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

Lecture 2: February 21

Lecturer: Anirban Bhattacharya & Debdeep Pati *Scribes: Brittany Alexander & Sandipan Pramanik*

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

2.1 Recap: Hypothesis testing and error rates (Lecam-Birge)

Let $y_1, \ldots, y_n \overset{iid}{\sim} p$ where p is a probability density function. If we consider the following simple null vs. simple alternative testing problem

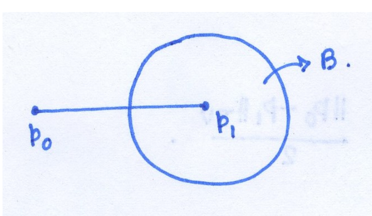
$$H_0 : p = p_0 \quad \text{vs.} \quad H_1 : p = p_1, \tag{2.1}$$

then Theorem 1.1 shows that there exists a test function $\mathbf{\Phi}_n$ such that it has exponentially decaying upper bounds to the Type-I and Type-II error probabilities; that is,

$$\mathbb{E}_{p_0} \mathbf{\Phi}_n \leq e^{-Cnh^2(p_0,p_1)} \tag{2.2}$$

$$\mathbb{E}_{p_1} [1 - \mathbf{\Phi}_n] \leq e^{-Cnh^2(p_0,p_1)}. \tag{2.3}$$

2.2 Extension to testing: Simple null vs. composite alternative



Assume the same setup as above. Now consider the following testing problem:

$$H_0 : p = p_0 \quad \text{vs.} \quad H_1 : p \in B, \tag{2.4}$$

where $B = \{p \,|\, \|p - p_1\|_{\text{TV}} \leq \|p - p_1\|_{\text{TV}}/2\}$ (there is nothing special about the '2' in the denominator; can be any positive number greater than 1). It can be shown that B is a convex set.

Theorem 2.1. *Under the described setup in 2.2, there exists a test function $\mathbf{\Phi}_n$ such that*

$$\mathbb{E}_{p_0}\mathbf{\Phi}_n \le e^{-n\|p-p_1\|_{\mathrm{TV}}^2/8} \tag{2.5}$$

$$\sup_{p\in B}\mathbb{E}_p[1-\mathbf{\Phi}_n] \le e^{-n\|p-p_1\|_{\mathrm{TV}}^2/8}. \tag{2.6}$$

Proof. Let

$$\phi(y) = \begin{cases} 1 & \text{if } p_1(y) > p_0(y) \\ 0 & \text{o.w.} \end{cases}$$

Define

$$\alpha = \mathbb{E}_{p_0}\,\phi(y_1) \qquad \text{, and} \qquad \gamma = \inf_{p\in B}\,\mathbb{E}_p\phi(y_1). \tag{2.7}$$

α and γ are the Type-I error probability and the minimum power based on one observation. Note that, α and γ can also be rewritten as follows:

$$\alpha = \int_{p_1>p_0} p_0\,d\lambda \tag{2.8}$$

$$\gamma = \inf_{p\in B}\int_{p_1>p_0} p\,d\lambda \tag{2.9}$$

where λ denotes the dominating measure.

Claim: $\gamma > \alpha$. That is, the test ϕ is unbiased.

Proof of claim: Fix $p \in B$. Then,

$$\int_{p_1>p_0} p\,d\lambda = \int_{p_1>p_0} p_1\,d\lambda - \int_{p_1>p_0}(p_1-p)\,d\lambda \tag{2.10}$$

$$\ge \int_{p_1>p_0} p_1\,d\lambda - \|p_1-p\|_{\mathrm{TV}} \tag{2.11}$$

$$\ge \int_{p_1>p_0} p_1\,d\lambda - \frac{\|p_1-p_0\|_{\mathrm{TV}}}{2} \tag{2.12}$$

which does not depend on p. Therefore,

$$\gamma-\alpha = \inf_{p\in B}\left[\int_{p_1>p_0} p\,d\lambda - \int_{p_1>p_0} p_0\,d\lambda\right] \tag{2.13}$$

$$= \inf_{p\in B}\left[\int_{p_1>p_0} p_1\,d\lambda - \frac{\|p_1-p_0\|_{\mathrm{TV}}}{2} - \int_{p_1>p_0} p_0\,d\lambda\right] \tag{2.14}$$

$$= \|p_1-p_0\|_{\mathrm{TV}}/2 > 0 \tag{2.15}$$

Hence the proof of the claim. ■

Back to main proof:

Define,

$$\mathbf{\Phi}_n = \begin{cases} 1 & \text{if } \dfrac{1}{n}\sum_{i=1}^n \phi(y_i) > \frac{\alpha+\gamma}{2} \\ 0 & \text{o.w.} \end{cases}$$

Theorem 2.2. ***Hoeffding's inequality.*** *Suppose $X_1, \cdots, X_n$ are independent random variable such that $X_i \in [0,1]$. Then, for any $t > 0$*

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i) > t\right] \leq e^{-2nt^2}. \tag{2.16}$$

Note: This can be thought of as a finite sample version the CLT.

