeling data that is likely coming from a non-atomic distribution.

2.4 The Schwartz theorem

In a breakthrough paper Schwartz (1965) provided general and nearly sharp sufficient conditions for posterior consistency in the IID setting of $D^n = (X_1, \ldots, X_n) \in \mathcal{X}^n$, with the model $X_i \stackrel{\text{IID}}{\sim} f$, $f \sim \Pi$, where Π is a probability measure on the set $\mathcal{F}$ of probability density functions wrt a given dominating measure, which we take to be the Lebesgue measure on $\mathcal{X}$. Let $d_{\text{KL}}(p, q) = \int p(x) \log\{p(x)/q(x)\}dx$ denote the Kullback-Leibler divergence. By a test function Φ_n we mean any $[0, 1]$-valued function of $\mathcal{X}^n$.

Suppose $f = f_0$ is the true density, and let P_0^∞ denote the joint density of $(X_1, X_2, \ldots)$ under f_0. We say that f_0 belongs to the KL support of Π if

$$\text{for every } \epsilon > 0,\ \Pi(\{f : d_{\text{KL}}(f_0, f) < \epsilon\}) > 0.$$

Theorem 4 (o). *If f_0 belongs to the KL support of Π and $U_n \subset \mathcal{F}$ are neighborhoods of f_0 such that there are test functions Φ_n, $n = 1, 2, \ldots$ satisfying*

$$\mathbb{E}_{f_0}\Phi_n \le Be^{-bn}, \quad \sup_{f \in U_n^c} \mathbb{E}_f(1 - \Phi_n) \le Be^{-bn}$$

for some $b, B > 0$, then $\Pi(U_n^c|D^n) \to 0$ almost surely $[P_0^\infty]$.

Before proving this theorem, we would look at a very useful consequence of this result. If we equip $\mathcal{F}$ with the weak convergence topology and the resulting metric (say the Lévy-Prokhorov metric, assuming $\mathcal{X}$ is separable), then the corresponding notion of posterior consistency at f_0 is referred to as *weak consistency* [more clearly, consistency in the weak topology].

Corollary 5. *If f_0 belongs to the KL support of Π then the posterior achieves weak consistency at f_0.*

Proof. Suffices to prove the condition of the Schwartz theorem for

$$U_n \equiv U = \left\{ f : \int \phi(x) f(x) dx < \int \phi(x) f_0(x) dx + \epsilon \right\}$$

for a given $\epsilon > 0$ and a continuous function $\phi : \mathcal{X} \to [0, 1]$, since these sets form a sub-base of the neighborhood system of f_0 under weak convergence. Take

$$\Phi_n(D^n) = I \left\{ \frac{1}{n} \sum_{i=1}^n \phi(X_i) > \int \phi(x) f_0(x) dx + \epsilon/2 \right\}.$$

By Hoeffding's inequality[3] $\mathbb{E}_{f_0}\Phi_n \le \exp\{-n\epsilon^2/2\}$. Also for any $f \notin U$, $\mathbb{E}_f(1-\Phi_n) \le \exp\{-n\epsilon^2/2\}$, again by Hoeffding's inequality[4]. Hence the condition of the Schwartz theorem is satisfied with $B = 1$, $b = \epsilon^2/2$. $\square$

This Corollary is a fairly useful result. When $\mathcal{X}$ is discrete, for which weak convergence topology matches with total variation topology, the above Corollary gives almost a complete characterization of posterior consistency. In particular when $\mathcal{X} = \mathbb{N}$, the situation considered in the first two examples of Section 2.3, we can say the posterior consistency is attained at every $\theta_0 \in \Delta_\infty$ in the KL support of Π, i.e., $\Pi\{\theta : \sum_j \theta_{0j} \log(\theta_{0j}/\theta_j) < \epsilon\} > 0$ for every $\epsilon > 0$. This readily generalizes the "finitely many non-zero element" condition of the first example in Section 2.3. Also, notice that in the counterexample by Freedman, the KL support condition fails at the $Geo(1/4)$ distribution.

Proving the KL support condition for more general cases, particularly when $\mathcal{X}$ is an interval in an Euclidean space, is not trivial, but some standard tools have emerged over the years, many relying on Taylor expansion of smooth functions and/or kernel convolution techniques.

