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

Lecture 25: April 23

Lecturer: Siva Balakrishnan *Scribe: Saurabh Garg*

25.1 Recap

In the last two lectures, we covered minimax hypothesis testing where we saw methods to construct lower bounds; a key technique/idea we learned is to transform to a simple null vs simple alternate. As an application, we saw the test for mean estimation where we design a Neyman-Pearson test with a Bayesian counterpart. Today, we will discuss a few interesting things in functional estimation.

25.2 Functional Estimation

Lets consider a unknown distribution f which we don't have access to, but rather we have access to samples $X_1, X_2, \ldots, X_n$ sampled iid from f. Define a functional $T(f)$ which captures a univariate property of f. Our main aim in functional estimation is to come up with an estimator for $T(f)$. We will consider the case when $T(f)$ takes the form $\int \phi(f)$, where ϕ is a smooth function. For example, $T(f)$ can be entropy, i.e., $\int f \log(f)$ or a quadratic function $\int f^2$.

We also consider "bivariate" functions $T(f, g)$ that are defined with two distributions f and g. The setting here is similar, we have access to samples $X_1, X_2, \ldots, X_n \sim f$ and $Y_1, Y_2, \ldots, Y_n \sim g$. Define $T(f, g) = \int \phi(f, g)$. As an example, $T(f, g)$ can take the form of divergence measure like KL divergence or Hellinger distance.

These are in some sense simple functionals. Another common application of functional estimation is in causal inference, where typically the functionals of interest are considerably more complex. One example of complex functional estimation is estimating *Average Treatment Effect* (ATE) τ in causal inference. The problem roughly is that we observe i.i.d samples $\{(X_1, A_1, Y_1), \ldots, (X_n, A_n, Y_n)\}$ where covariates X sampled from P_X, A is a Bernoulli random variable sampled with parameter gives by the propensity score function $\pi(X)$ and $Y(A)$ is the outcome. We wish to measure the casual association (the ATE) of the potential outcomes, i.e.

$$\tau = \mathbb{E}[Y(1) - Y(0)]\,,$$

which is the difference in outcomes if all units were treated versus all were in the control group. Notice that in this case the functional of interest is a difference of regression functions,

and the underlying distribution is complex (there are lots of different pieces, the covariate density, the propensity score function and the regression functions).

Returning to simple functionals: throughout the lecture we will focus on a canonical functional estimation problem – of estimating the integral of the square of the density. One naïve estimator to try is the so called plug-in estimator: (i) Estimate the function $\widehat{f}$ from n samples; (ii) Use $T(\widehat{f})$ as an estimate to $T(f)$.

Example. Consider $X_1, X_2, \ldots, X_n \sim f$, where f is a β-smooth function. Define $T(f) = \int f^2$. We will show that with using a plugin estimate $\widehat{f}$ of f can lead to undesirable bias in the estimate of the function $T(f)$.

Say $\widehat{f}$ is a kernel density estimator with fixed and decided parameters. Construct $\widehat{T} = T(\widehat{f})$.

$$\left|\widehat{T} - T\right| \leq \int \widehat{f}^2 - f^2 \approx \int (\widehat{f} - f)(\widehat{f} + f) \leq (\|\widehat{f}\|_\infty + \|f\|_\infty) \int |\widehat{f} - f|.$$

Assume densities are bounded in $\|\cdot\|_\infty$. Then it can be shown that $\mathbb{E}|\widehat{T} - T|^2 \approx Cn^{-2\beta/(2\beta+d)}$. This turns out to be far from the best possible estimate of $T(f)$. The rough problem is that when we square the estimator, we accumulate more bias, i.e., even if hypothetically $\widehat{f}$ was an unbiased estimator of f (it typically is not), $\widehat{f}^2$ would not be an unbiased estimator of f^2.

25.2.1 Bias Correction

Consider the Taylor expansion of $\int f^2$, i.e.

$$\int f^2 = \int [\widehat{f}^2 + 2\widehat{f}(f - \widehat{f}) + \underbrace{(f - \widehat{f})^2}_{\text{remainder term}}]$$

$$T(f) = T(\widehat{f}) + \int 2\widehat{f}(f - \widehat{f}) + \underbrace{\int (f - \widehat{f})^2}_{\approx n^{-2\beta/(2\beta+d)}}$$

Consider the middle term. It can be re-written as $2\int \widehat{f} f - 2\int \widehat{f}^2$. The first term is $2\mathbb{E}\widehat{f}$ which can be estimated at "$1/\sqrt{n}$ rates". This method is also known as *Influence function correction*. Consider the corrected estimator

$$\widehat{T} = \frac{2}{n} \sum_{i=1}^{n} \widehat{f}(x_i) - \int \widehat{f}^2 \,.$$

With the correction, we will obtain a rate on $\mathbb{E}[\widehat{T} - T]^2$ that looks like $n^{-4\beta/(2\beta+d)}$ for $\beta \leq d/2$. For $\beta > d/2$, the rate is n^{-1}.

Moreover, it turns out that the minimax rates for $\mathbb{E}[\widehat{T} - T]^2$ is $n^{-8\beta/(4\beta+d)}$ if $\beta \leq d/4$. For $\beta > d/4$, we have n^{-1} rate. With just first-order correction, we got the rate that looks closer to optimal than the plug-in.

