A short note on the Frequentist consistency of a nonparametric Bayes

Yen-Chi Chen
University of Washington
May 19, 2020

This note is to explain Proposition 11 of

> [VV2011] Van Der Vaart, A., & Van Zanten, H. (2011). Information rates of nonparametric Gaussian process methods. Journal of Machine Learning Research, 12(60), 2095-2119,

which is a simplified version of the grand theorem (Theorem 2.1) in the following paper:

> [GV2000] Ghosal, S., Ghosh, J. K., & Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. Annals of Statistics, 28(2), 500-531.

Note: I often call this theorem the fundamental theorem of Frequentist consistency of a nonparametric Bayes estimator.

Informally speaking, the proposition we will be discussing (see below Theorem 1) provides *sufficient conditions on establishing the rate of how the posterior distribution concentrates around the true model that generates the data under a fixed-design regression problem.*

Consider a simple fixed-design regression problem where we observe

$$(X_1, Y_1), \cdots, (X_n, Y_n)$$

from the model

$$Y_i = f_0(X_i) + \varepsilon_i,$$

where $\varepsilon_1, \cdots, \varepsilon_n \sim N(0, \sigma^2)$ are IID and $f_0(x) = \mathbb{E}(Y|X = x)$ is the true regression function. The covariates $X_1, \cdots, X_n$ are univariate and non-random.

The goal is to make inference of f_0. To simplify the notations, we denote $\mathcal{D}_n = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ as the data.

Suppose we use a Bayesian approach (nonparametric Bayes) such that we place a prior distribution $\Pi(f)$ of the regression function. Similar to the regular Bayesian problem, we denote $\Pi(f|\mathcal{D}_n)$ as the posterior distribution of f after observing the data. By the normal model of Y_i given $f(X_i)$, the posterior

$$d\Pi(f|\mathcal{D}_n) \propto \prod_{i=1}^{n} \phi\left(\frac{Y_i - f(X_i)}{\sigma}\right) d\Pi(f),$$

where $\phi(x)$ is the PDF of a standard normal distribution. We also denote

$$\phi_{n,f}(\mathcal{D}_n) = \prod_{i=1}^{n} \phi\left(\frac{Y_i - f(X_i)}{\sigma}\right).$$

Suppose we treat the nonparametric Bayes approach as a Frequentist procedure to construct an estimator. Then a natural way to discuss the consistency of such a procedure is to study the concentration of posterior distribution around f_0, the true data-generating process. In particular, we will consider the L_q concentration:

$$R_{q,n} = \mathbb{E}\left(\int \|f - f_0\|_n^q d\Pi(f|\mathcal{D}_n)\right),$$

where

$$\|f - f_0\|_n^q = \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - f_0(X_i)|^q.$$

A common choice is $q = 2$, leading to the L_2 concentration rate.

$R_{q,n}$ conveys the information on how fast the posterior distribution concentrates around the true data-generating function f_0. Note that the $R_{2,n}$ rate can be converted into a bound on the L_2 distance between the posterior mean and f_0, i.e.,

$$\mathbb{E}(\|\mathbb{E}(f|\mathcal{D}_n) - f_0\|_n^2) \leq R_{2,n}.$$

We are interested in the conditions of the prior Π such that the posterior distribution concentrates at a good rate. The Proposition 11 of [VV2011] provides a simple description on this purpose.

Let $D(\varepsilon, \mathcal{F}, \|\cdot\|_n)$ be an ε-packing number. Namely, the maximal number of elements in $\mathcal{F}$ such that pairwise distance under $\|\cdot\|_n$ norm is greater than or equal to ε.

