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

Lecture 8: March 21

Lecturer: Anirban Bhattacharya & Debdeep Pati *Scribes: Brittany Alexander & Huiling Liao*

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

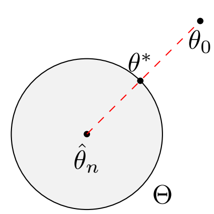
A new framework will be introduced to study the contraction rate. For more details, please refer to *Bayesian fractional posteriors* [ADY19].

8.1 Motivation

1. $\pi(\mathcal{P}_n^c) \leq e^{-n\varepsilon_n^2}$ and $\log(\varepsilon_n, \mathcal{P}_n, d) \leq n\varepsilon_n^2$ are not satisfied for π which have polynomial tails.
2. Don't deliver posterior "Risk bound". (follows using Jensen's Inequality)

$$\int_d (\theta, \theta_0)\pi(d\theta|X^{(n)}) \leq \varepsilon_n^2 \Rightarrow d(\hat{\theta}_n, \theta_0) \leq \varepsilon_n \text{ where } \hat{\theta}_n = \int \theta\pi(d\theta|X^{(n)})$$

3. Handle model misspecification.

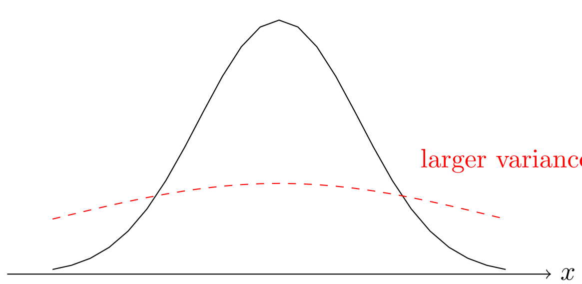


$$\hat{\theta}_n = \text{argmin}_{\theta\in\Theta} D(\theta||\theta_0)$$

Goal: Develop risk bounds if θ_0 is outside Θ.

First related paper: Kleijn & Van-der Vaart (2006, AoS).

Idea: we don't work with the likelihood but power-likelihood/ fractional likelihood $L_n^\alpha(\theta), 0 < \alpha < 1$.



8.2 Some Preliminaries

P, Q are probabilitic measures $<< \mu$, and $p = \frac{dP}{d\mu}, q = \frac{dQ}{d\mu}$.

Distances and divergences between probability measures:

$$h^2(p,q) = \frac{1}{2}\int(\sqrt{p}-\sqrt{q})^2 \mathrm{d}\mu$$
$$||p-q||_{TV} = \frac{1}{2}\int |p-q| d\mu$$
$$D(p||q) = \int p \log\frac{p}{q} d\mu$$

Renyi Divergence: For $\alpha \in (0,1)$,

$$D_\alpha(p,q) = \frac{1}{\alpha - 1}\log\int p^\alpha q^{(1-\alpha)} d\mu$$
$$A_\alpha(p,q) = \int p^\alpha q^{1-\alpha} d\mu \qquad (\alpha \text{ - affinity})$$

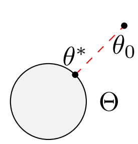
Properties:

1. $0 \leq A_\alpha(p,q) \leq 1$ if $\alpha \in (0,1) \Rightarrow D_\alpha(p,q) \geq 0$
2. $$D_{1/2}(p,q) = -2\log\int\sqrt{pq}d\mu$$ $$= -2\log\{1-h^2(p,q)\} \geq 2h^2(p,q) \qquad [\log(1+x) \leq x, x > -1]$$
3. For fixed p, q, $D_\alpha(p,q)$ is an increasing function $\alpha \in (0,1)$, i.e. if $\alpha_1 \leq \alpha_2, D_{\alpha_1}(p,q) \leq D_{\alpha_2}(p,q)$. For any $0 < \alpha \leq \beta < 1$, $$\frac{\alpha}{\beta}\frac{1-\beta}{1-\alpha}D_\beta \leq D_\alpha \leq D_\beta, 0 < \alpha \leq \beta < 1$$
4. By application of L'Hospital's rule $\lim_{\alpha\to 1} D_\alpha(p,q) = D(p,q)$

