36-709: Advanced Statistical Theory I **Spring 2020**

Lecture 19: April 2

Lecturer: Siva Balakrishnan *Scribe: Weichen Wu*

In the previous lecture, we consider the **Sieve maximum likelihood** estimator in the non-parametric density estimation problem. For $X_1, X_2, ..., X_n \sim p_0 \in \mathcal{P}$, we want an estimator of the distribution p_0. In order to do so, we construct a **sieve** $\{p_1, p_2, ..., p_{N(\varepsilon)}\}$ that forms a ε-Hellinger cover of $\mathcal{P}$. We then choose a distribution $\widehat{p}$ from the sieve that maximizes the likelihood. We will analyze the property of this sieve estimator in this lecture, and the main result is:

Theorem 19.1 *If $n\varepsilon^2 \geq 8 \log N(\varepsilon)$, then :*

$$\mathbb{P}(h(\widehat{p}, p_0) \geq 2\varepsilon) \leq 2\exp(-\frac{1}{8}n\varepsilon^2)$$

We will now prove this result.

19.1 Analysis of sieve estimator

A **Sieve** $\widetilde{\mathcal{P}}$ is an ε-Hellinger cover of $\mathcal{P}$, which means that $\forall p \in \mathcal{P}$, there exists $p^* \in \widetilde{\mathcal{P}}$ such that $h(p, p^*) \leq \varepsilon$. It should also satisfy one of the following conditions:

(1) $\exists p^* \in \widetilde{\mathcal{P}}$, such that $\frac{p_0}{p^*} \leq U$ for some constant $U \leq \frac{11}{8}$ universally;

(2) $\exists p^* \in \widetilde{\mathcal{P}}$, such that $\chi^2(p_0, p^*) \leq U\varepsilon^2$for some constant $U \leq \frac{3}{8}$ universally.

Notice that (1) actually implies (2). We will use the weaker condition (2) in the following proof of Theorem 19.1. We note in passing that the same proof can be modified (by adjusting various constants) when U is simply some universal constant (not upper bounded by 11/8 or 3/8 as above). Before this proof, recall the Wong-Shen Lemma that we proved in the previous lecture:

Lemma 19.2 (Wong-Shen Lemma) *If $h(p, p_0) \geq \delta$, then the likelihood ratio satisfies:*

$$\mathbb{P}_{X_1, X_2, ..., X_n \sim p_0}(\frac{\mathcal{L}_n(p)}{\mathcal{L}_n(p_0)} \geq \exp(-\frac{n\delta^2}{2})) \leq \exp(-\frac{n\delta^2}{4})$$

Proof: According to the definition of sieve, there exists $p^* \in \widetilde{\mathcal{P}}$, such that $h(p_0, p^*) \leq \varepsilon$. If $\widehat{p} = p^*$, then $h(\widehat{p}, p_0) \leq 2\varepsilon$ is satisfied. Therefore, if $h(\widehat{p}, p_0) > 2\varepsilon$, we must have $\widehat{p} \neq p^*$. This effectively means that there exists $p \in \widetilde{\mathcal{P}}$ such that $h(p, p^*) \geq (2-1)\varepsilon = \varepsilon$ and that $\mathcal{L}_n(p) \geq \mathcal{L}_n(p^*)$. This means:

$$\begin{aligned}
\mathbb{P}(h(\widehat{p}, p_0) \geq 2\varepsilon) \leq & \mathbb{P}(\sup_{p \in \widetilde{\mathcal{P}}, h(p,p^*) \geq \varepsilon} \frac{\mathcal{L}_n(p)}{\mathcal{L}_n(p^*)} \geq 1) \\
= & \mathbb{P}(\sup_{p \in \widetilde{\mathcal{P}}, h(p,p^*) \geq \varepsilon} \frac{\mathcal{L}_n(p)}{\mathcal{L}_n(p_0)} \frac{\mathcal{L}_n(p_0)}{\mathcal{L}_n(p^*)} \geq 1) \\
\leq & \mathbb{P}(\sup_{p \in \widetilde{\mathcal{P}}, h(p,p^*) \geq \varepsilon} \frac{\mathcal{L}_n(p)}{\mathcal{L}_n(p_0)} \geq \exp(-\frac{n\varepsilon^2}{2})) + \mathbb{P}(\frac{\mathcal{L}_n(p_0)}{\mathcal{L}_n(p^*)} \geq \exp(\frac{n\varepsilon^2}{2}))
\end{aligned}$$