Theorem 1 (Proposition 11 of Van Der Vaart and Van Zanten (2011)) *Consider the fixed design normal model as described in the above. Suppose the prior distribution satisfies the following two conditions:*

(F) *For some* $\varepsilon > 0, \sqrt{n}\varepsilon > 1$ *and for every* $r > 1$ *(integer), there exists a class of functions* $\mathcal{F}_r$ *such that*

$$D(\varepsilon, \mathcal{F}_r, \|\cdot\|_n) \leq \exp(n\varepsilon^2 r^2)$$
$$\Pi(\mathcal{F}_r) \geq 1 - \exp(-2n\varepsilon^2 r^2).$$

(T) $\Pi(f : \|f - f_0\|_n \leq \varepsilon) \geq \exp(-n\varepsilon^2)$.

Then

$$R_{q,n} = \mathbb{E}\left(\int \|f - f_0\|_n^q d\Pi(f|\mathcal{D}_n)\right) \leq C_0 \varepsilon^q$$

for some universal constant C_0.

The condition (F) requires that the prior has to put enough probability on a 'nice class' $\mathcal{F}_r$. The nice here means that $\mathcal{F}_r$ has a regular packing number so that it does not diverges too quickly. So in a sense, the prior places most of the mass on a smaller subset of the entire function space. Note that we do not require f_0 to be in any of these $\mathcal{F}_r$.

The only link that we need to connect our prior Π to f_0 is the second condition (T). This condition essentially requires that when we are thinking of an ε-ball around f_0 (and ε is shrinking), we have a sufficient among of mass ($e^{-n\varepsilon^2}$).

1 Two useful lemmas

We will prove this theorem with a small modification from the original proof in [VV2011] to make it more accessible. First of all, we recall two lemmas from [VV2011]. Note that we will not prove these two lemmas; interested readers can read [VV2011] for the proof of these two lemmas.

Lemma 2 (Normal Tests; Lemma 13 of Van Der Vaart and Van Zanten (2011)) . *Consider a multivariate normal $Z \sim N(\theta, \mathbf{I}) \in \mathbb{R}^n$ and $\theta \in \Theta \subset \mathbb{R}^n$. Let $M(s,\Theta)$ be the maximal number of points in Θ such that any pairwise distance is at least s. There exists a test $\psi(Z) \in \{0,1\}$ ($\psi(Z) = 1$ means rejecting H_0) such that for any $s > 1$ and $j \in \mathbb{N}$, we have*

$$\textit{(Type-1 error)} \quad \mathbb{E}(\psi(Z); Z \sim N(\theta_0, \mathbf{I})) \leq 9M(s/2,\Theta)\exp\left(-\frac{s^2}{8}\right),$$

$$\textit{(Type-2 error)} \quad \sup_{\theta\in\Theta; \|\theta-\theta_0\|\geq js} \mathbb{E}(1-\psi(Z); Z \sim N(\theta, \mathbf{I})) \leq \exp\left(-\frac{j^2 s^2}{8}\right).$$

One can think of this lemma as constructing location tests $H_0 : \theta = \theta_0$ against $H_{a,\ell} : \theta = \theta_\ell$ where θ_ℓ is the ℓ-th center in the packing set corresponding to the packing number $D(s/2,\Theta)$. And we pull all these tests together to form a final test. Thus, the type-1 error statement is essentially a Bonferroni correction to make sure if the data in deed from θ_0, then it is unlikely to falsely reject it. So this gives one side of 'concentration' around the true data-generating parameter. The second statement (type-2 error) shows that if the data is not coming from θ_0, then the worst case type-2 error is still small.

Lemma 3 (Density Ratio ; Lemma 14 of Van Der Vaart and Van Zanten (2011)) . *Consider a multivariate normal $Z \sim N(\theta, \mathbf{I}) \in \mathbb{R}^n$ and $\theta \in \Theta \subset \mathbb{R}^n$. For any prior distribution $Q(\theta)$ on $\mathbb{R}^n$ and any $s > 0$,*

$$\mathbb{E}\left(\int \frac{\phi_{n,\theta}(Z)}{\phi_{n,\theta_0}(Z)} dQ(\theta) \geq e^{-s^2} Q(\theta : \|\theta-\theta_0\| < s); Z \sim N(\theta_0, \mathbf{I})\right) \geq 1 - \exp\left(-\frac{s^2}{8}\right).$$

