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

Lecture 9: March 26

Lecturer: Anirban Bhattacharya & Debdeep Pati *Scribes: Eric Chuu, Zhao Tang Luo*

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

9.1 Corollary of the Main Theorem

Corollary 9.1. *For $D > 1$. Then with probability $1 - 2/\left((D-1)^2 n\epsilon^2\right)$,*

$$\int \frac{1}{n} D^{(n)}_{\theta_0,\alpha}(\theta, \theta^\star)\Pi_{n,\alpha}\left(d\theta \mid X^{(n)}\right) \leq \left(\frac{D\alpha + 1}{1-\alpha}\right)\epsilon^2 - \frac{1}{n(1-\alpha)}\log \Pi_n\left[B_n\left(\theta^\star, \epsilon; \theta_0\right)\right]$$

Proof. First we show that with probability $1 - 1/\left((D-1)^2 n\epsilon^2\right)$,

$$\int r_n(\theta, \theta^\star)\rho(d\theta) \leq Dn\epsilon^2 \tag{9.1}$$

Using the definition of ρ and applying Jensen's inequality, we obtain the following bounds

$$\begin{aligned} P^{(n)}_{\theta_0}\left[\int r_n\left(\theta, \theta^\star\right)\rho(d\theta) > Dn\epsilon^2\right] &\leq P^{(n)}_{\theta_0}\left[\int r_n\left(\theta, \theta^\star\right)\rho(d\theta) - E_{\theta_0}\int r_n\left(\theta, \theta^\star\right)\rho(d\theta) > (D-1)n\epsilon^2\right] \\ &\leq \frac{E_{\theta_0}\left[\int r_n\left(\theta, \theta^\star\right)\rho(d\theta)\right]^2}{(D-1)^2 n^2\epsilon^4} \\ &\leq \frac{\int E_{\theta_0} r_n^2\left((\theta, \theta^\star\right)\rho(d\theta)}{(D-1)^2 n^2\epsilon^4} \\ &\leq \frac{1}{(D-1)^2 n\epsilon^2} \\ &\leq \frac{\alpha D\epsilon^2}{1-\alpha} - \frac{1}{n(1-\alpha)}\log \Pi_n\left[B_n\left(\theta^\star, \epsilon; \theta_0\right)\right] \end{aligned}$$

with probability at least

$$1 - e^{-n\epsilon^2} - \frac{1}{(D-1)^2 n\epsilon^2} \geq 1 - \frac{2}{(D-1)^2 n\epsilon^2}$$

where we make use of the theorem from last lecture with $\delta = e^{-n\epsilon^2}$. □

9.2 Proof of the Main Theorem

Lemma 9.2. *(Variational) Let μ be a probability measure, and let h any measurable function such that $\int e^h d\mu < \infty$. Then*

$$\log \int e^h d\mu = \sup_{p \ll \mu} \left[\int h d\rho - D(\rho \parallel \mu) \right]$$

where the supremum is attained for

$$\frac{d\rho}{d\mu} = \frac{e^h}{\int e^h d\mu}$$

Remark. The right hand side of the theorem is minimized for $\rho \equiv \Pi_{n,\alpha}$ by taking $h = -\alpha r_n(\theta, \theta^\star)$ and $\mu \equiv \Pi_n$ in the variational lemma. Then

$$\log \int e^{-\alpha r_n(\theta,\theta^\star)} \Pi_n(d\theta) \leq \int -\alpha r_n(\theta, \theta^\star) \rho(d\theta) - D(\rho \parallel \Pi_n)$$

for all $\rho \ll \Pi_n$. Equality happens when $\rho \equiv \Pi_{n,\alpha}$.

Proof of Main Theorem. Note the following

$$E_{\theta_0}\left[e^{-\alpha r_n(\theta,\theta^\star)}\right] = A^{(n)}_{\theta_0,\alpha}(\theta, \theta^\star) = e^{-(1-\alpha) D^{(n)}_{\theta_0,\alpha}(\theta,\theta^\star)}$$
$$\Rightarrow E_{\theta_0}\left[e^{-\alpha r_n(\theta,\theta^\star) + (1-\alpha) D^{(n)}_{\theta_0,\alpha}(\theta,\theta^\star) - \log\frac{1}{\epsilon}}\right] = \epsilon$$

Integrating both sides w.r.t. Π_n and using Tonelli's theorem gives

$$E_{\theta_0}\left[\int \exp\left(-\alpha r_n(\theta, \theta^\star) + (1-\alpha) D^{(n)}_{\theta_0,\alpha}(\theta, \theta^\star) - \log\frac{1}{\epsilon}\right)\Pi_n(d\theta)\right] = \epsilon.$$

By the variational lemma, we have

$$E_{\theta_0}\left[\exp\left\{\sup_{p \ll \Pi_n} \int \left(-\alpha r_n(\theta, \theta^\star) + (1-\alpha) D^{(n)}_{\theta_0,\alpha}(\theta, \theta^\star) - \log\frac{1}{\epsilon}\right)\rho(d\theta)\right\} - D(\rho \parallel \Pi_n)\right] = \epsilon.$$

Setting $\rho \equiv \Pi_{n,\alpha}$, we further have

$$E_{\theta_0}\left[\exp\left\{\int \left(-\alpha r_n(\theta, \theta^\star) + (1-\alpha) D^{(n)}_{\theta_0,\alpha}(\theta, \theta^\star) - \log\frac{1}{\epsilon}\right)\Pi_{n,\alpha}(d\theta \mid X^{(n)})\right\} - D(\Pi_{n,\alpha} \parallel \Pi_n)\right] \leq \epsilon.$$

Hence, with $P^{(n)}_{\theta_0}$-probability at least $1-\epsilon$,

$$(1-\alpha) \int D^{(n)}_{\theta_0,\alpha}(\theta, \theta^\star) \Pi_{n,\alpha}(d\theta \mid X^{(n)}) \leq \alpha \int r_n(\theta, \theta^\star) \Pi_{n,\alpha}(d\theta \mid X^{(n)}) + D(\Pi_{n,\alpha} \parallel \Pi_n) + \log\frac{1}{\epsilon},$$

by the fact that $P(X \geq 0) \leq E\left(e^X\right)$ for a random variable X.