Example 5. Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{F}$ = all probability density functions on $\mathbb{R}$ (with respect to the Lebesgue measure). Let Π be a mixture of normals prior on $\mathcal{F}$ given by the law of the random density function

$$f(x) = (\phi * P)(x) = \int \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) dP(\mu, \sigma), \quad x \in \mathbb{R},$$

where $P \sim \tilde{\Pi}$, a probability measure on $\mathcal{P}(\mathbb{R} \times (0, \infty))$. For example $\tilde{\Pi}$ could be $\mathrm{DP}(\alpha, G_0)$ with $G_0 = N(a, b^2) \times IG(r, s^2)$. If the true density is $f_0 = \phi * P_0$ with $\mathrm{supp}(P_0) \subset (-a, a) \times (\underline{\sigma}, \bar{\sigma})$ for some finite $a \ge 0$ and $0 < \underline{\sigma} < \bar{\sigma}$, and P_0 belongs to the weak support of $\tilde{\Pi}$ [true if $\tilde{\Pi}$ is the DP] then f_0 belongs to the KL support of Π.

This can be proved by using fairly elementary tools, helped by the facts that f_0 admits a second moment [in fact it admits a moment generating function], and that for any compact subsets $K \subset \mathbb{R} \times (0, \infty)$ and $A \subset \mathbb{R}$, the collection of functions $\{h_x : (\mu, \sigma) \mapsto \sigma^{-1}\phi(\sigma^{-1}(x-\mu)) : x \in A\}$ on K is uniformly equicontinuous and hence forms a compact subset in the supremum norm. See Ghosal et al. (1999, Theorem 3) and Tokdar (2006, Lemma 3.1) for more details. $\square$

Proof of Schwartz's Theorem. Since $\Phi_n(D^n) \in [0, 1]$ we can write

$$\Pi(U_n^c | D^n) \le \Phi_n + \frac{(1-\Phi_n)\int_{U_n^c} \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} d\Pi(f)}{\int_{\mathcal{F}} \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} d\Pi(f)}$$

[3] Hoeffding's inequality: If $Y_1, \ldots, Y_n$ are bounded then $P(\bar{Y} - \mathbb{E}\bar{Y} \ge t) \le \exp\{-2nt^2\}$. Apply this with $Y_i = \phi(X_i)$, $\mathbb{E}\bar{Y} = \mathbb{E}Y_1 = \int \phi(x) f_0(x) dx$.

[4] applied to $Y_i = -\phi(X_i)$

Since $\mathbb{E}_{f_0}\Phi_n \leq Be^{-bn}$, it follows from Borel-Canteli lemma that for any $\beta < b$, $P_0^\infty(\Phi_n > e^{-n\beta}$ infinitely often$) = 0$, i.e., $\Phi_n \to 0$ a.s. $[P_0^\infty]$ and the convergence to zero is exponentially fast.

To show that the same happens to the second term on the right hand side of the above display it suffices to show that

1. $\mathbb{E}_{f_0}[(1-\Phi_n)\int_{U_n^c}\prod_{i=1}^n\{f(X_i)/f_0(X_i)\}d\Pi(f)] \leq Be^{-bn}$, and
2. for every $\beta > 0$, $e^{n\beta}\int_{\mathcal{F}}\prod_{i=1}^n\{f(X_i)/f_0(X_i)\}d\Pi(f) \to \infty$ almost surely $[P_0^\infty]$.

The first assertion holds since by interchanging the expectation and the integral (by Fubini's theorem)

$$\mathbb{E}_{f_0}\left[(1-\Phi_n)\int_{U_n^c}\prod_{i=1}^n \tfrac{f(X_i)}{f_0(X_i)}d\Pi(f)\right] = \int_{U_n^c}\mathbb{E}_f(1-\Phi_n)d\Pi(f) \leq Be^{-bn}$$

by the assumption on Φ_n.