The power of this lemma is to construct a density ratio bound. This implies a lower bound on the 'normalizing constant' in the posterior. To see this, suppose the inequality is true (with a high probability), we have

$$\int \frac{\phi_{n,\theta}(Z)}{\phi_{n,\theta_0}(Z)} dQ(\theta) = \frac{1}{\phi_{n,\theta_0}(Z)} \int \phi_{n,\theta}(Z) dQ(\theta) \geq e^{-s^2} Q(\theta : \|\theta-\theta_0\| < s).$$

Thus, we have

$$\Omega_0 = \int \phi_{n,\theta}(Z) dQ(\theta) \geq e^{-s^2} Q(\theta : \|\theta-\theta_0\| < s)\phi_{n,\theta_0}(Z).$$

The posterior

$$dQ(\theta|Z) \propto \phi_{n,\theta}(Z) d\Pi(\theta)$$

and

$$dQ(\theta|Z) = \frac{1}{\Omega_0}\phi_{n,\theta}(Z) d\Pi(\theta).$$

Thus, for any set $B \subset \mathbb{R}^n$, the posterior probability will be

$$Q(B|Z) = \frac{\int_B \phi_{n,\theta}(Z) dQ(\theta)}{\int \phi_{n,\theta}(Z) dQ(\theta)} \leq \frac{\int_B \phi_{n,\theta}(Z) dQ(\theta)}{e^{-s^2} Q(\theta : \|\theta - \theta_0\| < s) \phi_{n,\theta_0}(Z)} \tag{1}$$

when the inequality of Lemma 3 holds (with a high probability). Equation (1) shows a clear way that we can control the posterior probability using prior probability within a small ball $Q(\theta : \|\theta - \theta_0\| < s)$.

2 Proof of Theorem 1

The proof consists of several stages although the original proof in Van Der Vaart and Van Zanten (2011) is very concise. I reordered parts of the proofs and added more details to make the proof more accessible to readers who are not familiar with the process.

Converting into radius integral. Using the fact that for a positive random variable T,

$$\mathbb{E}(T^q) = \int_0^\infty q \cdot s^{q-1} P(T > s) ds,$$

we can rewrite

$$\begin{aligned} \int \|f - f_0\|_n^q d\Pi(f|\mathcal{D}_n) &= \int_0^\infty q \cdot s^{q-1} \Pi(f : \|f - f_0\|_n > s|\mathcal{D}_n) ds \\ &= (4\varepsilon)^q \int_0^\infty q \cdot r^{q-1} \Pi(f : \|f - f_0\|_n > 4\varepsilon r|\mathcal{D}_n) dr. \end{aligned}$$

Thus, the moment bound

$$\begin{aligned} R_{q,n} &= \mathbb{E}\left((4\varepsilon)^q \int_0^\infty q \cdot r^{q-1} \Pi(f : \|f - f_0\|_n > 4\varepsilon r|\mathcal{D}_n) dr \right) \\ &= \mathbb{E}\left((4\varepsilon)^q \int_0^2 q \cdot r^{q-1} \Pi(f : \|f - f_0\|_n > 4\varepsilon r|\mathcal{D}_n) dr + (4\varepsilon)^q \int_2^\infty q \cdot r^{q-1} \Pi(f : \|f - f_0\|_n > 4\varepsilon r|\mathcal{D}_n) dr \right) \\ &\leq (8\varepsilon)^q + (4\varepsilon)^q \mathbb{E}\left(\int_2^\infty q \cdot r^{q-1} \Pi(f : \|f - f_0\|_n > 4\varepsilon r|\mathcal{D}_n) dr \right). \end{aligned}$$

The first term is already at the correct rate, so we only need to make sure that the second term is also in the same order. Accordingly, what we need is

$$\mathbb{E}\left(\int_2^\infty q \cdot r^{q-1} \Pi(f : \|f - f_0\|_n > 4\varepsilon r|\mathcal{D}_n) dr \right) < \infty \tag{2}$$

