Frontiers of Statistics: Contraction theory for posterior distributions **Spring 2019**

Lecture 5: March 5

Lecturer: Anirban Bhattacharya & Debdeep Pati *Scribes: Huiya Zhou & Biraj Subhra Guha*

Note: *LaTeX template courtesy of UC Berkeley EECS dept & CMU's convex optimization course taught by Ryan Tibshirani.*

5.1 Posterior Convergence Rate

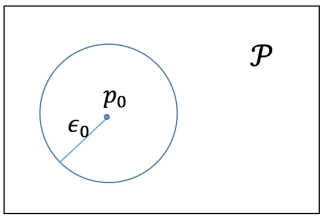
Let $x_1, \ldots, x_n \overset{iid}{\sim} p$ where p is a probability density function. The true density for observations is p_0.

Definition 5.1. *The posterior $\Pi_n(\cdot \mid x_1, \ldots, x_n)$ is said to converge at a rate ϵ_n for sequence of positive numbers $\epsilon_n \downarrow 0$, if*

$$\Pi_n(B(p_0; M_n\epsilon_n) \mid x_1, \ldots, x_n) \to 1 \text{ in probability}$$

for any sequence $M_n \to \infty$, as $n \to \infty$, where

$$B(p_0; \epsilon_n) = \{p \in \mathcal{P} : d(p, p_0) < \epsilon_n\}.$$



Remarks:

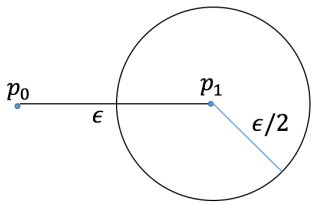
1. Why do we need M_n?
   In non-parametric problems, M_n is assumed to be a large $M > 0$.
   In parametric problems, $M_n = O(\log n)$
2. If ϵ_n is a convergence rate, $\delta_n \geq \epsilon_n$, where $\delta_n \downarrow 0$ is also a convergence rate. Typically we are interested in the "smallest/fastest" possible ϵ_n.
3. Typically convergence rate is obtained when the convergence happens "in probability".

5.2 Testing Problems

Let p_0 denote the true density and $\mathcal{P}$ is the subset of all possible density p.

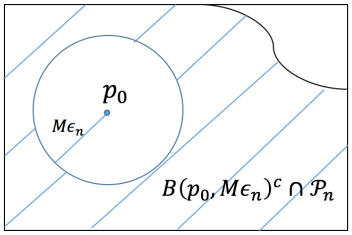
Firstly, we consider the following testing problem. If the hypothesis problem is

$$H_0 : p = p_0$$
$$H_1 : \left\{p \in \mathcal{P}, \|p - p_1\|_{\text{TV}} \leq \frac{\|p - p_0\|_{\text{TV}}}{2}\right\}$$



According to Theorem 2.1 of Lecture 2, we can construct ϕ_n such that

$$\mathbb{E}_n[\phi_n] \le e^{-\frac{n\|p-p_0\|_{\mathrm{TV}}^2}{8}}$$
$$\sup_{p\in B} \mathbb{E}_n[1-\phi_n] \le e^{-\frac{n\|p-p_0\|_{\mathrm{TV}}^2}{8}}$$



Now consider the alternative hypothesis is a region instead of a single ball.

$$H_0 : p = p_0$$
$$H_1 : p \in \mathcal{P}_n$$

where $\mathcal{P}_n \subseteq \mathcal{P}$ and $\mathcal{P}_n$ is "totally bounded".

Result: For a sequence of positive numbers $\epsilon_n \downarrow 0$, we want to use several balls to cover the region $B(p_0, M\epsilon_n)^c \cap \mathcal{P}_n$. Since $\mathcal{P} \cap B(p_0, \epsilon_n)^c$ is totally bounded. Then exists ϕ_n such that

$$\mathbb{E}_0[\phi_n] \le e^{-\frac{nM\epsilon_n^2}{8}} \mathcal{N}(\epsilon_n, \mathcal{P}_n, \|\cdot\|_{\mathrm{TV}})$$
$$\sup_{p\in\mathcal{P}_n\cap B(p_0,\epsilon_n)} \mathbb{E}_n[1-\phi_n] \le e^{-\frac{nM\epsilon_n^2}{8}}.$$

where $\mathcal{N}(\epsilon_n, \mathcal{P}_n, \|\cdot\|_{\mathrm{TV}})$ is the number of covering balls. Next, we will introduce the covering number in detail.