Using this theorem,

$$\mathbb{E}_0\mathbf{\Phi}_n = \mathrm{P}_0\left[\frac{1}{n}\sum_{i=1}^n \phi(y_i) > \frac{\alpha+\gamma}{2}\right] \tag{2.17}$$

$$= \mathrm{P}_0\left[\frac{1}{n}\sum_{i=1}^n (\phi(y_i) - \mathbb{E}_0\,\phi(y_i)) > \frac{\gamma-\alpha}{2}\right] \tag{2.18}$$

$$\leq e^{-2n(\gamma-\alpha)^2/2} = e^{-n\|p_1-p_0\|_{\mathrm{TV}}^2/8} \tag{2.19}$$

For proving (2.6), fix $p \in B$. Then

$$\mathbb{E}_p(1-\mathbf{\Phi}_n) = \mathrm{P}_p\left[\frac{1}{n}\sum_{i=1}^n \phi(y_i) < \frac{\alpha+\gamma}{2}\right] \tag{2.20}$$

$$= \mathrm{P}_p\left[\frac{1}{n}\sum_{i=1}^n (\phi(y_i) - \mathbb{E}_p\,\phi(y_i)) > \frac{\alpha+\gamma}{2} - \mathbb{E}_p\,\phi(y_1)\right] \tag{2.21}$$

$$\leq \mathrm{P}_p\left[\frac{1}{n}\sum_{i=1}^n (\phi(y_i) - \mathbb{E}_p\,\phi(y_i)) > \frac{\alpha-\gamma}{2}\right] \tag{2.22}$$

$$\leq e^{-n\|p_1-p_0\|_{\mathrm{TV}}^2/8}, \tag{2.23}$$

which does not depend on p. This completes the proof of the theorem. □

2.3 Posterior consistency

Let $X_1, \dots, X_n \overset{iid}{\sim} \mathrm{P}$ where P is a probability measure (or model) and suppose X_i's are defined on $(\Omega, \mathscr{A}, \mathrm{P})$. $X^{(n)} := (X_1, \dots, X_n)$ are defined on $(\Omega^n, \mathscr{A}^n, \mathrm{P}^n)$ as X_i's are independent. Let $\mathscr{P}$ denotes the class of all probability measures.

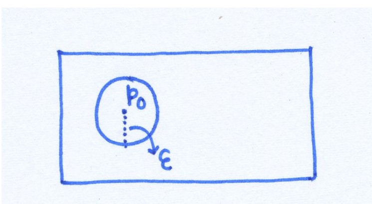
Quantity of interest: P . We further assume that P admits a density p.

True density: p_0, distribution P_0 (the true data generating density).

Let d denotes a distance metric on $\mathscr{P}$, and Π a prior on $\mathscr{P}$. Define

$$B(p_0, \varepsilon) := \{p \mid d(p_0, p) < \varepsilon\}, \tag{2.24}$$

a ball around p_0 of radius ε.



For any subset of densities B (that is, $B \subseteq \mathscr{P}$), define the posterior as

$$\Pi_n(B \mid X^{(n)}) = \frac{\int_B \prod_{i=1}^n p(x_i)\ \Pi(dp)}{\int_{\mathscr{P}} \prod_{i=1}^n p(x_i)\ \Pi(dp)}. \tag{2.25}$$

Definition: Posterior consistency. Π_n is said to be consistent at p_0 if for every $\varepsilon > 0$

$$\Pi_n\big[B(p_0, \varepsilon) \mid X^{(n)}\big] \to 1 \quad \text{in ?} \tag{2.26}$$

There are two notions of convergence; namely,

(i) "*Weak consistency*" if the above convergence happens 'in probability (i.p.)' under p_0.

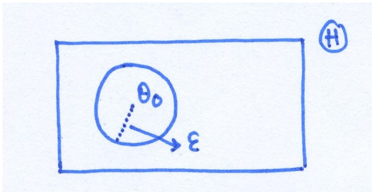
(ii) "*Strong consistency*" if the above convergence happens 'almost surely (a.s.)' under p_0.

Remark: (i) and (ii) implies $\Pi_n(\cdot \mid X^{(n)}) \xrightarrow{d} \delta_{\{p_0\}}$ a.s. or i.p. under p_0.

Posterior consistency for parameterized densities. Suppose the density p considered in the above setup is parameterized by $\theta \in \Theta$; that is, $p \equiv p_\theta$.