An interesting note is that by the Fubini's theorem, we can move the expectation into the integral, leading to the requirement

$$\int_2^\infty q \cdot r^{q-1} \mathbb{E}(\Pi(f : \|f - f_0\|_n > 4\varepsilon r|\mathcal{D}_n)) dr < \infty. \tag{3}$$

Here is an interesting note: in Lemma 2 and 3, all these probability bounds we have is an exponential (in fact, Gaussian) concentration (in terms of r). So if we plug them into the integral in equation (3), the integral will be finite. This is a key reason for the above construction.

Our strategy is to bound the posterior $\mathbb{E}(\Pi(f : \|f - f_0\|_n > 4\varepsilon r|\mathcal{D}_n))$ by 4 terms:

$$\begin{aligned}\Pi(f : \|f - f_0\|_n > 4\varepsilon r|\mathcal{D}_n) \le \psi(\mathcal{D}_n) + I(\mathcal{D}_n \in \mathcal{A}^c) + I(\mathcal{D}_n \in \mathcal{A})\Pi(f \notin \mathcal{F}_r) \\ + \Pi(f \in \mathcal{F}_r : \|f - f_0\|_n > 4\varepsilon r|\mathcal{D}_n)(1 - \psi(\mathcal{D}_n))I(\mathcal{D}_n \in \mathcal{A}),\end{aligned}$$

where $\psi = \psi(\mathcal{D}_n) \in \{0, 1\}$ is any test and $\mathcal{A}$ is the event

$$\mathcal{A} = \left\{ D_n : \int \frac{\phi_{n,f}(\mathcal{D}_n)}{\phi_{n,f_0}(\mathcal{D}_n)} d\Pi(\theta) \ge e^{-n\varepsilon^2 r^2} \Pi(f : \|f - f_0\|_n < \sqrt{n}\varepsilon r) \right\} \tag{4}$$

in Lemma 3 with $s = \sqrt{n}\varepsilon r$.

With this, we can bound the expectation using

$$\begin{aligned}\mathbb{E}(\Pi(f : \|f - f_0\|_n > 4\varepsilon r|\mathcal{D}_n)) &\le A(r) + B(r) + C(r) + D(r) \\ A(r) &= \mathbb{E}(\psi(\mathcal{D}_n)) \\ B(r) &= P(\mathcal{D}_n \in \mathcal{A}^c) \\ C(r) &= \mathbb{E}(I(\mathcal{D}_n \in \mathcal{A}) \cdot \Pi(f \notin \mathcal{F}_r|\mathcal{D}_n)) \\ D(r) &= \mathbb{E}(\Pi(f \in \mathcal{F}_r : \|f - f_0\|_n > 4\varepsilon r|\mathcal{D}_n)(1 - \psi(\mathcal{D}_n))I(\mathcal{D}_n \in \mathcal{A})).\end{aligned}$$

Note that the expectation is taken for the randomness of data $\mathcal{D}_n$ under the true model f_0. In what follows, we will bound each term.

Bounding A(r): type-1 error bound. Since we can chose any test ψ, we will choose the test to be the one that satisfies Lemma 2.

Note that we may not be able to do so if the packing number is too large. When we restrict ourselves to $\mathcal{F}_r$, condition (F) implies that the ε-packing number is

$$D(\varepsilon, \mathcal{F}_r, \|\cdot\|_n) \le e^{n\varepsilon^2 r^2}.$$

Consider the parameter space Θ is formed by $\theta = (f(X_1), \cdots, f(X_n))$ with $f \in \mathcal{F}_n$. For any $f_1, f_2 \in \mathcal{F}_r$,

$$\|\theta_1 - \theta_2\|^2 = \sum_{i=1}^{n} |f_1(X_i) - f_2(X_i)|^2 = n\|f_1 - f_2\|_n^2.$$

