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

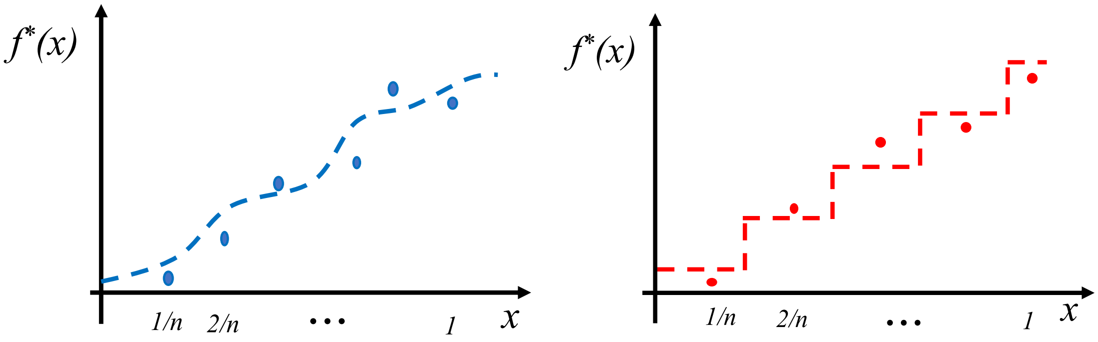
Lecture 22: April 14

Lecturer: Siva Balakrishnan *Scribe: Ankur Mallick*

In this lecture we will discuss the adaptivity of the non-parametric least squares estimator in the absence of special knowledge about the structure of the function. We will also study an oracle inequality when the true function does not belong to the class of functions used for least squares estimation.

22.1 Adaptivity of the Least Squares Estimator

Consider estimating a smooth monotone function f^* for which we observe noisy samples at fixed points on a grid as shown in Figure 22.1



(a) Monotone function

(b) Piecewise Constant function

Figure 22.1: We observe noisy samples (represented by dots) of the original monotone/piecewise constant function (represented by the dotted lines)

To analyse non-parametric least squares we note that its metric entropy is

$$\log \mathcal{N}(u) \leq \frac{cV}{u} \tag{22.1}$$

where $V = f_{\max} - f_{\min}$.

Due to this the rate at which we can estimate a monotone function as in Figure 22.1a is

$$\|\widehat{f} - f^*\|_n^2 \leq \left(\frac{\sigma^2 V}{N}\right)^{2/3} \tag{22.2}$$

which is the same as that for Lipschitz functions.

If f^* is monotone and has k constant pieces as in Figure 22.1b then

$$\|\widehat{f} - f^*\|_n^2 \lesssim \frac{k \log n}{n} \tag{22.3}$$

In this case least squares adapts to give the faster rate even if we don't know k because the local Gaussian Width is much smaller for such functions than for monotone functions [1]. This is remarkable because we did not tailor the estimator to any assumptions about structure. The original estimator adapts on its own.

This property that the local Gaussian width can be a highly variable function of the unknown parameter θ^* also underlies the adaptivity properties of the soft and hard thresholding estimators we have studied previously.

22.2 Oracle Inequality

Theorem 22.1 *Let δ_n be any positive constant solution to the inequality*

$$\frac{\mathcal{G}_n(\delta; \partial\mathcal{F})}{\delta} \leq \frac{\delta}{2\sigma} \tag{22.4}$$

Then there are universal positive constants (C, c_0, c_1, c_2) such that for any $t \geq \delta_n$ the non-parametric least squares estimate $\widehat{f}$ satisfies the bound

$$\mathbb{P}\left(\|\widehat{f} - f^*\|_n^2 \geq C \inf_{f \in \mathcal{F}} \|f - f^*\|_n^2 + ct\delta_n\right) \leq c_1 \exp\left(-c_2 \frac{nt\delta_n}{2\sigma^2}\right) \tag{22.5}$$

Thus the least squares estimator automatically achieves a bias-variance tradeoff.

Proof(Sketch): Recall the basic inequality

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \widehat{f}(X_i))^2 \leq \frac{1}{2n}\sum_{i=1}^{n}(y_i - \widetilde{f}(X_i))^2 \quad \forall \widetilde{f} \in \mathcal{F} \tag{22.6}$$

Since $y_i = f^*(X_i) + \sigma w_i$ where w_i's are standard normal random variables and σ is the noise standard deviation, we get the following modified basic inequality

$$\frac{1}{2}\|\widehat{\Delta}\|_n^2 \leq \frac{1}{2}\|\widetilde{f} - f^*\|_n^2 + \left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i(\widetilde{f} - \widehat{f})\right| \tag{22.7}$$

where $\widehat{\Delta} = \widehat{f} - f^*$. The rest of the proof involves considering the cases $\|\widehat{\Delta}\|_n \leq \sqrt{t\delta_n}$ and $\|\widehat{\Delta}\|_n \geq \sqrt{t\delta_n}$ separately. In the first case the probabilistic upper bound is zero while in the second case the bound is obtained by applying the standard Gaussian tail bound.

Note that all our analysis has been for the fixed design case due to which we only care about empirical norms. The results may not be the same for random design where we care about population norm.

Some examples of applications of the oracle inequality are

1. We can use it to bound the prediction error in paramteric regression as

$$\frac{\|\mathbf{X}\widehat{\theta} - \mathbf{X}\theta^*\|_2^2}{n} \lesssim \inf_{\|\theta\|_0 \leq s} \frac{\|\mathbf{X}\theta - \mathbf{X}\theta^*\|_2^2}{n} + \frac{s \log p}{n} \tag{22.8}$$

   where θ^* is arbitrary.

2. We can use it to improve the estimation of Sobolev functions with $2\alpha < d$. The naive estimate is given by

$$\widehat{f} = \arg\min_{f \in \mathcal{F}_\alpha} \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(X_i))^2 \tag{22.9}$$

   where $\mathcal{F}_\alpha$ is the function class. Functions in this class can be highly irregular (in particular, they can be discontinuous). This estimate will be an interpolation i.e. it will satisfy $\widehat{f}(X_i) = y_i$ and will not be consistent i.e. even if $f^* \in \mathcal{F}_\alpha$ the least squares estimate will do poorly. Therefore we can instead construct a simpler class of functions $\widetilde{\mathcal{F}}_\alpha$ where the least squares estimate is sensible and then use the oracle inequality to characterise how far $\widehat{f}$ is from f^* through the bias term (we can also try to balance the bias-variance tradeoff). This idea is closely related to the idea of using sieves that we discussed previously. In particular, one can view the space $\widetilde{\mathcal{F}}_\alpha$ as a sieve of the space $\mathcal{F}_\alpha$ – so sieves, in addition to simplifying the analysis of our estimators can also provide explicit bias-variance control that the direct least squares estimator might not have.

References

[1] Adityanand Guntuboyina, Bodhisattva Sen, et al. Nonparametric shape-restricted regression. *Statistical Science*, 33(4):568–594, 2018.