To prove the second assertion, let $K = \{f : d_{\mathrm{KL}}(f_0, f) < \beta\}$. Notice that

$$e^{n\beta}\int_{\mathcal{F}}\prod_{i=1}^n \tfrac{f(X_i)}{f_0(X_i)}d\Pi(f) \geq e^{n\beta}\int_{K}\prod_{i=1}^n \tfrac{f(X_i)}{f_0(X_i)}d\Pi(f)$$

and also that if $d_{\mathrm{KL}}(f_0, f) < \beta$ then

$$e^{n\beta}\prod_{i=1}^n \tfrac{f(X_i)}{f_0(X_i)} = \exp\left\{n\left(\beta - \tfrac{1}{n}\textstyle\sum_{i=1}^n \log\tfrac{f_0(X_i)}{f(X_i)}\right)\right\} \to \infty \text{ a.s. } [P_0^\infty]$$

by the strong law of large numbers. From which, by another application of Fubini's theorem[5], we may conclude $e^{n\beta}\int_K\prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)}d\Pi(f) \to \infty$ almost surely $[P_0^\infty]$. □

2.5 Extensions of Schwartz's theorem: use of sieves

When $\mathcal{F}$ is equipped with a stronger metric d, such as the total variation or the Hellinger metric, and we take $U_n \equiv U = \{f : d(f, f_0) \leq \epsilon\}$ for some $\epsilon > 0$, it is no longer possible (or at least easy) to construct tests Φ_n satisfying the requirements of the Schwartz's theorem.

To further probe existence of tests that can well separate a given f_0 from a set $V \subset \mathcal{F}$ based on n observations $X_1, \ldots, X_n$, define the minimax risk of testing as:

$$\pi_n(f_0, V) = \inf_{\Phi:\mathcal{X}^n\to[0,1]}\left\{\mathbb{E}_{f_0}\Phi + \sup_{f\in V}\mathbb{E}_f(1-\Phi)\right\},$$

[5] we need this because the null set where divergence to infinity does not hold may depend on f. However, if we let $E = \{(X^\infty, f) : e^{n\beta}\prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} \to \infty\}$ and take Π_K to be the restriction of Π to K then $(P_0^\infty \times \Pi_K)(E) = 1$ and hence for almost every X^∞ (under P_0^∞), $\Pi_K(\{f : e^{n\beta}\prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} \to \infty\}) = 1$.

which is same as the sum total of the type I and maximum type II error probabilities for testing $H_0 : f = f_0$ vs. $H_1 : f \in V$. It follows from somewhat first-principles calculations that

$$\pi_1(f_0, V) = 1 - \inf_{f \in \mathrm{conv}(V)} \frac{1}{2}\|f_0 - f\|_1 \le \sup_{f \in \mathrm{conv}(V)} \rho_{1/2}(f_0, f) =: \rho_{1/2}(f_0, \mathrm{conv}(V)) \quad (2)$$

$$\pi_n(f_0, V) \le \rho_{1/2}^n(f_0, \mathrm{conv}(V)) \quad (3)$$

where $\rho_{1/2}(f_1, f_2) = \int \sqrt{f_1(x) f_2(x)}dx = 1 - \frac{1}{2}d_H^2(f_1, f_2)$ denotes the Hellinger affinity between f_1 and f_2. So, if $d_H(f_0, \mathrm{conv}(V)) := \inf_{f \in \mathrm{conv}(V)} d_H(f_0, f) > \epsilon$ then there exists a test $\Phi_{n,V}$ such that

$$\mathbb{E}_{f_0}\Phi_{n,V} \le e^{-n\epsilon^2}, \quad \sup_{f \in V} \mathbb{E}_f(1 - \Phi_{n,V}) \le e^{-n\epsilon^2}$$

since $\rho_{1/2}(f_0, \mathrm{conv}(V)) < (1 - \epsilon^2/2)^n \le e^{-n\epsilon^2}$.

So if $U^c \subset V_1 \cup \cdots \cup V_N$, for some $N < \exp(\xi n\epsilon^2)$ with $\xi \in (0, 1/2)$, where each V_j is convex with $d_H(f_0, V_j) > \epsilon$, then the test

$$\Phi_n = \max_{1 \le j \le N} \Phi_{n,V_j}$$

satisfies the requirements of the Schwartz theorem with $B = 1$, $b = (1/2 - \xi)\epsilon^2$. Unfortunately, for $U = \{f : d(f, f_0) \le \epsilon\}$ with d being L_1 or Hellinger, one cannot contain U^c within such a finite intersection. To handle such cases, one can extend the Schwartz theorem by using a sequence of compact subsets $\mathcal{F}_n \subset \mathcal{F}$ as follows.