Thus, the number $M(s, \Theta)$ can be bounded by the packing number via

$$M(\sqrt{n}s, \Theta) \le D(s, \mathcal{F}_r, \|\cdot\|_n).$$

Using the fact that $M(s_1, \Theta) \le M(s_2, \Theta)$ if $s_1 > s_2$ and the choice $s = 4\sqrt{n}\varepsilon r$ in Lemma 2 and $r > 1$, we can find a test ψ such that

$$\begin{aligned}A(r) = \mathbb{E}(\psi(\mathcal{D}_n)) &\le M(2\sqrt{n}\varepsilon r, \Theta)e^{-2n\varepsilon^2 r^2} \\ &\le M(\sqrt{n}\varepsilon, \Theta)e^{-2n\varepsilon^2 r^2} \\ &\le D(\varepsilon, \mathcal{F}_r, \|\cdot\|_n)e^{-2n\varepsilon^2 r^2} \\ &\le e^{-n\varepsilon^2 r^2}.\end{aligned} \tag{5}$$

We will choose ψ to be the test with the above property. Note that the second property in Lemma 2 will be used later in bounding $D(r)$.

Bounding B(r): outside the good event. This result is straight forward from Lemma 3 due to the definition of $\mathcal{A}$ in equation (4). Thus, we immediately obtain

$$B(r) \leq e^{-n\varepsilon^2 r^2/8}. \tag{6}$$

Note that under the choice $s = \sqrt{n}\varepsilon r$ in the event $\mathcal{A}$, the posterior in equation (1) will be modified into

$$\begin{aligned}\Pi(B|\mathcal{D}_n) &\leq \frac{\int_B \phi_{n,f}(\mathcal{D}_n)d\Pi(f)}{e^{-n\varepsilon^2 r^2}\Pi(f : \|f - f_0\| < \sqrt{n}\varepsilon r)\phi_{n,f_0}(\mathcal{D}_n)} \\ &\leq e^{n\varepsilon^2(r^2+1)} \int_B \frac{\phi_{n,f}(\mathcal{D}_n)}{\phi_{n,f_0}(\mathcal{D}_n)} d\Pi(f),\end{aligned} \tag{7}$$

where the last inequality uses condition (T): $\Pi(f : \|f - f_0\| < \sqrt{n}\varepsilon r) \geq e^{-n\varepsilon^2}$.

Bounding C(r): outside the good set $\mathcal{F}_r$. To bound this, here is an interesting note:

$$\mathbb{E}\left(\frac{\phi_{n,f}(\mathcal{D}_n)}{\phi_{n,f_0}(\mathcal{D}_n)}\right) = 1$$

since the data $\mathcal{D}_n$ is generated from ϕ_{n,f_0}. Thus, by the Fubini theorem and equation (7) and the fact that $I(\mathcal{D}_n \in \mathcal{A}) \leq 1$,

$$\begin{aligned}C(r) &= \mathbb{E}\left(I(\mathcal{D}_n \in \mathcal{A})\Pi(\mathcal{F}_r^c|\mathcal{D}_n)\right) \\ &\leq \mathbb{E}\left(e^{n\varepsilon^2(r^2+1)} \int_{\mathcal{F}_r^c} \frac{\phi_{n,f}(\mathcal{D}_n)}{\phi_{n,f_0}(\mathcal{D}_n)} d\Pi(f)\right) \\ &\leq e^{n\varepsilon^2(r^2+1)} \int_{\mathcal{F}_r^c} \mathbb{E}\left(\frac{\phi_{n,f}(\mathcal{D}_n)}{\phi_{n,f_0}(\mathcal{D}_n)}\right) d\Pi(f) \\ &\leq e^{n\varepsilon^2(r^2+1)}\Pi(\mathcal{F}_r^c) \\ &\leq e^{-n\varepsilon^2(r^2-1)},\end{aligned} \tag{8}$$

where the last inequality uses condition (F): $\Pi(\mathcal{F}_r^c) \leq e^{-2n\varepsilon^2 r^2}$. $C(r)$ is still shrinking since we are in the regime where $r \geq 2$.

