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

Lecture 1: February 19

Lecturer: Anirban Bhattacharya & Debdeep Pati *Scribes: Satwik Acharyya*

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

1.1 Introduction

We'll consider the high dimensional set up here. The data set is denoted with the usual notation x and θ denotes corresponding high (or infinite) dimensional parameter.

$$\begin{aligned}
&x \mid \theta, \lambda \sim f(x \mid \theta) \\
&\theta \mid \lambda \sim \pi(\theta \mid \lambda) \\
&\lambda \sim p(\lambda)
\end{aligned}$$

Here λ is the hyper-parameter. Now the marginal posterior distribution of θ is given as $\pi(\theta|x) = \int \pi(\theta, \lambda|x)d\lambda$. Posterior mean is defined as $\boldsymbol{\hat{\theta} = \int \theta \pi(\theta|x) d\theta}$. For the whole course, we are going to denote the true data generating parameter as θ_0. Ideally, we want $\pi(. \mid x)$ to "concentrate" around θ_0 as sample size increases.

Comment 1 : In high dimensional set up, we make subjective assumptions on priors because objective choice of prior is difficult in that scenario.

Comment 2 : Bernstein von Mises theorem states that the posterior distribution takes a asymptotic normal shape in case of regularized parameter model. To prove similar kind of results, more assumptions are required in high dimensional set up.

Question : How well does the posterior mean $(\hat{\theta})$ perform in "recovering" the true data generating parameter θ_0? First we need to define a loss function or distance function to answer the recovery rate of a parameter.

Notation : $d(\hat{\theta}, \theta_0)$: Measures distance between estimator & true value of the parameter.

For posterior mean, at the least we want $E_{\theta_0}(\hat{\theta}, \theta_0) \to 0$ as $n \to \infty$ where E_{θ_0} denotes the expectation under true parameter θ_0. We are also going to focus on the convergence rate of $E_{\theta_0}(\hat{\theta}, \theta_0)$ towards 0. The notion of fast convergence is discussed through fundamental information theoretic lower bound.

Definition : We say ϵ_n to be the minimax rate w.r.t. loss function d & parameter space Θ if

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} E_\theta(\hat{\theta}, \theta) \asymp \epsilon_n$$

Comment : Infimum is taken over all estimators of θ and maximum risk is considered over the space Θ. $E_\theta(\hat{\theta}, \theta)$ represents the risk of the estimator $\hat{\theta}$. A particular estimator $\tilde{\theta}$ is said to attain the minimax lower bound if $\sup_{\theta \in \Theta} E_\theta(\hat{\theta}, \theta) \asymp \epsilon_n$.

Notation : $a_n \asymp b_n \Rightarrow 0 < c_1 < a_n/b_n < c_2 \;\forall$ large n.

Example : (Sparse Mean Estimation)

Suppose $Y \mid \mu \sim N_n(\mu, I_n)$ and μ is sparse which means $\mu \in l_0[s;n] = \{\theta \in \mathrm{R}^n : \text{at most s coordinates are non-zero}\}$. Now the distance from sparse mean estimator is provide by

$$d(\hat{\mu}, \mu) = \frac{\|\hat{\mu} - \mu\|^2}{n} = \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_i - \mu_i]^2.$$

The minimax lower bound is of the order $\frac{s}{n}\log(\frac{n}{s})$.

Comment 1 : First we are going to discuss about the log term in the minimax rate. Consider an example where we know that first s coordinates are non zero and rest of the coordinates are zero which means $\mu_1 \neq \cdots \neq \mu_s \neq 0$ and $\mu_{s+1} = \cdots = \mu_n = 0$. The corresponding estimator is provided below.

$$\hat{\mu}_j = \begin{cases} Y_j, & \text{if } j = 1, \dots, s \\ 0, & \text{if } j = s+1, \dots, n \end{cases}$$

$$d(\hat{\mu}, \mu) = \frac{\|\hat{\mu} - \mu\|^2}{n} = \frac{s}{n}$$

The logarithmic term appears because of not knowing the location of sparsity. Combinatorial price is adjusted in logarithmic order.

Comment 2 : Minimax rate is adaptive to s. It gives us the minimax rate without knowledge of s.

Question : What will be the minimax rate in case of miss specified model ?

Ans : We usually assume that the models are correct in case of determining the minimax rate.

Next we are going to talk about distances and divergences between probability measures. Let P and Q are the probability measures with densities $p = dP/d\mu$ and $q = dQ/d\mu$ w.r.t. dominating measure μ.

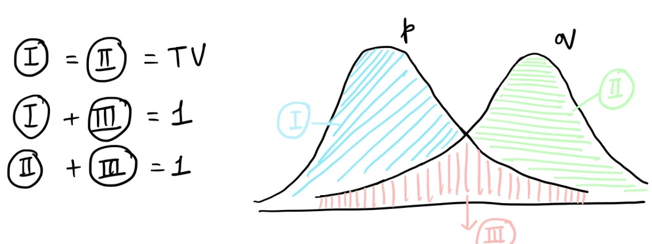
1. **Hellinger Distance** :

$$h(p,q) = h^2(p,q)^{\frac{1}{2}}$$
$$h^2(p,q) = \frac{1}{2}\int(\sqrt{p} - \sqrt{q})^2 d\mu = 1 - \int\sqrt{pq}\,d\mu = 1 - A(p,q)$$

   Hellinger Affinity $= A(p,q) = \int\sqrt{pq}\,d\mu$
   Comment : $0 \leq h(p,q) \leq 1$.