Proposition 6. *Suppose the metric d on $\mathcal{F}$ is bounded from above by d_H. Fix any $\epsilon > 0$. If there exist a sequence of compact subsets $\mathcal{F}_n \subset \mathcal{F}$ and constants $\delta > 0$, $0 < \xi < 1/2$, $C > 0$ such that*

1. $\log N(\delta, \mathcal{F}_n, d) \le \xi n\epsilon^2$
2. $\Pi(\mathcal{F}_n^c) \le e^{-Cn}$

then $\Pi(\{f : d(f_0, f) > \epsilon\}|D^n) \to 0$ a.s. $[P_0^\infty]$ for every f_0 in the KL support of Π.

Proof. Fix an f_0 in the KL support of Π and let $U = \{f : d(f_0, f) \le \epsilon\}$. Clearly $U^c \cap \mathcal{F}_n$ can be contained in some $V_1 \cup \cdots \cup V_N$ where $N \le e^{\xi n\epsilon^2}$, each V_j is convex and $d_H(f_0, V_j) \ge d(f_0, V_j) > \epsilon$. So there exists a test Φ_n satisfying

$$\mathbb{E}_{f_0}\Phi_n \le e^{-bn} \text{ and } \sup_{f \in U^c \cap \mathcal{F}_n} \mathbb{E}_f(1 - \Phi_n) \le e^{-bn}$$

As in the proof of the Schwartz theorem, we write

$$\Pi(U^c|D^n) \le \Phi_n + \frac{(1 - \Phi_n)\int_{U^c \cap \mathcal{F}_n} \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} d\Pi(f)}{\int_{\mathcal{F}} \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} d\Pi(f)} + \frac{\int_{\mathcal{F}_n^c} \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} d\Pi(f)}{\int_{\mathcal{F}} \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} d\Pi(f)}.$$

As shown in the proof of the Schwartz theorem, the first two terms decay to 0 exponentially fast almost surely $[P_0^\infty]$. To show the same for the last term, we only need to show that $\mathbb{E}_{f_0} \int_{\mathcal{F}_n^c} \prod_{i=1}^n \{f(X_i)/f_0(X_i)\} d\Pi(f) \leq e^{-b'n}$ for some $b' > 0$. This is trivially true because by exchanging the expectation and the integration, this term precisely equals $\Pi(\mathcal{F}_n^c)$. □

Getting hold of the sequence $\{\mathcal{F}_n\}$, often referred to as a sieve, requires some understanding of the function space $\mathcal{F}$. Here are some examples for general function spaces $\mathcal{F}$, not necessarily a space of probability measures.

Example 6. For any $d \in \mathbb{N}$ and $\epsilon > 0$, $\log N(\epsilon, \Delta_d, \|\cdot\|_1) \leq (d-1)\log(5/\epsilon)$. □

So, the ϵ-entropy of Δ_d, as for any compact Euclidean subset, grows logarithmically in $1/\epsilon$. For nonparametric spaces, the growth rate could be much faster.

Example 7. The Hölder norm of order $\alpha > 0$ of a continuous function $f : \mathcal{X} \to \mathbb{R}$ on abounded subset $\mathcal{X} \subset \mathbb{R}^p$ is defined as

$$\|f\|_\alpha := \max_{k \in \mathbb{N}_0^p : k. < \lfloor \alpha \rfloor} \sup_{x \in \mathcal{X}} |D^k f(x)| + \max_{k \in \mathbb{N}_0^p : k. = \lfloor \alpha \rfloor} \frac{|D^k f(x) - D^k f(y)|}{\|x - y\|^{\alpha - \lfloor \alpha \rfloor}}$$

where $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$, $k. = k_1 + \ldots + k_p$ for $k \in \mathbb{N}_0^p$, $\lfloor \alpha \rfloor$ is the largest integer strictly smaller than α and

$$D^k = \frac{\partial^{k.}}{\partial x_1^{k_1} \cdots \partial x_p^{k_p}}, \quad k = (k_1, \ldots, k_p) \in \mathbb{N}_0^p$$

is the mixed partial derivative of order (multi-index) k. It turns out that there exists a constant $K = K_{\alpha,p}$ such that

$$\log N(\epsilon, \{f : \|f\|_\alpha \leq M\}, \|\cdot\|_\infty) \leq K \mathrm{vol}(\mathcal{X})(M/\epsilon)^{p/\alpha}.$$