Bounding D(r): the nice regime and type-2 error. First, we will invoke equation(7) again to upper bound the posterior probability:

$$\begin{aligned}D(r) &= \mathbb{E}(\Pi(f \in \mathcal{F}_r : \|f - f_0\|_n > 4\varepsilon r|\mathcal{D}_n)(1 - \psi(\mathcal{D}_n))I(\mathcal{D}_n \in \mathcal{A})) \\ &\leq \mathbb{E}\left(e^{n\varepsilon^2(r^2+1)} \int_{\mathcal{F}_r \cap \{f:\|f-f_0\| \geq 4\varepsilon r\}} \frac{\phi_{n,f}(\mathcal{D}_n)}{\phi_{n,f_0}(\mathcal{D}_n)} d\Pi(f)(1 - \psi(\mathcal{D}_n))\right) \\ &= e^{n\varepsilon^2(r^2+1)} \int_{\mathcal{F}_r \cap \{f:\|f-f_0\| \geq 4\varepsilon r\}} \mathbb{E}\left(\frac{\phi_{n,f}(\mathcal{D}_n)}{\phi_{n,f_0}(\mathcal{D}_n)}(1 - \psi(\mathcal{D}_n))\right) d\Pi(f),\end{aligned}$$

where the second equality is due to the Fubini's theorem.

A powerful result of the density ratio is that

$$\mathbb{E}\left(\frac{\phi_{n,f}(\mathcal{D}_n)}{\phi_{n,f_0}(\mathcal{D}_n)}(1-\psi(\mathcal{D}_n))\right) = \mathbb{E}(1-\psi(\mathcal{D}_n) : \mathcal{D}_n \sim \phi_{n,f}).$$

Namely, the 'data' is now drawn from f, not the true parameter f_0. This key result allows us to connect back to the second bounds in Lemma 2. Thus, we obtain a refine bound on $D(r)$:

$$D(r) \leq e^{n\varepsilon^2(r^2+1)} \int_{\mathcal{F}_r \cap \{f:\|f-f_0\| \geq 4\varepsilon r\}} \mathbb{E}(1-\psi(\mathcal{D}_n) : \mathcal{D}_n \sim \phi_{n,f}) d\Pi(f)$$

Using the second bound in Lemma 2 with the fact that $\|f-f_0\|_n \geq 4\varepsilon r$ implies $\|\theta-\theta_0\| \geq 4\sqrt{n}\varepsilon r$, j in Lemma 2 can be chosen as $j=4$ (implying $js = 4\sqrt{n}\varepsilon r$ since $s = \sqrt{n}\varepsilon r$ is chosen in event $\mathcal{A}$) and we obtain a bound

$$\sup_{f \in \mathcal{F}_r : \|f-f_0\| \geq 4\varepsilon r} \mathbb{E}(1-\psi(\mathcal{D}_n) : \mathcal{D}_n \sim \phi_{n,f}) \leq e^{-2n\varepsilon^2 r^2}.$$

Thus,

$$D(r) \leq e^{n\varepsilon^2(r^2+1)} e^{-2n\varepsilon^2 r^2} \Pi(f \in \mathcal{F}_r : \|f-f_0\|_n \geq 4\varepsilon r) \leq e^{-n\varepsilon^2(r^2-1)}. \tag{9}$$

Note that we use the fact that $\Pi(f \in \mathcal{F}_r : \|f-f_0\|_n \geq 4\varepsilon r) \leq 1$ in the last inequality.