5.3 Covering Number

Let (X, d) denote the metric space, where d is a metric on X.

- A δ–covering / δ–net of X relative to d is any set $\{\theta_1, \ldots, \theta_N\} \subseteq X$, $\exists\ i \in \{1, \ldots, N\}$ such that $d(\theta, \theta_i) < \delta$.
- The δ–covering number of (X, d) denoted by $\mathcal{N}(\delta, X, d)$ is the cardinality of the smallest δ–covering.
- Covering number is unique but there can be multiple δ–coverings that have the same covering number.

Examples:

If $X = [-1, 1]$, $d(x, y) = |x - y|$, then $\mathcal{N}(\delta, [-1, 1], d) \approx c/\delta$.

If $X = [-1, 1]^p$, $d(x, y) = \|x - y\|$, then $\mathcal{N}(\delta, [-1, 1]^p, d) \approx (c/\delta)^p$.

Next, consider the δ–covering number $\mathcal{N}(\epsilon_n, \mathcal{P}_n, \|\cdot\|_{\mathrm{TV}})$ in our testing problem.

$$B(p_0; M\epsilon_n)^c \cap \mathcal{P}_n = \{p : \|p - p_0\|_{\mathrm{TV}} < M\epsilon_n\}$$

5.4 Main Theorem on Posterior Contraction for i.i.d setup:

Let $X_1.X_2, \dots X_n \underset{i.i.d}{\sim} p$ with p_0 denoting the true density and $p, p_0 \in \mathcal{P}$. We shall denote by $P^{(n)}$ the probability distribution of $(X_1, X_2 \dots X_n)$ under p.

Let a single dominating measure μ satisfy $p << \mu \; \forall p \in \mathcal{P}$. Denote $D(p_0||p) := \int p_0 \log\left(\frac{p_0}{p}\right) d\mu,\; V(p_0||p) := \int p_0 \left(\log \frac{p_0}{p}\right)^2 d\mu$ with the assumption that $D(p_0||p) < \infty, V(p_0||p) < \infty$.

Suppose there exist sieves $\mathcal{P}_n \subset \mathcal{P}$ equipped with distance d satisfying $d \lesssim h$, with h denoting the Hellinger metric and also a sequence $\{\epsilon_n\}_{n=1}^{\infty}$ with $\epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$ and constants $C_1, C_2 > 0$ such that the following three conditions hold:

A1) **Parameter Space Complexity:**

$$\log \mathcal{N}(\epsilon_n, \mathcal{P}_n, d) \leq C_1 n \epsilon_n^2$$

A2) **Prior Thickness/Concentration:**

$$\Pi\left\{p : D(p_0||p) \leq \epsilon_n^2,\; V(p_0||p) \leq \epsilon_n^2\right\} \geq \exp\left(-C_2 n \epsilon_n^2\right)$$

A3) **Negligibility of sieve complement in terms of prior mass:**

$$\Pi(\mathcal{P} - \mathcal{P}_n) \leq \exp\left(-(C_2 + 4) n \epsilon_n^2\right)$$

Theorem: Under conditions A1-A3, for a sufficiently large constant $M > 0$, we have:

$$\Pi\left\{p : d(p_0, p) > M\epsilon_n \;\middle|\; X_1.X_2, \dots X_n\right\} \xrightarrow[\text{in probability}]{} 0$$

w.r.t the distribution of $X_1.X_2, \dots X_n$ under the truth p_0. This shows, ϵ_n is the contraction rate of the posterior $\Pi_n(p)$ corresponding to $(\mathcal{P}, d)$.

Remarks: The following three remarks detail how the conditions help arriving at the conclusion of the theorem.

1) Condition A1 enforces that the sieve $\mathcal{P}_n$ is not 'too big'.
2) Condition A2 enforces that 'KL neighborhood' of the truth receives sufficient prior mass.
3) Condition A3 enforces that the sieve $\mathcal{P}_n$ is the 'effective parameter space', prior mass outside it being negligible.

