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

Lecture 20: April 7

Lecturer: Siva Balakrishnan *Scribe: Tanya Marwah*

In the previous lectures we studied about the Sieve estimators in the Hellinger metric as well as the TV/ℓ_1 metric (where we introduced the Yatracos Estimator). In both the cases we considered the case where $X_1, X_2, \ldots, X_n \sim p_0 \in \mathcal{P}$ and we wished to find $\widehat{p}$ such that either $h(\widehat{p}, p_0) \leq 2\epsilon$ or $\text{TV}(\widehat{p}, p_0) \leq 2\epsilon$. In general when it comes to density estimation type problems, a way to go would be to put down a net and find a member of the net using an appropriate rule (KL or Hellinger perhaps go with likelihood, TVgo with Yatracos sets.)

In today's lecture we will first talk about robust estimation, where we consider cases when $X_1, X_2 \ldots X_n \sim p_0 \notin \mathcal{P}$. We will then move our discussion forward to non paramteric least squares.

20.1 Robust Estimation

We assume that perhaps the data comes from a different that doesn't belong to the class $\mathcal{P}$, i.e., $X_1, X_2, \ldots, X_n \sim p_0 \notin \mathcal{P}$.

There are two ways to think about this setup,

- Model Misspecification: Where we assume that even though $p_0 \notin \mathcal{P}$ it is still close in some metric like TV/H^2.
- Corrupted Data: Some ϵ fraction of $X_1, \ldots, X_n$ is replaced by some arbitrary value. An example for this is the Huber's contamination model where we assume that $X \sim (1-\epsilon)P + \epsilon Q$ where ϵ is a small positive and $P \in \mathcal{P}$ and Q is some arbitrary noise such that $Q \notin \mathcal{P}$. This mixture model has the property that $\text{TV}((1-\epsilon)P + \epsilon Q, P) \leq \epsilon$.

  This is a special case of Model Misspecification.

The broad goal of robust estimation is to find estimators whose performance degrades "gracefully" with ϵ (i.e., estimators with high breakdown point). To elaborate, given that $p_0 \notin \mathcal{P}$ we wish to estimator $\widehat{p} \in \mathcal{P}$ such that for metric ρ,

$$\rho(\widehat{p}, p_0) \leq c \inf_{p \in \mathcal{P}} \rho(p, p_0) + m$$

Where m is some form of complexity measure of the class $\mathcal{P}$ which is usually governed by some sort of Le Cam's equation.

Suppose we consider the Yatracos estimator we discussed in the last lecture. Formally, we find $\eta > 0$ such that,

$$n\eta^2 \gg \log N(\eta),$$

where $N(\eta)$ is the TV covering number of $\mathcal{P}$. Now, we construct an η covering of $\mathcal{P}$ which we denote by $\{p_1, \ldots, p_N\}$. We now construct the collection of Yatracos sets $\mathcal{A} = \{A_{ij} : p_i > p_j\}$, and select:

$$\widehat{p} = \arg \inf_{p \in \{p_1, \ldots, p_N\}} \sup_{A \in \mathcal{A}} |p_n(A) - p(A)|.$$

Our analysis from last lecture with minor modifications then shows that with probability at least $1 - \delta$,

$$\mathrm{TV}(p_0, \widehat{p}) \leq C_1 \inf_{p \in \mathcal{P}} \mathrm{TV}(p, p_0) + C_2 \sqrt{\frac{\log(N/\delta)}{n}}.$$

You will explore this more carefully in your HW. It is worth noting that likelihood based procedures typically are poorly behaved when model is misspecified. This is intuitively due to the fact that the likelihood (which is a product of probabilities), and likelihood ratios, are not very robust. For instance, the likelihood of a dataset with a single corruption can be drastically different from the likelihood of its uncorrupted counterpart.