Notice that the entropy is calculated with respect to the supremum norm. A proof can be given by using Taylor's expansion. □

Example 8. For a give $d \in \mathbb{N}$, let ϕ_σ denote the $N(0, \sigma^2 \mathrm{I}_d)$ pdf on $\mathbb{R}^d$. For any probability measure P on $\mathbb{R}^d$, let $\phi_{P,\sigma}$ denote the normal location mixture density:

$$\phi_{P,\sigma}(y) = \int \phi_\sigma(y - \mu) dP(\mu), \ y \in \mathbb{R}^d.$$

For any $\epsilon > 0$, $H, M \in \mathbb{N}$, and $a, \underline{\sigma} > 0$ define

$$\mathcal{G} = \left\{ \phi_{P,\sigma} : P = \sum_{h=1}^\infty \pi_h \delta_{\mu_h}; \mu_h \in [-a, a]^d, h \leq H; \sum_{h > H} \pi_h < \epsilon; 1 < \frac{\sigma}{\underline{\sigma}} < (1+\epsilon)^M \right\}$$

then

$$\log N(\epsilon, \mathcal{G}, \|\cdot\|_1) \leq dH \log \frac{a}{\underline{\sigma}\epsilon} + H \log \frac{1}{\epsilon} + \log M.$$

The set $\mathcal{G}$ essentially contains P that are well approximated by a discrete measure with $H+1$ atoms. Each of the first H atoms μ_h need to be put into a $\underline{\sigma}\epsilon$ covering of $[-a, a]^p$, giving rise to the first term on the bound above. The $(H+1)$-dimensional simples contributes the second term to the entropy calculation. The last term comes from an ϵ-covering the interval $(\underline{\sigma}, \underline{\sigma}(1+\epsilon)^M)$ in the logarithmic scale. For a complete proof see http://arxiv.org/abs/1111.4148, and also see Shen et al. (2013) for extension where the spherical normal kernel is replaced with a $N(0, \Sigma)$, and the last condition on the sieve is stated in terms of eigenvalues of Σ. □

3 Posterior contraction rates

3.1 Definition and basic theorem

When the posterior contracts to the truth, it is possible to quantify the rate of this convergence, or at least to find useful bounds on this rate. Again we consider the parameter space to be a metric space (Θ, d) and the truth to be $\theta = \theta_0 \in \Theta$.

Definition 2. The posterior $\Pi_n(\cdot|D^n)$ is said to contract to δ_{θ_0} at the rate $\epsilon_n \to 0$ (or faster) if for every $M_n \to \infty$, $\Pi_n(\{\theta : d(\theta, \theta_0) > M_n\epsilon_n|D^n\}) \to 0$ in probability when $D^n \sim p_n(\cdot|\theta_0)$.

The extra sequence $M_n \to \infty$ is needed sometimes for technical convenience. Note that $M_n \to \infty$ very slowly. In fact, for many nonparametric methods, a constant $M_n \equiv M$ suffices. While the definition only quantifies a bound on the contraction rate, this is often because a bound on the converse side could be derived from the minimax estimation theorem. Suppose the posterior contracts at a rate $\epsilon_n \to 0$ or faster at every $\theta_0 \in \Theta_0 \subset \Theta$. Then there exists an estimator $\hat{\theta}_n$ that is consistent for $\theta \in \Theta_0$, and converges at least as fast as ϵ_n. However, we know that no estimator can converge faster than the minimax estimation error rate for Θ_0. Therefore, if ϵ_n is close to the minimax rate of Θ_0, then it is indeed a sharp quantification of the posterior contraction rate.