Noticing that

$$\begin{aligned}
&\alpha \int r_n(\theta, \theta^\star)\Pi_{n,\alpha}(d\theta \mid X^{(n)}) + D(\Pi_{n,\alpha} \parallel \Pi_n) \\
&= -\int \log \left(\frac{p_\theta(X^{(n)})}{p_{\theta^\star}(X^{(n)})} \right)^\alpha \Pi_{n,\alpha}(d\theta \mid X^{(n)}) - \int \log \frac{\Pi_{n,\alpha}(\theta \mid X^{(n)})}{\Pi_n(\theta)} \Pi_{n,\alpha}(d\theta \mid X^{(n)}) \\
&= -\int \log \frac{\int \left(p_{\theta'}(X^{(n)})\right)^\alpha \Pi_n(d\theta')}{\left(p_{\theta^\star}(X^{(n)})\right)^\alpha} \Pi_{n,\alpha}(d\theta \mid X^{(n)}) \\
&= -\log \int e^{-\alpha r_n(\theta,\theta^\star)} \Pi_n(d\theta),
\end{aligned}$$

where the second equality is due to the definition of $\Pi_{n,\alpha}$, we now have

$$\begin{aligned}
&\int \frac{1}{n} D^{(n)}_{\theta_0,\alpha}(\theta, \theta^\star)\Pi_{n,\alpha}(d\theta \mid X^{(n)}) \\
&= -\frac{1}{n(1-\alpha)} \log \int e^{-\alpha r_n(\theta,\theta^\star)} \Pi_n(d\theta) + \frac{1}{n(1-\alpha)} \log \frac{1}{\epsilon} \\
&\leq \frac{\alpha}{n(1-\alpha)} \int r_n(\theta, \theta^\star)\rho(d\theta) + \frac{1}{n(1-\alpha)} D(\rho \parallel \Pi_n) + \frac{1}{n(1-\alpha)} \log \frac{1}{\epsilon}
\end{aligned}$$

for all $\rho \ll \Pi_n$ by the remark after the variational lemma.

9.3 An Example

Consider a convex function regression $y_i = \mu(x_i) + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0,1)$, where $x_i \in [0,1]^d$ is fixed and $\mu \in \mathcal{F} = \{\text{all convex functions}\}$. Let $\mu_0(\cdot)$ be the true mean function. There are two cases: (1) $\mu_0 \in \mathcal{F}$, (2) $\mu_0 \notin \mathcal{F}$.

Write $p_\theta^{(n)} = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(y_i - \mu(x_i)\right)^2\right)$. Then

$$D(p_{\theta_0}^{(n)} \parallel p_\theta^{(n)}) = \frac{n}{2} \left\|\mu_0 - \mu\right\|_{2,n}^2 = \frac{1}{2} \sum_{i=1}^n \left[\mu_0(x_i) - \mu(x_i)\right]^2.$$

and

$$\mu^* = \theta^* = \arg\min_{\mu \in \mathcal{F}} \left\|\mu_0 - \mu\right\|_{2,n}^2.$$

The misspecified Renyi divergence is given by

$$\begin{aligned}
D^{(n)}_{\theta_0,\alpha}(\theta, \theta^\star) &= D^{(n)}_{\mu_0,\alpha}(\mu, \mu^*) \\
&= \frac{n\alpha}{2(1-\alpha)} \left[(1-\alpha)\left\|\mu_0 - \mu\right\|_{2,n}^2 + 2\langle \mu - \mu^*, \mu - \mu_0 \rangle_{2,n}\right].
\end{aligned}$$

A sufficient condition for $D^{(n)}_{\mu_0,\alpha} \geq 0$ is that the set $\{p_\mu^{(n)} : \mu \in \mathcal{F}\}$ is convex, which doesn't hold for this problem. However, from Figure 9.1 we know that $\langle \mu - \mu^*, \mu - \mu_0 \rangle_{2,n} \geq 0$ and thus $D^{(n)}_{\mu_0,\alpha} \geq 0$.

The prior for μ is specified as a uniform distribution on the maximum of hyperplanes $\max_{1\leq k\leq K}\{a_k^T x + b_k\}$, and a Poisson prior is placed on K.

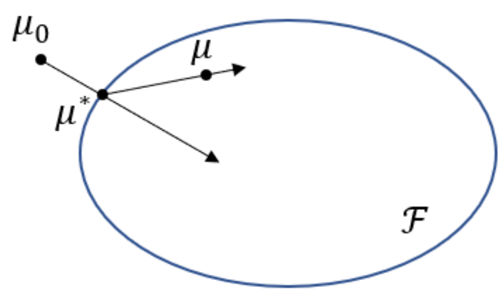

Figure 9.1: The angle between $\mu - \mu^*$ and $\mu - \mu_0$ is less than $\pi/2$ and thus $\langle \mu - \mu^*, \mu - \mu_0 \rangle_{2,n} \geq 0$.