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

Lecture 24: April 21

Lecturer: Siva Balakrishnan *Scribe: Mike Stanley*

24.1 Recap of Minimax Hypothesis Testing

In the previous lecture, we introduced minimax hypothesis testing for the situation where we are given some data $X_1, \ldots, X_n \in P$ and we want to distinguish the following general hypotheses:

$$H_0 : P \in \mathcal{P}_0$$
$$H_1 : P \in \mathcal{P}_1(\epsilon) := \mathcal{P}_1 \cap \{P : \rho(P, \mathcal{P}_0) > \epsilon\}.$$

Further, we defined a test function $\phi_T(X_1, \ldots, X_n) \in \{0, 1\}$ ($\phi_T = 1$ means reject the null), which allowed us to define the risk of our test as a function of ϵ:

$$R_\epsilon(T) = \sup_{P \in \mathcal{P}_0} \mathbb{E}_P \phi_T + \sup_{P \in \mathcal{P}_1(\epsilon)} \mathbb{E}_P(1 - \phi_T) \tag{24.1}$$

and hence the minimax test is the one that minimizes the above risk, i.e.

$$\widehat{T} = \arg\min_T R_\epsilon(T) \tag{24.2}$$

We then defined the **critical radius** as some ϵ^* such that for some constants C_1 and C_2, $\inf_T R_{C_1\epsilon^*}(T) \leq \frac{1}{6}$ and $\inf_T R_{C_2\epsilon^*}(T) \geq \frac{1}{3}$. Roughly, when the null and alternate are separated by at least some large constant times the critical radius then the risk of the best test can be made small, and when the null and alternate are separated by less than some small constant times ϵ^* then no test can distinguish them (in the minimax sense). The choice of constants in this definition turn out to be unimportant. One way to see this is to see that if you had a non-trivial test (i.e. with risk < 1) then you could repeat the test a few times (on split samples) and combine the results – using a Hoeffding bound you can see that the risk of the test will get exponentially smaller in the number of times you repeat the test.

Additionally, we proved an upper bound on the risk for the gaussian mean testing problem showing that for the test statistic $T = \sum_{i=1}^d y_i^2$ and test function $\phi_T = \mathbf{1}\{T \geq \mathbb{E}_0 T + C\sqrt{\text{Var}_0 T}\}$, if $\mathbb{E}_1 T - \mathbb{E}_0 T \geq C\Big(\sqrt{\text{Var}_0(T)} + \sqrt{\text{Var}_1(T)}\Big)$, then $R(T) \leq \frac{1}{C^2}$.

Perhaps most importantly, we compared the scaling of the critical radius between this testing problem and estimation and noted that for the minimax testing problem

$$\epsilon^* \asymp \sigma \frac{d^{1/4}}{\sqrt{n}} \tag{24.3}$$

whereas for the estimation problem

$$\epsilon^* \asymp \sigma \sqrt{\frac{d}{n}} \tag{24.4}$$

In this lecture, we further explore why the discrepancy in these rates exists. In particular, we draw connections between the aforementioned results an functional estimation.

24.2 Looking at the test statistic as a functional to estimate

Suppose that we observe $Y \sim \mathcal{N}(\theta^*, \sigma^2 \mathbf{I}_d)$ and we are interested in estimating the functional $T = \sum_{i=1}^d \theta_i^{*2} = \|\theta^*\|_2^2$. Of course, we could directly estimate T using the plug-in estimator, i.e.

$$\widehat{T} = \sum_{i=1}^d y_i^2 \tag{24.5}$$

However, $\widehat{T}$ is biased, i.e. $\mathbb{E}\widehat{T} = T + \sigma^2 d$ and hence we obtain the scaling of the risk

$$\begin{aligned}
\mathbb{E}(\widehat{T} - T)^2 &= (\mathbb{E}\widehat{T} - T)^2 + \mathrm{Var}(\widehat{T}) \\
&= ((T + \sigma^2 d) - T)^2 + \mathrm{Var}(\widehat{T}) \\
&\asymp \sigma^4 d^2 + \mathrm{Var}(\widehat{T})
\end{aligned}$$