Power posterior: $\mathcal{X}^{(n)}, \mathcal{G}^{(n)}, \mathbb{P}_\theta^{(n)}$ statistical experiments. $\theta \in \Theta$, π_n: prior.

$$\mathbf{x}^{(n)} = (x_1, x_2, \ldots, x_n),\ p_\theta^{(n)} = \frac{dP_\theta^{(n)}}{d\mu}$$
$$L_{n,\alpha}(\theta) = [p_\theta^{(n)}(\mathbf{x}^{(n)})]^\alpha$$
$$\Pi_{n,\alpha}(\theta) = \frac{L_{n,\alpha}(\theta)\Pi_n(d\theta)}{\int_\Theta L_{n,\alpha}(\theta)\Pi_n(d\theta)}$$

Goal:



$d(\theta,\theta')$ is a distance metric on Θ.

$\int d(\theta,\theta')\pi_{n,\alpha}(d\theta|\mathbf{x}^{(n)})$: fractional posterior risk

$\theta^* = \mathrm{argmin}_{\theta\in\Theta} D(P_{\theta_0}^{(n)}||P_\theta^{(n)})$: KL-divergence

Let $\Pi_{n,\alpha}(\cdot)$ denote the posterior distribution obtained by combining the fractional likelihood $L_{n,\alpha}$ with the prior Π_n, that is, for any measurable set $B \in \mathcal{B}$,

$$\Pi_{n,\alpha} = \frac{\int_B e^{\alpha r_n(\theta,\theta^*)}\Pi_n(d\theta)}{\int_\Theta e^{-\alpha r_n(\theta,\theta^*)}\Pi_n(d\theta)}, \text{ where } r_n(\theta,\theta^*) = \log \frac{p_{\theta^*}^{(n)}(\mathbf{x}^{(n)})}{p_\theta^{(n)}(\mathbf{x}^{(n)})}$$

Misspecified Renyi Divergence

$$D_{\theta_0,\alpha}^{(n)} = \frac{1}{\alpha-1}\log A_{\theta_0,\alpha}^{(n)}(\theta,\theta^*) \text{ where } A_{\theta_0,\alpha}^{(n)}(\theta,\theta^*) = \int \{\frac{p_\theta^{(n)}}{p_{\theta^*}^{(n)}}\}^\alpha p_{\theta_0}^{(n)} d\mu$$

If $\theta^* = \theta_0$ (well specified case), then $A_{\theta_0,\alpha}^{(n)}(\theta,\theta^*) = A_\alpha(\theta,\theta^*) = A_\alpha(\theta,\theta_0)$.

Result: If θ^* as defined before and $\{p_\theta^{(n)} : \theta \in \Theta\}$ is a convex set, then $0 \le A_{\theta_0,\alpha}^{(n)}(\theta,\theta^*) \le 1$.

Recall that $B(\theta_0;t) = \{\theta : \int p_{\theta_0} \log \frac{p_{\theta_0}}{p_\theta} < \varepsilon^2, \int p_{\theta_0}(\log frac p_0 p_\theta)^2 \le \varepsilon^2\}$. Define a θ^*-specific KL-neighbourhood as following:

$$B(\theta^*,\varepsilon;\theta_0) = \{\theta \in \Theta : \int p_{\theta_0}^{(n)} \log \frac{p_{\theta^*}^{(n)}}{p_\theta^{(n)}} d\mu < n\varepsilon^2, \int p_{\theta_0}^{(n)} (\log \frac{p_{\theta^*}^{(n)}}{p_\theta^{(n)}})^2 d\mu < n\varepsilon^2\}$$