2. **Total Variation Distance** :

$$\| p - q \|_{TV} = \sup_{\text{B:Borel set}} | P(B) - Q(B) | = \frac{1}{2}\int | p - q | \, d\mu$$
$$\int_{p>q}(p-q)d\mu = \int_{q>p}(q-p)d\mu = 1 - \int \min(p,q)d\mu$$



   Comment : $0 \leq \| p - q \|_{TV} \leq 1$.

3. **Kullback Leibler Divergence** :

$$D(p \mid\mid q) = \int p\log\left(\frac{p}{q}\right)d\mu$$

Example : $p \equiv \text{N}(0,1)$ and $q \equiv \text{N}(0,1)$

$$\mid\mid p - q \mid\mid_{TV} = 2\Phi\left(\frac{\mu}{2}\right) - 1$$
$$\mid\mid p - q \mid\mid_{TV} = \begin{cases} 1, & \text{if } \mu = \pm\infty \\ 0, & \text{if } \mu = 0 \end{cases}$$
$$D(p \mid\mid q) = \frac{\mu^2}{2}$$
$$D(p \mid\mid q) = \begin{cases} 0, & \text{if } \mu = 0 \\ \infty, & \text{if } \mu = \infty \end{cases}$$

Inequalities

$$\begin{aligned} \mid\mid p - q \mid\mid_{TV}^2 \lesssim h^2(p,q) \lesssim \mid\mid p - q \mid\mid_{TV} \\ h^2(p,q) \lesssim D(p \mid\mid q) \lesssim h^2(p,q)\left[1 + \log \mid\mid p/q \mid\mid_\infty\right] \end{aligned} \tag{1.1}$$

Notation : $a \lesssim b$ mean $a \leq Cb$ for a positive constant C.

Product Measures Now we are are going to define product measures and establish their connection to distance measures.

$p = p_1 \bigotimes \cdots \bigotimes p_m$ & $p(y_1, \ldots y_m) = \prod_{i=1}^m p_i(y_i)$ similarly $q = q_1 \bigotimes \cdots \bigotimes q_m$ & $q(y_1, \ldots y_m) = \prod_{i=1}^m q_i(y_i)$.

$$D(p \mid\mid q) = \sum_{i=1}^m D(p_i \mid\mid q_i)$$

Example : $p \equiv N(\mu, I_m)$ and $q \equiv N(0, I_m) \Rightarrow D(p \mid\mid q) = \sum_{i=1}^m \mu_i^2/2 = \mid\mid \mu \mid\mid^2 /2$.

$$\mid\mid p - q \mid\mid_{TV} \leq \sum_{i=1}^m \mid\mid p_i - q_i \mid\mid_{TV}$$

$$h^2(p,q) = 1 - A(p,q) = 1 - \prod_{i=1}^m A(p_i, q_i) = 1 - \prod_{i=1}^m [1 - h^2(p_i, q_i)] \leq \sum_{i=1}^m h^2(p_i, q_i)$$

1.2 Hypothesis Testing and error rates

Let $y_1, \ldots, y_n \overset{\text{iid}}{\sim} p$. We are going to set up simple null vs simple alternative test which is $H_0 : p = p_0$ vs $H_1 : p = p_1$. Let

$$\Phi_n(y_1, \ldots, y_n) = \begin{cases} 1, & \text{if } \prod_{i=1}^n \frac{p_1(y_i)}{p_0(y_i)} > 1 \\ 0, & \text{ow} \end{cases}$$

The cut off is arbitrarily considered as 1. Our area of interest will be of type-I and type-II error rates.

Theorem 1.1. *Under previously mentioned set, we can obtain exponential rates for type-I and type-II error.*

$$E_{p_0}[\Phi_n] \leq e^{-Cnh^2(p_0,p_1)}$$

$$E_{p_1}[1-\Phi_n] \leq e^{-Cnh^2(p_0,p_1)}$$

Proof.

$$\begin{aligned}
&E_{p_0}[\Phi_n] = P_{p_0}\left[\prod_{i=1}^{n}\frac{p_1(y_i)}{p_0(y_i)} > 1\right] = P_{p_0}\left[\prod_{i=1}^{n}\sqrt{\frac{p_1(y_i)}{p_0(y_i)}} > 1\right] \\
&\leq E_{p_0}\left(\prod_{i=1}^{n}\sqrt{\frac{p_1(y_i)}{p_0(y_i)}}\right) = \left\{E_{p_0}\left(\prod_{i=1}^{n}\sqrt{\frac{p_1(y_i)}{p_0(y_i)}}\right)\right\}^n = \{A(p_0,p_1)\}^n \\
&= e^{n\log A(p_0,p_1)} = e^{n\log(1-h^2(p_0,p_1))} \leq e^{-nh^2(p_0,p_1)}
\end{aligned}$$

$\square$