Example 1. $X_1, \dots, X_n \overset{iid}{\sim} \mathrm{N}(\mathbf{0}, \Sigma)$. In this case $\theta = \Sigma$.

Example 2. $X_1, \dots, X_n \overset{iid}{\sim} \mathrm{N}(\boldsymbol{\mu}, \mathbf{I})$. In this case $\theta = \boldsymbol{\mu}$.



Now the d should be considered as a distance metric on Θ and Π is a prior on Θ. We can similarly define a ball of radius ε around θ_0 as

$$B(\theta_0, \varepsilon) := \{\theta \in \Theta \,|\, d(\theta_0, \theta) < \varepsilon\}. \tag{2.27}$$

Then Π_n is said to be consistent at θ_0 if

$$\Pi_n\big[B(\theta_0, \varepsilon) \,|\, X^{(n)}\big] \to 1 \qquad \text{a.s. or i.p. under } p_{\theta_0}. \tag{2.28}$$

Properties of point estimators. Suppose $X_1, \ldots, X_n \overset{iid}{\sim} p_\theta$ (true value of θ is θ_0).

Goal: To come up with "point estimators" $\hat{\theta}_n$ depending on the posterior such that $d(\hat{\theta}_n, \theta) \to 0$ a.s. or i.p. under p_{θ_0}. If

$$\hat{\theta}_n = \int \theta \; \Pi_n(d\theta \,|\, X^{(n)}), \tag{2.29}$$

then under additional assumptions apart from posterior consistency, $d(\hat{\theta}_n, \theta) \to 0$.

Theorem 2.3. ***Alternative point estimate.*** *Suppose $\Pi_n(\cdot \,|\, X^{(n)})$ is consistent at θ_0 (with respect to (wrt) d on Θ). Let $\hat{\theta}_n$ be the center of the smallest ball in d that contains posterior mass at least 1/2. That is,*

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \hat{r}_n(\theta) \tag{2.30}$$

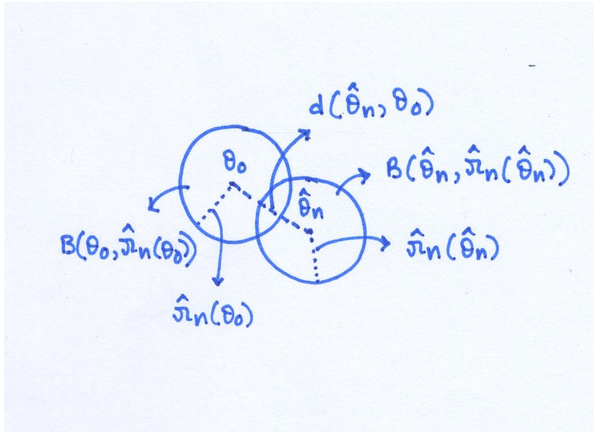
where $\hat{r}_n(\theta) = \inf\{r \,|\, \Pi_n(B(\theta, r) \,|\, X^{(n)}) \geq 1/2\}$. Then $d(\hat{\theta}_n, \theta) \to 0$ as $n \uparrow \infty$ a.s. (or i.p.) wrt p_{θ_0}.

Proof. Fix $\varepsilon > 0$. Then consistency of $\Pi_n(\cdot \,|\, X^{(n)})$ is at θ_0 implies that there exists $n_0 \equiv n_0(\varepsilon) \in \mathbb{N}$ such that

$$\Pi_n(B(\theta_0, \varepsilon) \,|\, X^{(n)}) \geq 1/2 \qquad \forall \; n \geq n_0 \tag{2.31}$$

a.s. wrt p_{θ_0}. So by definition of $\hat{r}_n(\theta)$, we get $\hat{r}_n(\theta_0) < \varepsilon$ for all $n \geq n_0$ a.s. wrt p_{θ_0}. Further, by definition of $\hat{\theta}_n$ we also get

$$\hat{r}_n(\hat{\theta}_n) \leq \hat{r}_n(\theta_0) < \varepsilon. \tag{2.32}$$



Now let us focus on the two balls $B(\hat{\theta}_n, \hat{r}_n(\hat{\theta}_n))$ and $B(\theta_0, \hat{r}_n(\theta_0))$ centered around $\hat{\theta}_n$ and θ_0, respectively. Because of the consistency condition of $\Pi_n(\cdot \,|\, X^{(n)})$ at θ_0, the two balls should overlap. Therefore, using (2.33) this implies that

$$d(\hat{\theta}_n, \theta) \leq \hat{r}_n(\hat{\theta}_n) + \hat{r}_n(\theta_0) < 2\varepsilon \qquad \forall\; n \geq n_0 \tag{2.33}$$

a.s. wrt p_{θ_0}. Since ε is arbitrary this completes the proof.

□