We could de-bias our estimator using $\widehat{T} = \sum_{i=1}^d y_i^2 - \sigma^2 d$. We will explore this idea of de-biasing further in the next lecture. It is however worth noting that for the purpose of testing this bias is irrelevant (it simply shifts the distribution of the statistic under both the null and the alternate but does not affect the risk of the test in any meaningful way).

24.3 Lower Bounds

Now, we want to argue that there isn't a better test. First, we define a different testing problem that allows us to operate in the world of testing a simple null against a simple alternative.

$$H_0 : \theta^* = 0$$
$$H_1 : \theta^* \sim \epsilon \mathbb{S}^{d-1}$$

where we then have $Y \sim \mathcal{N}(\theta^*, \sigma^2 \mathbf{I}_d)$. All of the vectors on $\epsilon \mathbb{S}^{d-1}$ are by definition ϵ away from $\mathcal{P}_0$.

Let us denote the distribution under the null as P_0 (just $N(0, \sigma^2 \mathbf{I}_d)$) and under the alternate by P_1 (this is the infinite mixture distribution defined above by sampling $\theta^* \sim \epsilon \mathbb{S}^{d-1}$ and $Y \sim \mathcal{N}(\theta^*, \sigma^2 \mathbf{I}_d)$.

Let's investigate the risk of the optimal test T. By the Neyman-Pearson lemma we know the optimal test is the likelihood ratio test (and we know its risk exactly).

$$\begin{aligned} R(T) &= 1 - \mathbf{TV}(P_0, P_1) \\ &= 1 - \frac{1}{2}\mathbb{E}_0 \mid \frac{p_1}{p_0} - 1 \mid \\ &= 1 - \frac{1}{2}\mathbb{E}_0 \mid L - 1 \mid \\ &\geq 1 - \frac{1}{2}\sqrt{\mathrm{Var}_0(L)} \\ &= 1 - \frac{1}{2}\sqrt{\chi^2(P_0, P_1)} \end{aligned}$$

where L is the likelihood ratio between the alternate and null distributions. In order to lower bound the minimax risk, it is thus sufficient to show that if the null and alternate are not sufficiently well-separated then $\mathrm{Var}_0(L) \to 0$.

We would like to understand the variance of L, i.e. $\mathbb{E}_0 L^2 - 1$. Let's re-write L in terms of the actual probability density functions.

$$\begin{aligned}
L &= \frac{p_1(y)}{p_0(y)} \\
&= \frac{\mathbb{E}_{\theta^* \sim \epsilon\mathbb{S}^{d-1}} \exp\{\frac{-\|y-\theta^*\|^2}{2\sigma^2}\}}{\exp\{\frac{-\|y\|^2}{2\sigma^2}\}} \\
&= \mathbb{E}_{\theta^* \sim \epsilon\mathbb{S}^{d-1}} \exp\{\frac{-\|\theta^*\|^2}{2\sigma^2} + \frac{y^T\theta^*}{\sigma^2}\} \\
&= \exp\{\frac{-\epsilon^2}{2\sigma^2}\}\mathbb{E}_{\theta^* \sim \epsilon\mathbb{S}^{d-1}} \exp\{\frac{y^T\theta^*}{\sigma^2}\}
\end{aligned}$$