20.2 Non Parametric Least Squares

Setup: We have $(x_1, y_1), \ldots (x_n, y_n) \sim P_{XY}$ and

$$y_i = f^*(x_i) + \sigma w_i, \quad w_i \sim \mathcal{N}(0, 1) \tag{20.1}$$

And we wish to estimate f^*.

There can be different measures of the quality of the estimate f of the regression function,

- $L^2(\mathbb{P})$ metric: $\mathbb{E}_X[(f(x) - f^*(X))^2]$ also written as $\|f^* - f\|_{L^2(\mathbb{P})}$ This metric is usually used in the random design setup when we know that the x_i's are sampled from some distribution.
- $L^2(\mathbb{P}_n)$ metric: We use this metric when we have fixed number of samples $x_1, \ldots x_n$, also known as fixed design setup where we assume that the samples are fixed and the only randomness is being added by the w_i's . We define an empirical distribution $\mathbb{P}_n$ and the associated $L^2(\mathbb{P}_n)$ norm is given by

$$\|f - f^*\|_{L^2(\mathbb{P}_n)} = [\frac{1}{n} \sum_{i=1}^{n} (\widehat{f}(x_i) - f^*(x_i))^2]^{1/2}$$

In this class we will mainly focus on the errors in $L^2(\mathbb{P}_n)$. Note that we will use the shorthand $\|\widehat{f} - f^*\|_n$ for $L^2(\mathbb{P}_n)$ norm.

The estimator that we will be analysing for non paramteric least squares is defined as,

$$\widehat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - f(x_i))^2 \right\} \tag{20.2}$$

We will talk about the cases with regularization and situations when $f^* \notin \mathcal{F}$ in the subsequent lectures. Possible examples for the collection of functions $\mathcal{F}$ can be Lipschitz functions, or all convex functions on the domain $[0, 1]$. Hence, note that the error between the estimated function $\widehat{f}$ and f^* should depend upon the size of the collection of functions $\mathcal{F}$ and now the size affects the rate of convergence.

The complexity term that we will be working with is called *local Gaussian width* and is defined as,

$$\mathcal{G}_n(\delta) = \mathbb{E}_w \left[\sup_{\substack{f \in \widetilde{\mathcal{F}} \\ \|f\|_n \leq \delta}} \left| \frac{1}{n} \sum_{i=1}^{n} w_i f(x_i) \right| \right]. \tag{20.3}$$

Where $\widetilde{\mathcal{F}} = \mathcal{F} - f^* = \{f : f = g - f^*, g \in \mathcal{F}\}$. Constrasting it to *global Gaussian width* with is defined as,

$$\mathcal{G}_n(\delta) = \mathbb{E}_w \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} w_i f(x_i) \right| \right]. \tag{20.4}$$

That is we are looking at the Gaussian width around the functions $f \in \widetilde{\mathcal{F}}$ that have small $L^2(\mathbb{P}_n)$ norm. (A way to think of it would be that we are likely to be confused by functions that are around f^* and we would like to know many functions are there in this small set around f^*.)

Now we will introduce *Critical Radius*, which will play a central role in our analysis. The critical radius is defined as the set of positive scalars δ that satisfy the *critical inequality*,

$$\frac{\mathcal{G}_n(\delta)}{\delta} \leq \frac{\delta}{2\sigma}. \tag{20.5}$$

In the equation above the RHS is an increasing function in δ and we will verify later that the LHS will be decreasing in δ under the assumption that $\widetilde{\mathcal{F}}$ is a star-shaped function.

Star Shaped function classes: A function class $\widetilde{\mathcal{F}}$ is star-shaped around f^* if for any $\alpha \in [0, 1]$, the function $\alpha f^* \in \widetilde{\mathcal{F}}$. In other words, all the functions that lie on the line connecting f and f^* should also lie in $\widetilde{\mathcal{F}}$. A stronger assumption is that the class of functions is convex, as convexity will imply that the function class is star shaped (around every possible function $f^* \in \widetilde{\mathcal{F}}$.