Another extension of Schwartz's theorem provides sufficient conditions to find contraction rates. Again, we restrict to the IID case, with $D^n = (X_1, \ldots, X_n)$, $X_i \overset{\text{IID}}{\sim} f(x_i|\theta)$. Let $K(\theta_0; \theta) = d_{\text{KL}}(f(\cdot|\theta_0), f(\cdot|\theta)) = \mathbb{E}_{\theta_0} \log\{f(X_1|\theta_0)/f(X_1|\theta)\}$ and $V(\theta_0; \theta) = \mathbb{E}_{\theta_0} \log^2\{f(X_1|\theta_0)/f(X_1|\theta)\}$.

Theorem 7. *Let $\epsilon_n \to 0$ such that $n\epsilon_n^2 \to \infty$ and there exist sets $\Theta_n \subset \Theta$, $n \geq 1$ and constants $c_1, c_2 > 0$ satisfying*

1. $\log N(\epsilon_n, \Theta_n, d) \leq c_1 n\epsilon_n^2$
2. $\Pi(\Theta_n^c) \leq e^{-(4+c_2)n\epsilon_n^2}$.

Then the posterior $\Pi(\cdot|D^n)$ contracts at the rate ϵ_n or faster at every θ_0 satisfying

$$\Pi\left(\{\theta : K(\theta_0; \theta) < \epsilon_n^2, V(\theta_0; \theta) < \epsilon_n^2\}\right) \geq e^{-c_2 n\epsilon_n^2}.$$

3.2 Two applications

3.2.1 Density estimation with DP location mixture of multivariate normals

Consider estimating a density function on $\mathbb{R}^d$ from IID data $X_1, \ldots, X_n$. For any $d \times d$ positive definite matrix Σ, let ϕ_Σ denote the pdf of the $N(0, \Sigma)$ distribution and for any probability measure P on $\mathbb{R}^d$, define

$$\phi_{P,\Sigma}(x) = \int_{\mathbb{R}^d} \phi_\Sigma(x - \mu) dP(\mu).$$

Let Π denote the probability law of the random pdf $\phi_{P,\Sigma}$ when $(P, \Sigma) \sim \mathrm{DP}(\alpha, N(m, S)) \times \mathrm{IW}(r, \Sigma_0)$ for some m, S, r and Σ_0.

Consider the model $X_i \overset{\text{IID}}{\sim} f$, $f \sim \Pi$. We can characterize the posterior contraction rate at a true f_0 by some basic smoothness and tail properties of f_0. For any $\beta > 0$, $\tau_0 \geq 0$ and $L : \mathbb{R}^d \to \mathbb{R}_+$, define the locally β-Hölder class with envelope L, denoted $\mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$, to be the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ with finite mixed partial derivatives $D^k f$ ($k \in \mathbb{N}_0^d$) of all orders up to $k. \leq \lfloor \beta \rfloor$, and for every $k \in \mathbb{N}_0^d$ with $k. = \lfloor \beta \rfloor$ satisfying

$$|(D^k f)(x + y) - (D^k f)(x)| \leq L(x) e^{\tau_0 \|y\|^2} \|y\|^{\beta - \lfloor \beta \rfloor}, \quad x, y \in \mathbb{R}^d. \tag{4}$$

Theorem 8. *Suppose that $f_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$ is a probability density function satisfying*

$$P_0 \left(|D^k f_0|/f_0\right)^{(2\beta+\epsilon)/k.} < \infty, \quad k \in \mathbb{N}_0^d, k. \leq \lfloor \beta \rfloor, \qquad P_0 \left(L/f_0\right)^{(2\beta+\epsilon)/\beta} < \infty \tag{5}$$

for some $\epsilon > 0$ where $P_0 g = \int g(x) f(x) dx$ denotes expectation of $g(X)$ under $X \sim f_0$. Also suppose there are positive constants a, b, c, τ such that

$$f_0(x) \leq c \exp(-b\|x\|^\tau), \quad \|x\| > a. \tag{6}$$

Then with Π as in above, the posterior contracts at f_0, in the Hellinger or the L_1 metric, with rate $\epsilon_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$, where $t = \{d^*(1 + 1/\tau + 1/\beta) + 1\}/(2 + d^*/\beta)$ and $d^* = \max(d, 2)$.