In general, the first-order corrected estimator will have the property that when $s > d/2$ we get n^{-1} rates but furthermore the estimator will have an associated CLT and will have minimal possible variance (amongst sufficiently regular estimators). There is an entire part of statistical theory that has been developed to reason about optimality in functional estimation type problems (this is the area of semi-parametric inference). Recall and contrast this with Influence Functions and Regular Asymptotically Linear Estimators from Lecture 19, 705.

25.2.2 Minimax Rates

One might ask "how one gets optimal rates in these settings". We will try to do this for the integral of the square of the density by using a truncated series estimator.

Consider the function f in an orthogonal basis ϕ s.t. $\int \phi_i^2 = 1$ and for all $j \neq i$, we have $\int \phi_j \phi_i = 0$. Assume $f = \sum_i \theta_i \phi_i$, where $\theta_i = \int \phi_i f$ and define the estimator for θ_i as $\widehat{\theta}_i = \sum_{j=1}^n \phi_i(x_j)/n$.

While constructing an estimator for f, we decide the number of terms to include by trading off the bias with variance. But while constructing an estimator for $T(f) = \int f^2$ in the naïve way, we still would have the bias issue because we will be squaring $\widehat{\theta}_i$s.

To build some intuition it will be useful to first consider the problem of estimating a simple parametric functional. Consider a simple problem where we have $X_1, X_2, \ldots, X_{2n} \sim \mathcal{N}(\theta^*, I_d)$ and we want to estimate $T(\theta^*) = \sum_{i=1}^d \theta_i^{*2}$. There are several possible ways to estimate this functional without any bias. The classical way would be to design an appropriate U-statistic, but we will do something even simpler (but less practical).

Split the samples into two equal parts and consider sample mean on each part separately, i.e., let $\widehat{\theta}_1 = \sum_{i=1}^n x_i/n$ and $\widehat{\theta}_2 = \sum_{i=1}^n x_{i+n}/n$. Construct $\widehat{T} = \sum_{j=1}^d \widehat{\theta}_{1j}\theta_{2j}$. Clearly, with sample splitting we have eliminated the bias, i.e., $\mathbb{E}(\widehat{T}) = \sum_{j=1}^d \theta_j^{*2} = T$. Thus with this, we get

$$\begin{aligned}\mathbb{E}[\widehat{T}-T]^2 &= \mathrm{Var}(\widehat{T}) \\ &= \mathbb{E}[\widehat{T}]^2 - T^2 \\ &= \mathbb{E}\left[\sum_{j=1}^{d}\sum_{k=1}^{d}\widehat{\theta}_{1j}\widehat{\theta}_{1k}\widehat{\theta}_{2j}\widehat{\theta}_{2k} - \sum_{j=1}^{d}\sum_{k=1}^{d}\theta_j^{*2}\theta_k^{*2}\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{d}\widehat{\theta}_{1j}^2\widehat{\theta}_{2j}^2 - \sum_{j=1}^{d}\theta_j^{*4}\right].\end{aligned}$$

Now consider the term $\mathbb{E}[\widehat{\theta}_{ij}^2]$. This expectation of square of a random variable sampled from $\mathcal{N}(\theta_j^*, 1/n)$ which is nothing but $\theta_j^{*2} + 1/n$. Hence, we get $\mathbb{E}[\widehat{T}-T]^2 = d/n^2 + 4\sum_j \theta_j^{*2}/n$. Suppose that we assume that $\|\theta^*\|_2 \le C$ (for some universal constant). Then as long as $d < n$, we obtain the same rate as we would for a one dimensional problem which is $1/n$.

The fact that we obtain rates where the dimension dependent term scales as d/n^2 is related to the minimax testing rates we obtained in the previous lectures. In particular, if you used $\widehat{T}$ as a test statistic, you will be able to convince yourself that you can reproduce the same upper bounds as we did before using the χ^2 test, i.e. you will see that if $\|\theta^*\| \gg d^{1/4}/\sqrt{n}$ a test based on $\widehat{T}$ will have non-trivial power.

Going back to our density estimation example where we want to estimate a functional $T(f) = \int f^2$. We will just sketch the high-level ideas. Using the idea discussed here, we truncate the series with K terms and pay for the bias separately, i.e., we estimate $\sum_{i=1}^{K}\theta_i^{*2}$ with sample splitting (notice that our above argument did not really use normality, and would continue to work with our estimator for the coefficients of the truncated series). With this, we get

$$\mathbb{E}[\widehat{T}-T]^2 = \underbrace{\Big(\sum_{i=K+1}^{\infty}\theta_i^{*2}\Big)^2}_{\text{bias}} + \underbrace{\frac{K}{n^2} + \frac{4\sum_i \theta_i^{*2}}{n}}_{\text{variance}}$$

Consider *Sobolev ellipsoids* where $\sum_{j=1}^{\infty}\theta_j^* j^{2\beta} \le 1$ and $\sum_{j=1}^{\infty}\theta_j^{*2} \le C$ for some constant C. The (bias)^2 term evaluates to $K^{-4\beta}$, since

$$\sum_{i=K+1}^{\infty}\theta_i^{*2} \le K^{-2\beta}\sum_{i=K+1}^{\infty}\theta_i^{*2} i^{2\beta} \le K^{-2\beta}.$$

Now, balancing the variance and bias squared in the usual way we would choose $K \approx n^{2/(4\beta+1)}$. Thus, we get the desired optimal rates which is $\max\left\{n^{-1}, n^{-8\beta/(4\beta+1)}\right\}$. The

argument also extends in a straightforward way to higher dimensions (just by appropriately modifying the definition of the Sobolev ellipsoid).