Hence, if we can find a δ that the critical inequality is satified then we can find the rate of convergence of the least squares error in terms of this δ.

Theorem 20.1 *suppose that our function class $\widetilde{\mathcal{F}}$ is star-shaped, and let δ_n be any any solution to the critical inequality defined in equation 20.5. Then for any $t \geq \delta_n$, the non-parametric least-sqaures estimate $\widehat{f}$ satisfies the bound,*

$$\Pr(\|\widehat{f} - f\|_n^2 \geq 16t\delta_n) \leq \exp\{-\frac{nt\delta_n}{2\sigma^2}\}. \tag{20.6}$$

In this lecture we will just go through how the proof goes about heuristically,

Proof Sketch: We first start with the basic inequality,

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(x_i) - y_i)^2 \leq \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2. \tag{20.7}$$

Given that $y_i = f^(x_i)+\sigma w_i$, let $\widehat{\Delta} = \widehat{f} - f^*$ and after rearranging some terms in equation 20.7, we get,

$$\frac{1}{2}\|\widehat{f} - f^*\|_n^2 \leq \frac{\sigma}{n}\sum_{i=1}^{n} w_i\widehat{\Delta}(x_i). \tag{20.8}$$

Given that $\widehat{\Delta} \in \widetilde{\mathcal{F}}$ we can upper bound the term in the RHS by taking a supremum over all functions $g \in \widetilde{\mathcal{F}}$ with $\|g\|_n \leq \|\widehat{\Delta}\|_n$, and therefore after taking expectation over both the sides in equation 20.8, we have,

$$\frac{\mathbb{E}\|\widehat{\Delta}\|_n^2}{2} \leq \mathbb{E}\Bigg[\sup_{\substack{\|g\|_n \leq \|\widehat{\Delta}\|_n \\ g \in \widetilde{\mathcal{F}}}} \left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i g_i\right|\Bigg]. \tag{20.9}$$

Now, reasoning heuristically, if we assume that $\mathbb{E}\|\widehat{\Delta}\|_n^2 = \delta_n^2$ and in the RHS of equation 20.9, instead of $\|g\|_n \leq \|\widehat{\Delta}\|_n$ where $\|\widehat{\Delta}\|_n$ is a quantity we don't know, we replace it by δ_n, we get,

$$\frac{\delta_n^2}{2} \leq \mathbb{E}\Bigg[\sup_{\substack{\|g\|_n \leq \delta_n \\ g \in \widetilde{\mathcal{F}}}} \left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i g_i\right|\Bigg]$$

$$\implies \frac{\delta_n^2}{2} \leq \underbrace{\sigma\,\mathbb{E}\Bigg[\sup_{\substack{\|g\|_n \leq \delta_n \\ g \in \widetilde{\mathcal{F}}}} \left|\frac{1}{n}\sum_{i=1}^{n} w_i g_i\right|\Bigg]}_{\text{the local Gaussian width}}.$$

Hence from our current heuristic argument we get

$$\frac{\delta_n^2}{2} \lesssim \sigma\mathcal{G}_n(\delta_n). \tag{20.10}$$

Therefore if we can find a δ that satisfies the critical inequality in equation 20.5, it will potentially be an upper bound on the δ_n that satisfies equation 20.10, i.e., if $\frac{\mathcal{G}_n}{\delta} \leq \frac{\delta}{2\sigma}$ and $\frac{\delta_n^2}{2} \lesssim \sigma \mathcal{G}_n(\delta_n)$ implies that $\delta_n \leq \delta$.

Assuming that δ^* is the smallest solution to equation 20.5, then we have something like $\delta_n \leq \delta^*$, and given that we assumed that $\delta_n^2 := \mathbb{E}\|\widehat{\Delta}\|_n^2$ it implies that $\mathbb{E}\|\widehat{\Delta}\|_n^2 \leq \delta^*$.