According to Wong-Shen lemma and union bound, the first term is upper-bounded by:

$$\begin{aligned}
& \mathbb{P}(\sup_{p \in \widetilde{\mathcal{P}}, h(p,p^*) \geq \varepsilon} \frac{\mathcal{L}_n(p)}{\mathcal{L}_n(p_0)} \geq \exp(-\frac{n\varepsilon^2}{2})) \leq N(\varepsilon) \exp(-\frac{n\varepsilon^2}{4}) \\
& \leq \exp(\frac{n\varepsilon^2}{8}) \exp(-\frac{n\varepsilon^2}{4}) = \exp(-\frac{n\varepsilon^2}{8})
\end{aligned}$$

Meanwhile, according to Markov inequality, the second term is upper-bounded by:

$$\begin{aligned}
\mathbb{P}(\frac{\mathcal{L}_n(p_0)}{\mathcal{L}_n(p^*)} \geq \exp(\frac{n\varepsilon^2}{2})) \leq & \mathbb{E}_{p_0} \prod_{i=1}^{n} \frac{p_0(X_i)}{p^*(X_i)} \exp(-\frac{n\varepsilon^2}{2}) \\
= & \exp(-\frac{n\varepsilon^2}{2}) (\int \frac{p_0^2}{p^*})^n \\
= & \exp(-\frac{n\varepsilon^2}{2}) (1 + \chi^2(p_0, p^*))^n \\
\leq & \exp(-\frac{n\varepsilon^2}{2}) \exp(n\chi^2(p_0, p^*)) \\
\leq & \exp(-(\frac{1}{2} - U)\varepsilon^2) \leq \exp(-\frac{n\varepsilon^2}{8})
\end{aligned}$$

In conclusion, we have:

$$\mathbb{P}(h(\widehat{p}, p_0) \geq 2\varepsilon) \leq 2 \exp(-\frac{1}{8} n\varepsilon^2)$$

Q.E.D. ∎

19.2 Sieve Estimation in the TV/ℓ_1 metric

In this section, we discuss a similar problem using a method devised by Yannis Yatracos.

19.2.1 Problem and method

Again, we have n samples $X_1, X_2, ..., X_n \sim p_0 \in \mathcal{P}$. Now, we want a non-parametric estimator $\widehat{p}$ such that the total variance between $\widehat{p}$ and p_0 , $TV(\widehat{p}, p_0)$, is small. In order to do this, we use the following method:

(1) Construct an ε- TV cover of $\mathcal{P}$: $\widetilde{\mathcal{P}} = \{p_1, p_2, ..., p_{N(\varepsilon)}\}$, such that $\forall p \in \mathcal{P}$, there exists a $p^* \in \widetilde{\mathcal{P}}$ such that $TV(p, p^*) \leq \varepsilon$.

(2) Construct a family of **Yatracos sets** A_{ij} for $1 \leq i < j \leq n$, with $A_{ij} = \{x : p_i(x) > p_j(x)\}$. Denote the family of all Yatracos sets as $\mathcal{A}$. Apparently, $|\mathcal{A}| < N^2(\varepsilon)$, and we have

$$TV(p_i, p_j) = |p_i(A_{ij}) - p_j(A_{ij})|.$$

(3) Pick $\widehat{p} = \text{argmin}_{p \in \widetilde{\mathcal{P}}} \Delta_n(p)$, in which

$$\Delta_n(p) = \sup_{A \in \mathcal{A}} |p(A) - p_n(A)|$$

with $p_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \in A)$.

In the following parts, we will show that this is a good way to guarantee that $TV(\widehat{p}, p_0)$ is small.