Theorem 8.1. *(Risk bound) Fix $\alpha \in (0,1), \varepsilon \in (0,1)$,*

$$\int \frac{1}{n} D_{\theta_0,\alpha}^{(n)} \Pi_{n,\alpha}(d\theta|\mathbf{x}^{(n)}) \le \frac{\alpha}{n(1-\alpha)} \int r_n(\theta,\theta^*)\rho(d\theta) + \frac{1}{n(1-\alpha)} D(\rho||\Pi_n) + \frac{1}{n(1-\alpha)}\log\frac{1}{\varepsilon}$$

for all probabilistic measures $\rho << \Pi_n$ with $p_{\theta_0}^{(n)}$-probability at least $1-\varepsilon$.

(We are going to choose rho minimizes the first 2 terms. The first term behaves like the likelihood, how well it fits the data. The second term is for how far the prior is away from the θ^*.)

Illustration: Let's assume $\rho(d\theta) = \Pi_{n,\alpha}(d\theta|\mathbf{x}^{(n)})$.

Sum of the first 2 terms:

$$\begin{aligned} &\frac{1}{n(1-\alpha)} \int (\alpha r_n(\theta,\theta^*) + \log \frac{\rho(\theta)}{\Pi_n(\theta)})\rho(d\theta) \\ =&\frac{1}{n(1-\alpha)} \int \log \frac{\rho(\theta)}{\Pi_n(\theta) e^{-\alpha r_n(\theta,\theta^*)}}\rho(d\theta) \\ =&-\frac{1}{n(1-\alpha)}\{\log \int \exp\{-\alpha r_n(\theta,\theta^*)\}\Pi_n(d\theta)\} \end{aligned}$$

Corollary 8.2. *With probability at least $1 - \frac{2}{(D-1)^2 n\varepsilon^2}, D > 1$,*

$$\int \frac{1}{n} D_{\theta_0,\alpha}^{(n)} \Pi_{\theta_0,\alpha}(d\theta|\mathbf{x}^{(n)}) \le \frac{D\alpha+1}{1-\alpha}\varepsilon^2 - \frac{1}{n(1-\alpha)}\log \Pi_n(B_n(\theta^*,\varepsilon;\theta_0))$$

Assume $\Pi_n(B(\theta^*,\varepsilon;\theta_0)) \ge e^{-n\varepsilon^2}$, then

$$\int \frac{1}{n} D_{\theta_0,\alpha}^{(n)} \Pi_{\theta_0,\alpha}(d\theta|\mathbf{x}^{(n)}) \le \frac{D\alpha+1}{1-\alpha}\varepsilon^2 + \frac{\varepsilon^2}{1-\alpha} = C\cdot\varepsilon^2$$

If we choose ρ to be very concentrated around θ^*, $\rho = \frac{\P_n 1_C}{\Pi_n(C)}$, then

$$D(\rho||\Pi_n) = \int_C \frac{\Pi_n(d\theta)}{\Pi_n(C)} \log \frac{\Pi_n(\theta)}{\Pi_n(C)\Pi_n(\theta)} = -\log \Pi_n(C), C = B(\theta^*, \varepsilon; \theta_0)$$

To simplify the argument: $\theta^* = \theta_0$

$$B(\theta^*, \varepsilon; \theta_0) = \{\theta : ||\theta - \theta_0|| < \varepsilon^2\}$$

$$\log P(||\theta - \theta_0|| < \varepsilon) \doteq -\frac{1}{2}\theta_0^T \Sigma^{-1}\theta_0 + \log \underbrace{P(||\theta|| < \varepsilon)}_{\varepsilon^d}$$

$$\log P(||\theta - \theta_0|| < \varepsilon) \doteq d \log \varepsilon$$

References

[ADY19] A. Bhattacharya , D. Pati and Y. Yang, "Bayesian fractional posteriors," *The Annals of Statistics*, 47(1), 2019, pp. 39–66.