See Shen et al. (2013) for a proof. The sieve described in Example 8 works, with suitably chosen parameters. To show the (augmented) KL support condition of Theorem 7, one needs to show that for all small $\sigma > 0$, one can approximate f_0 by a $\phi_{P,\sigma^2 \mathrm{I}_d}$ with an approximation error of the order of σ^β. A good choice for P is the probability measure P_0 associated with the pdf f_0 itself! Notice that, $\phi_{P_0,\sigma^2 \mathrm{I}_d} \to f_0$ pointwise as $\sigma \to 0$. However the resulting approximation error is $\asymp \sigma$, which decays at a slower rate in $\sigma \to 0$ relative to σ^β when $\beta > 1$. It turns out that the signed measure P_1 associated with the function

$$f_1 = f_0 - \sum_{\substack{k \in \mathbb{N}_0^d \\ 1 \leq k. \leq \lfloor \beta \rfloor}} d_k \sigma^{k.} D^k f_0$$

provides the correct approximation order of σ^β, where the numbers d_k, $k \in \mathbb{N}_0^d$ are found recursively as follows: if $k. = 1$, set $c_k = 0$ and $d_k = -m_k/k!$ and for $k. \geq 2$ define

$$c_k = -\sum_{\substack{k=l+m \\ l.\geq 1,\ m.\geq 1}} \frac{(-1)^{m.}}{m!}\mu_m d_l, \qquad d_k = \frac{(-1)^{k.}\mu_k}{k!} + c_k. \tag{7}$$

where for any $k \in \mathbb{N}_0^d$, $\mu_k = \int y^k \phi_{\mathrm{I}_d}(y)dy$ is the k-th moment of the d dimensional standard normal distribution, $y^k := y_1^{k_1} \cdots y_d^{k_d}$, $k! := k_1! \cdots k_d!$. Further calculations show that P_1 can be replaced with a probability measure P_2 without affecting the approximation order.

Several comments are due on the statement of Theorem 8. The stated rate ϵ_n, without the $(\log n)^t$ term is the minimax estimation error rate for the β-Hölder class (Yang and Barron, 1999) when $d \geq 1$. And hence the DP mixture model offers nearly optimal estimation on such a class. However, since the same happens for every $\beta > 0$, the resulting method is "adaptive" – the same prior specification works for all smoothness classes and the posterior automatically adapts to the correct smoothness level. This is fundamental since smoothing based classical nonparametric methods require some external help with the bandwidth selection to be able to adapt to the correct smoothness level. When $d = 1$, the inverse-Wishart (now inverse-Gamma) prior on $\Sigma = ((\sigma^2))$ cannot be shown to give the optimal rate (it might, but the current proof technique does not work), but an inverse-Gamma prior on σ can be shown to work.

Also, same posterior contraction rates apply for the simpler DP mixture prior where one restricts $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, with inverse-Gamma priors on each σ_j. In fact, the even simpler model where $\Sigma = \sigma^2 \mathrm{I}_d$, with an inverse gamma prior on $\sigma^{\min(2,d)}$ also gives the same contraction rates. But clearly finite sample properties of these priors will be quite different. When working in asymptopia[6], some of these important differences get absorbed in the constants leading the contraction rate.

Theorem 8 follows the conventional path of characterizing posterior contraction rates by the smoothness class the true density belongs to. This is somewhat artificial – but there is very little available in terms of alternative formulations that are more relevant to statistical modeling.

3.2.2 Gaussian process regression

Consider paired data $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, $i = 1, \ldots, n$ where the conditional behavior of Y_is given X_is is modeled through the nonparametric regression model:

$$Y_i = f(X_i) + \epsilon_i,\ \epsilon_i \overset{\mathrm{IID}}{\sim} N(0, \sigma^2)$$

with $(f, \sigma) \in C(\mathcal{X}) \times \mathbb{R}_+$ unknown. We assume $\mathcal{X}$ is a compact subset of $\mathbb{R}^p$. Consider the following GP prior Π on f:

$$f|\psi \sim \mathrm{GP}(0, C(\cdot, \cdot|\psi)), \quad \psi^p \sim Ga(a, b),$$

[6]a term coined by David Pollard

where $C(s,t) = \exp\{-\psi^2\|s-t\|^2\}$ is the isotropic square-exponential covariance function with a single scalar correlation-range parameter $\psi > 0$. Also suppose that the prior on σ is a density H on $\mathbb{R}_+$ with $\text{supp}(H) \subset [c,d]$ where $0 < c < d < \infty$ [this technically rules out the inverse Gamma prior, but one can use a truncated version of it with a very small c and a very large d].