19.2.2 A heuristic condition

In this part, we consider a heuristic condition in which p_0 is included in $\widetilde{\mathcal{P}}$. In this case, we have:

$$\Delta_n(p_0) = \sup_{A \in \mathcal{A}} |p_0(A) - \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \in A)| \leq C \sqrt{\frac{\log(N^2(\varepsilon)/\delta)}{2n}}$$

with probability of at least $1 - \delta$, according to Hoeffding's inequality and union bound.

On the other hand, consider every $p \in \widetilde{\mathcal{P}}, p \neq p_0$, such that $TV(p, p_0) \geq \varepsilon$. Further define $A^* = \{x : p(x) > p_0(x)\}$, then we get:

$$\begin{aligned}\Delta_n(p) =& \sup_{A\in\mathcal{A}} |p(A) - p_n(A)| \\ \geq& |p(A^*) - p_n(A^*)| \\ \geq& |p(A^*) - p_0(A^*)| - |p_0(A^*) - p_n(A^*)| \\ \geq& \varepsilon - \sqrt{\frac{\log N(\varepsilon)}{n}}\end{aligned}$$

with high probability, using Hoeffding's bound. Notice that $|p(A^*) - p_0(A^*)| = TV(p, p_0) \geq \varepsilon$. Therefore, as long as $\varepsilon - \sqrt{\frac{\log N(\varepsilon)}{n}} \gg \sqrt{\frac{\log N(\varepsilon)}{n}}$, which is implied by $n\varepsilon^2 \gg \log N(\varepsilon)$, $\Delta_n(p) > \Delta_n(p_0)$, which means p cannot be $\widehat{p}$. In conclusion, as long as we can find ε such that $n\varepsilon^2 \gg \log N(\varepsilon)$, we have $TV(\widehat{p}, p_0) \leq \varepsilon$ with high probability.

19.2.3 General case

The previous part considers a rare condition in which the real distribution p_0 is accidentally in $\widetilde{\mathcal{P}}$. If this is not the case, we can still guarantee that $TV(\widehat{p}, p_0) \leq 2\varepsilon$ with high probability. According to the definition of $\widetilde{\mathcal{P}}$, we can find a $p^* \in \widetilde{\mathcal{P}}$ such that $TV(p_0, p^*) \leq \varepsilon$. Now we go on to argue that for any $p \in \widetilde{\mathcal{P}}$ such that $TV(p, p_0) \geq 2\varepsilon$, p cannot be $\widehat{p}$ as $\Delta_n(p) > \Delta_n(p^*)$. This is because:

(1) $\Delta_n(p^*) = \sup_{A\in\mathcal{A}} |p_n(A) - p^*(A)| \leq TV(p_0, p^*) + \sup_{A\in\mathcal{A}} |p_n(A) - p_0(A)| \leq \varepsilon + C\sqrt{\frac{2\log N(\varepsilon)}{n}}$ with high probability, according to Hoeffding's bound and union bound.

(2) Due to the same reason as that in the previous part, $\Delta_n(p) \geq 2\varepsilon - C\sqrt{\frac{\log N(\varepsilon)}{n}}$ with high probability.

So in conclusion, no matter whether $p_0 \in \widetilde{\mathcal{P}}$, as long as $n\varepsilon^2 \gg \log N(\varepsilon)$, we have $TV(\widehat{p}, p_0) \leq 2\varepsilon$ with high probability.

19.3 Robust density estimation

In the problem of robust density estimation, the true distribution is not in the family of candidates. Formally, we have $X_1, X_2, ..., X_n \sim p_0 \notin \mathcal{P}$, but we still want to find an estimator $\widehat{p} \in \widetilde{\mathcal{P}}$, such that

$$TV(\widehat{p}, p_0) \leq C \min_{p\in\widetilde{\mathcal{P}}} TV(p_0, p) + \sqrt{\frac{\log N(\varepsilon)}{n}}$$

For example, in Huber's model, we have $p_0 = (1-\varepsilon)P + \varepsilon Q$, where ε is a small positive number, $P \in \mathcal{P}$, $Q \notin \mathcal{P}$ is some arbitrary noise. If $\widehat{p}$ is constant, then we get $TV(\widehat{p}, p_0) \leq \varepsilon$. We will discuss this in the following lecture.