Thus putting equations (5), (6), (8), and (9) altogether, we conclude

$$\mathbb{E}(\Pi(f : \|f-f_0\|_n > 4\varepsilon r | \mathcal{D}_n)) \leq e^{-n\varepsilon^2 r^2} + e^{-n\varepsilon^2 r^2/8} + 2e^{-n\varepsilon^2(r^2-1)}.$$

All of them are Gaussian concentrations with respect to r so clearly, we have equation (3)

$$\int_2^\infty q \cdot r^{q-1} \mathbb{E}(\Pi(f : \|f-f_0\|_n > 4\varepsilon r | \mathcal{D}_n)) dr < \infty,$$

and thus,

$$\begin{aligned} R_{q,n} &\leq (8\varepsilon)^q + (4\varepsilon)^q \int_2^\infty q \cdot r^{q-1} \mathbb{E}(\Pi(f : \|f-f_0\|_n > 4\varepsilon r | \mathcal{D}_n)) dr \\ &\leq (8\varepsilon)^q + (4\varepsilon)^q C_1 \end{aligned}$$

for some constant C_1. So the result follows.

3 Remarks

- The good event $\mathcal{A}$ is to ensure that we can upper bound the posterior probability by the density ratio and a prior probability. The density ratio lemma (Lemma 3) shows that the good event holds with a overwhelming probability.
- The function space $\mathcal{F}_r$ is used in three quantities: $A(r), C(r)$, and $D(r)$. It appears in $A(r)$ to upper bound the packing number in the normal model $M(s,\Theta)$, which is where the first condition of (F) was used and the type-1 error rate in Lemma 2. In $C(r)$, it controls the probability that is not in this good set (condition (T)). Finally, we use it in $D(r)$ in the type-2 error rate controls. Under the good event $\mathcal{A}$, the density ratio property (Lemma 3) converts the expectation under f_0 into the expectation under f so $\mathbb{E}(1-\psi)$ becomes a type-2 error rate problem and the result of Lemma 2 applies.

- Here is the grand theorem (Theorem 2.1) of Ghosh and Van Der Vaart (2000) [GV2000]. One will see the similarity between it and Theorem 1:

  Theorem 4 (Theorem 2.1 of Ghosh and Van Der Vaart (2000); simplified) *Let d be a metric of distribution function space $\mathcal{P}$ and consider data Z generated from a PDF $p_0 \in \mathcal{P}$ and denote P_0 the CDF corresponding to p_0. Suppose that the prior distribution Π on $\mathcal{P}$ satisfies the following conditions*

  (F') *for a sequence ε_n with $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$, there is a constant $C > 0$ and sets $\mathcal{P}_n \subset \mathcal{P}$ with*

  $$D(\varepsilon_n, \mathcal{P}_n, d) \leq \exp(n\varepsilon_n^2)$$
  $$\Pi(\mathcal{P} \backslash \mathcal{P}_n) \leq \exp(-n\varepsilon_n^2(C+4)).$$

  (T') $\Pi\left(P : -\mathbb{E}\left(\log \frac{p(Z)}{p_0(Z)}\right) \leq \varepsilon_n^2, \mathbb{E}\left(\left(\log \frac{p(Z)}{p_0(Z)}\right)^2\right)\right) \geq \exp(-n\varepsilon_n^2 C)$, *where p is the PDF corresponds to any distribution P.*

  Then for sufficiently large M, we have

  $$\Pi(p : d(P, P_0) \geq M\varepsilon_n | Z_1, \cdots, Z_n) \to 0$$

  in probability.

  - Note that Theorem 2.1. of [GV2000] even allow the prior to be Π_n but here I take it to be fixed for simplicity.
  - The proof of the above theorem follows a similar strategy as the proof of Theorem 1. In particular, Theorem 7.1. of [GV2000] is essentially Lemma 2 and Lemma 8.1 of [GV2000] plays the role of Lemma 3.
  - Many papers prove the convergence rate of a nonparametric Bayes procedure by verifying the conditions (F') and (T') that a prior distribution has.

- Finally, I would like to note that this result (Theorem 1 or 4) is essentially treating a Bayesian procedure as a Frequentist method. We are *assuming that there is a true model f_0 or p_0 that generates our data* and we want to show that the posterior concentrates around it. This convergence property is a Frequentist property although the estimator is constructed from a Bayesian procedure.