ECE598: Information-theoretic methods in high-dimensional statistics Spring 2016

Lecture 17: Density Estimation

Lecturer: Yihong Wu *Scribe: Jiaqi Mu, Mar 31, 2016 [Ed. Apr 1]*

In last lecture, we studied the minimax risk of a parameterized density estimation and its upper bound. We are given n i.i.d. samples X_1, ..., X_n generated from P_θ, where $P_\theta \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is the density to be estimated. Let the loss function between a true distribution P_θ and an estimated distribution $\hat{P}$ be their KL-divergence, i.e.,

$$\ell(P_\theta, \hat{P}) = D(P_\theta \| \hat{P}).$$

One can bound the minimax risk R^* of this estimation problem by,

$$R_n^* = \inf_{\hat{P}} \sup_{\theta \in \Theta} {}_\theta D(P_\theta \| \hat{P}) \le \frac{C_n}{n}, \tag{17.1}$$

where C_n is the capacity over θ and X^n, i.e.,

$$C_n = \sup_{\pi \in M(\Theta)} I(\theta; X^n) = \inf_{\epsilon > 0} \{n\epsilon + \log N_{KL}(\epsilon)\},$$

where $N_{KL}(\epsilon)$ is the covering number of $\mathcal{P}$.

Further, we can use the chain rule in mutual information to learn the properties of C_n. For any prior π over Θ, one has,

$$R_\pi^* = I(\theta; X_{n+1} | X^*) = I(\theta; X^{n+1}) - I(\theta; X^n).$$

Taking the supremum over π on both sides, one has,

$$\begin{aligned} R_n^* = \sup R_\pi^* &= \sup_\pi \left(I(\theta; X^{n+1}) - I(\theta; X^n) \right) \\ &\ge \sup_\pi I(\theta; X^{n+1}) - \sup_\pi I(\theta; X^n) = C_{n+1} - C_n. \end{aligned}$$

Therefore we can have a lower bound over R_n^* as well.

Remark 17.1. There are some properties of $\{C_n\}$:

- $\{C_n\}$ is subadditive and increasing, i.e.,

  $$C_{n+m} \le C_n + C_m, \qquad \forall m,\ n \in \mathbb{Z}_+.$$

  and therefore $\frac{C_n}{n}$ has a limit for $n \to \infty$. By Fekete's lemma,

  $$\lim_{n \to \infty} \frac{C_n}{n} = \inf_{n \ge 1} \frac{C_n}{n}.$$

- If we let $\Delta_n = C_{n+1} - C_n$, one can rewrite C_n by,

$$C_n = \sum_{k=1}^{n-1} \Delta_k,$$

and therefore,

$$\Delta_n \leq \frac{\sum_{k=1}^{n-1} \Delta_k}{n}.$$

In today's lecture, we use the bound in (17.1) to study the minimax risk of a nonparameterized density estimation.

17.1 Density Estimation

We are interested in estimating a smooth probability density function. To be precise, we are interested in estimating a pdf $f \in \mathcal{P}_\beta$ with smoothness parameter $\beta > 0$, where f belongs to $\mathcal{P}_\beta$ iff,

- f is a pdf on $[0, 1]$ and is upper bounded by a constant, say, 2.
- $f^{(m)}$ α-Hölder continuous, i.e.,

$$|f^{(m)}(x) - f^{(m)}(y)| \leq |x - y|^\alpha, \qquad \forall\ x, y \in (0, 1),$$

where $\alpha \in (0, 1]$, $m \in \mathbb{Z}$ and $\beta = \alpha + m$.

Note: For example, if $\beta = 1$, then $\mathcal{P}_1$ is simply the set of pdfs which are Lipshitz and bounded by 2.

Theorem 17.1. *Given n i.i.d. samples X_1, ..., X_n randomly generated from a pdf $f \in \mathcal{P}_\beta$, the minimax risk of an estimation $\hat{f}$ of f under the quadratic loss function $\ell(f, \hat{f}) = \|f - \hat{f}\|_2^2 = \int_0^1 (f(x) - \hat{f}(x))^2 dx$ satisfies*

$$R^*(\mathcal{P}_\beta) = \inf_{\hat{f}} \sup_{f \in \mathcal{P}_\beta} \|f - \hat{f}\|_2^2 \asymp n^{-\frac{2\beta}{1+\beta}}. \tag{17.2}$$

Before we goes into the proof for Theorem 17.1, we makes some remarks.

Remark 17.2. The larger β is, the smoother the pdfs are, the faster R^* decays with n.

Remark 17.3. If f is defined over $[0, 1]^d$, the bound turns into,

$$R_n^*(\mathcal{P}_\beta) = \inf_{\hat{f}} \sup_{f \in \mathcal{P}_\beta} \|f - \hat{f}\|_2^2 \asymp n^{-\frac{2\beta}{d+\beta}}$$

Now we prove Theorem 17.1.