We record an important Lemma now that shall help us proving the above theorem:

Denominator Lemma: For every $\epsilon > 0$ and probability measure Q_B on the set

$$B := \left\{p : D(p_0||p) \leq \epsilon^2,\; V(p_0||p) \leq \epsilon^2\right\}$$

the set

$$\Omega_n := \left((X_1, X_2 \dots X_n) : \int \prod_{i=1}^{n} \frac{p(X_i)}{p_0(X_i)} Q_B(dp) > \exp\left(-(C+1)n\epsilon^2\right)\right)$$

satisfies for every $C > 0$:

$$P_0^{(n)}(\Omega_n) \geq 1 - \frac{1}{C^2 n \epsilon^2} \tag{5.1}$$

Proof: All expectations and probabilities here are w.r.t P_0. By Jensen's Inequality applied to logarithm, it suffices to show:

$$P_0^{(n)}\left(\sum_{i=1}^{n}\int \log\frac{p\left(X_i\right)}{p_0\left(X_i\right)}Q_B(dp) \leq -(C+1)n\epsilon^2\right) \leq \frac{1}{C^2 n\epsilon^2} \tag{5.2}$$

Define $Z_i := \int \log\frac{p(X_i)}{p_0(X_i)}(dp),\ i = 1, 2 \ldots n$ i.i.d so that $EZ_1 = -D\left(p_0||p\right), EZ_1^2 = V\left(p_0||p\right)$ so that we can rewrite the event in the LHS of (5.2) using the fact that Q is defined on B:

$$\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Z_i - EZ_i\right) \leq -(C+1)\sqrt{n}\epsilon^2 + \sqrt{n}D\left(p_0||p\right)\right\} \subset \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Z_i - EZ_i\right) \leq -C\sqrt{n}\epsilon^2\right\} \tag{5.3}$$

Now, we simply apply Chebyshev's inequality and again use that Q is defined on B:

$$P_0^{(n)}\left(\sum_{i=1}^{n}\int \log\frac{p\left(X_i\right)}{p_0\left(X_i\right)}Q_B(dp) \leq -(C+1)n\epsilon^2\right) \leq \frac{Var\left(Z_1\right)}{C^2 n\epsilon^4} \leq \frac{1}{C^2 n\epsilon^2} \tag{5.4}$$

The lemma is proved.

There exists several ways to build a 'point vs ball-complement' test, as mentioned in class. We shall look at just one of them without proof details, and the interested reader should look up *Convergence Rates of Posterior Distributions, AoS 2000.*

Testing Lemma: For $d \lesssim h$ as mentioned before, A1 forces the existence of tests ϕ_n with the following error bounds:

$$\begin{aligned} E_{p_0}\phi_n &\leq \exp\left(C_1 n\epsilon_n^2\right).\frac{\exp\left(-KnM^2\epsilon_n^2\right)}{1-\exp\left(-KnM^2\epsilon_n^2\right)} \\ \sup_{p\in\mathcal{P}_n:\, d(p,p_0)>M\epsilon_n} E_p\left(1-\phi_n\right) &\leq \exp\left(-KnM^2\epsilon_n^2\right) \end{aligned} \tag{5.5}$$

Here M is a large constant that can be chosen suitably later, and K is a universal testing constant. E_{p_0}, E_p here denote the corresponding expectations.

Proof of Theorem:

Start by choosing $\epsilon = \epsilon_n$ and call $B_n := B$ in the Denominator lemma.

Denote $T_n := \Pi\left\{p : d\left(p_0, p\right) > M\epsilon_n \;\middle|\; X_1.X_2, \ldots X_n\right\} = \int_{p\in\mathcal{P}:\, d(p,p_0)>M\epsilon_n}\Pi_n(dp) \leq 1$ and consider the following decomposition and inequality:

$$E_{p_0}T_n \leq E_{p_0}\left[\phi_n\right] + E_{p_0}\left(T_n\mathbf{1}_{\Omega_n}\right)\left(1-\phi_n\right) + E_{p_0}\left[\mathbf{1}_{\Omega_n^c}\right] \tag{5.6}$$

Our proof of the theorem will be done if we can show that the three terms in (5.6) go to zero.