For posterior contraction at a true (f_0, σ_0), there are several possible choices. A standard one (van der Vaart and van Zanten, 2008, 2009) is to treat the X_is as fixed design points and take the (stochastic) metric: $d((f,\sigma),(f_0,\sigma_0)) = \|f - f_0\|_n + |\sigma - \sigma_0|$ where $\|f\|_n = [(1/n)\sum_{i=1}^n |f(X_i) - f_0(X_i)\}^2]^{1/2}$ is the empirical L_2 norm based on the observed predictors. This requires extending Theorem 7 to the independent but not-identically distributed case, as done in Ghosal and Vaart (2007). One can also consider a stochastic design situation, where $X_i \overset{\text{IID}}{\sim} q$, a pdf on $\mathcal{X}$ that is bounded from above [otherwise arbitrary] and take d to be the Hellinger distance between the joint densities of (X_i, Y_i) induced by (f, σ, q) and (f, σ, q). This distance also characterizes average prediction error at a new X^* drawn from q. In either scenario, we can get minimax optimal posterior contraction rates (up to $\log n$ terms) as stated below. In the following, let $C^\alpha(\mathcal{X})$ denote the class of all continuous functions with finite Hölder norm of order α [see Example 7 for definition].

Theorem 9. *If* $f_0 \in C^\alpha(\mathcal{X})$ *and* $\sigma_0 \in \text{supp}(H)$ *then* $\Pi(\{(f,\sigma) : d((f,\sigma),(f_0,\sigma_0)) > \epsilon_n\}|D^n) \to 0$ *in* P_0^∞ *probability with* $\epsilon_n = n^{-1/(2+d/\alpha)}(\log n)^t$ *where* $t = 1 - 1/(2 + 4\alpha/d)$.

Once again, without the $(\log n)^t$, the rate is the minimax rate of estimation for $C^\alpha(\mathcal{X})$ (Yang and Barron, 1999). Since α in the theorem is arbitrary, we again see that the single GP regression method automatically adapts to the optimal contraction rate for every Hölder smoothness class – without any further user intervention or external *a priori* knowledge of the the smoothness of f_0.

Several extensions of Theorem 9 now exist, including anisotropic covariance function (Bhattacharya et al., 2014), and also to the case of large p small n regression where the GP regression is augmented with variable selection prior (Yang and Tokdar, 2015).

References

Bhattacharya, A., D. Pati, and D. Dunson (2014). Anisotropic function estimation using multi-bandwidth Gaussian processes. *Annals of statistics 42*(1), 352–381.

Diaconis, P. and D. Freedman (1986). On the consistency of bayes estimates. *The Annals of Statistics 14*(1), 1–26.

Freedman, D. A. (1963). On the asymptotic behavior of bayes' estimates in the discrete case. *The Annals of Mathematical Statistics 34*(4), 1386–1403.

Ghosal, S., J. K. Ghosh, R. Ramamoorthi, et al. (1999). Posterior consistency of dirichlet mixtures in density estimation. *The Annals of Statistics 27*(1), 143–158.

Ghosal, S. and V. D. Vaart (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics 35*(1), 192–223.

Schwartz, L. (1965). On bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 4*(1), 10–26.

Shen, W., S. T. Tokdar, and S. Ghosal (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika 100*(3), 623–640.

Tokdar, S. T. (2006). Posterior consistency of dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics 67*(4), 90–110.

van der Vaart, A. and J. van Zanten (2008). Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics 36*(3), 1435–1463.

van der Vaart, A. W. and J. H. van Zanten (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *The Annals of Statistics 37*(5B), 2655–2675.

Yang, Y. and A. Barron (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics 27*, 1564–1599.

Yang, Y. and S. T. Tokdar (2015). Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics 43*(2), 652–674.