Proof. First, we claim we can use the minimax risk over a set of lower bounded pdfs to bound $R_n^*(\mathcal{P}_\beta)$. The idea is in Lemma 17.1

Lemma 17.1. *Let $\mathcal{F}$ be the set of pdfs that are lower bounded, i.e., $\mathcal{F} = \left\{f : f \geq \frac{1}{2}\right\}$. Let $\mathcal{P}$ be an arbitrary set of pdfs on $[0,1]$ and let $\tilde{\mathcal{P}} = \mathcal{P} \cap \mathcal{F}$. Then*

$$R_n^*(\mathcal{P}) \leq R_n^*(\tilde{\mathcal{P}}) \leq 16 R_n^*(\mathcal{P}).$$

Proof of Lemma 17.1. Since $\tilde{\mathcal{P}}_\beta \subset \mathcal{P}_\beta$, the lower bound is obvious,

$$R_n^*(\tilde{\mathcal{P}}_\beta) \leq R_n^*(\mathcal{P}_\beta). \tag{17.3}$$

We will construct an estimator to show,

$$R_n^*(\mathcal{P}_\beta) \leq 16 R_n^*(\tilde{\mathcal{P}}_\beta). \tag{17.4}$$

Let X_1, ..., X_n be the n i.i.d. samples from $f \in \mathcal{P}_\beta$ we have, and let U_1, ..., U_n be n i.i.d. samples uniformly generated from $[0,1]$. We define n i.i.d. random variables Z_1, ..., Z_n as,

$$Z_i = \begin{cases} U_i & \text{w.p. } \frac{1}{2}, \\ X_i & \text{otherwise.} \end{cases}$$

Thus, it is equivalent to think Z_1, ..., Z_n are i.i.d. samples from $g = \frac{1}{2}(1+f) \in \tilde{\mathcal{P}}_\beta$. Let $\hat{g}$ be an estimator of g from Z^n. Let $\tilde{g}$ be its projection in $\mathcal{F}$, i.e.,

$$\tilde{g} = \arg\min_{h \in \mathcal{F}} \|h - \hat{g}\|.$$

Note $g \in \mathcal{F}$, and we can bound the distance between $\tilde{g}$ and g by,

$$\|\tilde{g} - g\| \leq \|\hat{g} - g\| + \|\tilde{g} - \hat{g}\| \leq 2\|\hat{g} - g\|.$$

Let $\hat{f} = 2\tilde{g} - 1$, which is a valid pdf since $\tilde{g}$ is lower bounded by $\frac{1}{2}$. As a result, for every pdf $f \in \mathcal{P}_\beta$, there is a corresponding $g = \frac{1}{2}(1+f) \in \tilde{\mathcal{P}}_\beta$ which has a good estimator $\hat{g}$, and one can construct a good estimator $\hat{f}$ from $\hat{g}$ in the sense that,

$$\|\hat{f} - f\| = 2\|\tilde{g} - g\| \leq 4\|\hat{g} - g\|.$$

Therefore,

$$\begin{aligned} R_n^*(\mathcal{P}_\beta) &= \inf_{\hat{f}} \sup_{f \in \mathcal{P}_\beta} \|\hat{f} - f\|_2^2 \\ &\leq 16 \inf_{\hat{g}} \sup_{f \in \mathcal{P}_\beta} \left\| \hat{g} - \frac{1}{2}(1+f) \right\|_2^2 \\ &\leq 16 \inf_{\hat{g}} \sup_{g \in \tilde{\mathcal{P}}_\beta} \|\hat{g} - g\|_2^2 = R_n^*(\tilde{\mathcal{P}}_\beta), \end{aligned}$$

where the first inequality is due to the construction of $\hat{f}$, and the second inequality is due to $\left\{\frac{1}{2}(1+f) : f \in \mathcal{P}_\beta\right\} \subset \tilde{\mathcal{P}}_\beta$. Therefore from (17.3) and (17.4) Lemma 17.1 follows. □

It is then equivalent to prove,

$$R_n^*(\tilde{\mathcal{P}}_\beta) = \inf_{\hat{f}} \sup_{f \in \mathcal{P}_\beta} \|f - \hat{f}\|_2^2 \asymp n^{-\frac{2\beta}{d+\beta}}$$

Upper bound First we use the capacity to upper bound the minimax risk. On one hand, It is known that for any bounded pdf f and g,

$$\|f-g\|_1^2 \gtrsim \|f-g\|_2^2,$$

and the total variation between f and g is bounded by its KL-divergence,

$$D(f\|g) \geq 2d_{\mathrm{TV}}^2(f,g) = \frac{1}{2}\|f-g\|_1^2.$$

Therefore, we have for any bounded pdf f and g,

$$\|f-g\|_2^2 \lesssim D(f\|g)$$

As a result,

$$R_n^*(\tilde{\mathcal{P}}_\beta) = \inf_{\hat{g}} \sup_{g\in\tilde{\mathcal{P}}_\beta} \|g-\hat{g}\|_2^2 \leq \inf_{\hat{g}} \sup_{g\in\tilde{\mathcal{P}}_\beta} D(g\|\hat{g}) = R_{n,KL}^*(\tilde{\mathcal{P}}_\beta). \tag{17.5}$$