The first term goes to zero by (5.5), as we can choose M large enough to make $1-\exp\left(-KnM^2\epsilon_n^2\right) > \frac{1}{2}$ and $KM^2 > C_1 + 1$, so that $E_{p_0}\left[\phi_n\right] \leq 2\exp\left(-n\epsilon_n^2\right)$.

For the other terms, in the Denominator lemma, use $C = 1$, $B_n := B$ with $\epsilon = \epsilon_n$, $Q_B = \Pi_{B_n}$(the restriction of Π to B_n) and define $S_n' := \int\prod_{i=1}^{n}\frac{p(X_i)}{p_0(X_i)}\Pi_{B_n}(dp)$ so that:

$$\Omega_n := \left((X_1, X_2 \ldots X_n) : S_n' > \exp\left(-2n\epsilon^2\right)\right) \tag{5.7}$$

Now, write $T_n = \frac{U_n}{S_n}$. For the denominator S_n, we have the following:

$$S_n = \int \prod_{i=1}^{n} \frac{p(X_i)}{p_0(X_i)} \Pi(dp) \geq \int_{B_n} \prod_{i=1}^{n} \frac{p(X_i)}{p_0(X_i)} \Pi(dp) = \Pi(B_n) \int \prod_{i=1}^{n} \frac{p(X_i)}{p_0(X_i)} \Pi_{B_n}(dp) =: \Pi(B_n) S_n' \tag{5.8}$$

From (5.7), (5.8) and A2, we have:

$$T_n \mathbf{1}_{\Omega_n} \leq \frac{\exp\left(2n\epsilon^2\right)}{\Pi(B_n)} U_n \leq \exp\left((2 + C_2) n\epsilon_n^2\right) \tag{5.9}$$

We decompose the numerator $U_n := \int_{p \in \mathcal{P}:\, d(p,p_0) > M\epsilon_n} \prod_{i=1}^{n} \frac{p(X_i)}{p_0(X_i)} \Pi(dp)$ as follows:

$$U_n = \int_{p \in \mathcal{P}_n:\, d(p,p_0) > M\epsilon_n} \prod_{i=1}^{n} \frac{p(X_i)}{p_0(X_i)} \Pi(dp) + \int_{p \in (\mathcal{P} - \mathcal{P}_n):\, d(p,p_0) > M\epsilon_n} \prod_{i=1}^{n} \frac{p(X_i)}{p_0(X_i)} \Pi(dp) =: U_{n,1} + U_{n,2} \tag{5.10}$$

Now, (5.5), (5.10), A3 and the observations $E_{p_0}\left[\frac{p}{p_0}\right] = 1$ and $E_{p_0}\left[(1 - \phi_n)\frac{p}{p_0}\right] = E_p(1 - \phi_n)$ together imply:

$$\begin{aligned} E_{p_0} U_n \mathbf{1}_{\Omega_n} (1 - \phi_n) &\leq E_{p_0} U_{n,1} (1 - \phi_n) + E_{p_0} U_{n,2} \\ &\leq \int_{p \in \mathcal{P}_n:\, d(p,p_0) > M\epsilon_n} E_p (1 - \phi_n) \Pi(dp) + \Pi(\mathcal{P} - \mathcal{P}_n) \\ &\leq \exp\left(-KnM^2\epsilon_n^2\right) + \exp\left(-(C_2 + 4) n\epsilon_n^2\right) \\ &\leq 2\exp\left(-(C_2 + 4) n\epsilon_n^2\right) \end{aligned} \tag{5.11}$$

where in the last step, $M > \sqrt{\frac{C_2+4}{K}}$ was chosen. Now (5.9) and (5.11) together imply:

$$E_{p_0}(T_n \mathbf{1}_{\Omega_n})(1 - \phi_n) \leq \exp\left((2 + C_2) n\epsilon_n^2\right) . \exp\left(-(4 + C_2) n\epsilon_n^2\right) \leq \exp\left(-(2 + C_2) n\epsilon_n^2\right) \tag{5.12}$$

Hence, by (5.12), the second term in (5.6) goes to zero.

Also, the third term of (5.6) goes to zero directly by (5.1), where Ω_n is defined by (5.7).

This completes the proof of posterior contraction rate theorem for i.i.d data case.