Turning our attention back to the computation of $\mathbb{E}_0 L^2$:

$$\begin{aligned}
\mathbb{E}_0 L^2 &= \mathbb{E}_0\left[\exp\left\{\frac{-\epsilon^2}{\sigma^2}\right\}\mathbb{E}_{\theta^*,\widetilde{\theta}} \exp\left\{\frac{y^T(\widetilde{\theta}+\theta^*)}{2\sigma^2}\right\}\right] \\
&= \exp\left\{\frac{-\epsilon^2}{\sigma^2}\right\}\mathbb{E}_{\theta^*,\widetilde{\theta}} \exp\left\{\frac{\|\widetilde{\theta}+\theta^*\|_2^2}{2\sigma^2}\right\} \\
&= \exp\left\{\frac{-\epsilon^2}{\sigma^2}\right\}\mathbb{E}_{\theta^*,\widetilde{\theta}} \exp\left\{\frac{\epsilon^2}{2\sigma^2} + \frac{\epsilon^2}{2\sigma^2} + \frac{2\widetilde{\theta}^T\theta^*}{2\sigma^2}\right\} \\
&= \mathbb{E}_{\theta^*,\widetilde{\theta}} \exp\left\{\frac{\widetilde{\theta}^T\theta^*}{\sigma^2}\right\}
\end{aligned}$$

Now, we want to use the above form to argue that the variance of the likelihood ratio is small.

$$\begin{aligned}
\mathbb{E}_0 L^2 &= \mathbb{E}_{\theta^*,\widetilde{\theta}\sim\epsilon\mathbb{S}^{d-1}} \exp\left\{\frac{\widetilde{\theta}^T\theta^*}{\sigma^2}\right\} \\
&= \mathbb{E}_{\theta^*,\widetilde{\theta}\sim\mathbb{S}^{d-1}} \exp\left\{\frac{\epsilon^2\widetilde{\theta}^T\theta^*}{\sigma^2}\right\}
\end{aligned}$$

If $\theta^* \sim \sqrt{d}\mathbb{S}^{d-1}$, this implies that θ^* is C-subgaussian. For more information on this, see [2]. Thus, for a fixed vector, v,

$$\begin{aligned}
\mathbb{E}_0 L^2 &= \mathbb{E}_{\theta^* \sim \mathbb{S}^{d-1}} \exp\left\{\frac{\epsilon^2 \theta^{*T} v}{\sigma^2}\right\} \\
&= \mathbb{E}_{\theta^* \sim \sqrt{d}\mathbb{S}^{d-1}} \exp\left\{\frac{\epsilon^2 \theta^{*T} v}{\sqrt{d}\sigma^2}\right\} \\
&\leq \exp\left\{\frac{\epsilon^4 \|v\|^2 C^2}{2d\sigma^4}\right\} \\
&\leq 1 + \frac{2C^2\epsilon^4}{2d\sigma^4}
\end{aligned}$$

where the second to last line holds from the subgaussian condition, and the last line holds if $\frac{C^2\epsilon^4}{2d\sigma^4} \leq 1$. We justify this last line of reasoning by recalling that $\exp(x) \leq 1 + 2x$ if $x \leq 1$.

If $\frac{C^2\epsilon^4}{d\sigma^4} \to 0$, then $\mathrm{Var}_0(L) \to 0$ and $R(T) \geq 1 - \frac{1}{2}\sqrt{\mathrm{Var}_0(L)}$. I.e., if $\frac{\epsilon^4}{d\sigma^4} \leq C$, then $R(T) \geq \frac{1}{3}$, and hence if $\epsilon^* \leq C\sigma d^{\frac{1}{4}}$, then $R(T) \geq \frac{1}{3}$. This is precisely the lower bound we set out to show.

For more exploration of lower bound techniques like the one above, look at Ingster and Suslina [1].

24.4 Brief Recap

Here, we have explored why testing rates are better than estimation rates. Namely, with estimation we observed

$$\epsilon^* \asymp \sqrt{\frac{d}{n}} \tag{24.6}$$

whereas for testing, we observed

$$\epsilon^* \asymp \frac{d^{1/4}}{\sqrt{n}} \tag{24.7}$$

References

[1] Yu. I. Ingster and Irina A. Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. Springer, 2003.

[2] Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018.