On the other hand, one can bound the minimax risk under KL-divergence by (17.1), where the capacity between g and X^n can be computed via,

$$\begin{aligned} C_n &\leq \inf_{\epsilon>0}\{\log N_{KL}(\epsilon) + n\epsilon\} \\ &\asymp \inf_{\epsilon>0}\{\log N_2(\sqrt{\epsilon}) + n\epsilon\} \\ &\asymp \inf_{\epsilon>0}\{\epsilon^{-\frac{1}{2\beta}} + n\epsilon\} = n^{\frac{1}{1+2\beta}}. \end{aligned}$$

The first equality is due to the connection between the KL-divergence and the L_2 distance. The second equality comes from Kolomogrov-Tikhomirov's Theorem. Therefore with (17.5), the upper bound is proved by showing,

$$R_n^*(\tilde{\mathcal{P}}_\beta) \lesssim \frac{C_n}{n} \leq n^{\frac{1}{1+2\beta}-1} = n^{-\frac{2\beta}{1+2\beta}}. \tag{17.6}$$

Lower bound Next we lower bound $R_n^*(\mathcal{P}_\beta)$ by Fano's inequality. Due to the relation between covering and packing numbers, we know,

$$\log M(\tilde{\mathcal{P}}_\beta, \|\cdot\|_2, \epsilon) \asymp \log N(\tilde{\mathcal{P}}_\beta, \|\cdot\|_2, \epsilon) \asymp \epsilon^{-1/\beta},$$

where the second equality is due to Kolomogrov-Tikhomirov's Theorem. Let $\epsilon = n^{-\frac{\beta}{1+2\beta}}$. Fano's inequality tells us,

$$\begin{aligned} R_n^*(\tilde{\mathcal{P}}_\beta) &\gtrsim \epsilon^2\left(1 - \frac{I(g;X^n)+\log 2}{\log M(\tilde{\mathcal{P}}_\beta, \|\cdot\|_2, \epsilon)}\right) \\ &\gtrsim \epsilon^2\left(1 - \frac{C_n}{\log M(\tilde{\mathcal{P}}_\beta, \|\cdot\|_2, \epsilon)}\right) \\ &\gtrsim \epsilon^2\left(1 - \frac{n^{\frac{1}{1+2\beta}}}{\epsilon^{-\frac{1}{\beta}}}\right) \\ &\asymp \epsilon^2 = n^{-\frac{2\beta}{1+2\beta}}. \end{aligned} \tag{17.7}$$

The proof is done via (17.6) and (17.7). □

We make some remarks on the proof.

Remark 17.4. We have learned two ways to construct a density estimator:

- The mean of predictive density estimators;
- The maximum likelihood estimator.

None of those is computationally efficient. In practice, kernel density estimator (KDE) is proposed: let $X_1, ..., X_n$ be the n samples, one can estimate the density by its histogram,

$$\hat{\pi} = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

This estimator, however, is not a pdf. To address this issue, one put a kernel instead of a spike over each sample points, i.e.,

$$\hat{f}_\Omega = \Omega \otimes \hat{\pi},$$

where Ω is the kernel function and is chosen to satisfy the smooth constraint of the pdfs.

Remark 17.5. The result in Theorem 17.1 can be generalized to L_p for $p < \infty$, following the same analysis.

17.2 Estimator Based On Pairwise Comparison

When we prove the lower bound in Theorem 17.1, we use the property that if an ϵ-covering of a set Θ cannot be tested, then $\theta \in \Theta$ cannot be estimated. On the contrary, if the ϵ-covering can be tested, can $\theta \in \Theta$ be estimated? First we study when the ϵ-covering can be tested.

In a binary classification problem over two distributions, where $\phi = 0$ indicates the data comes from P and $\phi = 1$ indicates the data comes from Q. The error is lower bounded by the total variation between P and Q, i.e.,

$$\min_{\hat{\phi}} p(\hat{\phi} \neq \phi) = 1 - d_{\mathrm{TV}}(P, Q).$$

In a binary classification problem over two sets of distributions, where $\phi = 0$ indicates the data comes from $P \in \mathcal{P}$ and $\phi = 1$ indicates the data comes from $Q \in \mathcal{Q}$. The error is lower bounded by the total variation between P and Q, i.e.,

$$\min_{\hat{\phi}} p(\hat{\phi} \neq \phi) = 1 - \min_{P \in \mathrm{co}(\mathcal{P}), Q \in \mathrm{co}(\mathcal{Q})} d_{\mathrm{TV}}(P, Q).$$

Remark 17.6. In general, the minimization of $d_{\mathrm{TV}}(P, Q)$ over the convex hulls of $\mathcal{P}$ and $\mathcal{Q}$ is a hard problem.

References