Lecture 24: Robust Hypothesis testing between TV balls

Lecturer: Jiantao Jiao *Scribe: Nived Rajaraman*

We continue the discussion on robust hypothesis testing in this lecture. We discuss the work of [CGR16] based on ideas drawn from Yatracos [Yat85] to the setting of robust testing between TV-balls. The objective will be to construct a hypothesis test between two distributions, when the samples do not necessarily come from either distribution, but from a distribution lying within a TV-distance ε around either distribution. Yatracos' test for the finite sample setting involves a simple idea motivated by the robust hypothesis problem at the population level - construct an appropriate metric $\widetilde{\mathsf{TV}}$ which would serve to discriminate between the two hypotheses based on the empirical distribution.

1 Robust hypothesis testing

The setting of robust testing between TV balls is the following. Given two distributions P_0 and P_1 define $\mathcal{P}_0$ (and likewise $\mathcal{P}_1$) as $\{P : \mathsf{TV}(P, P_0) \leq \varepsilon\}$ - in other words, it is the TV-ball centered at P_0 of radius ε. Then, given samples $(X_1, X_2, \cdots, X_n) \overset{\text{i.i.d.}}{\sim} P$ where $P \in \mathcal{P}_0 \bigcup \mathcal{P}_1$, the robust hypothesis testing problem between TV-balls is to test between the hypotheses: $\mathcal{H}_0 : P \in \mathcal{P}_0$ and $\mathcal{H}_1 : P \in \mathcal{P}_1$. The problem is only meaningful if $P_0 \bigcap P_1 = \emptyset$, so we assume that $\mathsf{TV}(P_0, P_1) \geq 2\varepsilon$ which is a necessary and sufficient condition.
Recall the following definitions that were introduced in the previous lecture:

Definition 1. *Given a sample X, a test $\phi : \mathcal{X} \to [0, 1]$ returns the probability of rejecting the null hypothesis.*

Definition 2. *The level of a simple hypothesis test is defined as $\mathbb{E}_{X\sim P_0}[\phi(X)]$ and is the probability of incorrectly rejecting when the sample indeed comes from the null hypothesis. In particular, the level of a test is its type-I error.*

Definition 3. *The power of a test, β is defined as $\mathbb{E}_{X\sim P_1}[\phi(X)]$ and is the probability of correctly accepting the alternate when the sample indeed comes from the alternate hypothesis. In particular, the type-II error of the test is $1 - \beta$.*

Recall that in the more general setting considered by Huber [Hub64], the robust test constructed crucially relies on the knowledge of the parameter ε. In a typical setting, one however cannot assume that ε is known. This motivates the construction of a test ϕ that is agnostic to the knowledge of ε. Indeed this is the problem considered by [CGR16], based on ideas drawn from the work of Yatracos [Yat85] and will be the subject of the discussion to follow.

Let us begin with the simple case of the population setting to motivate the finite sample setting. Note that by population setting, we mean that we are given access to a distribution $P \in \mathcal{P}_0 \bigcup \mathcal{P}_1$, and the objective is to test between the hypotheses: $\mathcal{H}_0 : P \in \mathcal{P}_0$ and $\mathcal{H}_1 : P \in \mathcal{P}_1$. Observe that in this simple setting, the optimal test is to simply output the class $\mathcal{P}_0$ (respectively $\mathcal{P}_1$) if $\mathsf{TV}(P, P_0) < \mathsf{TV}(P, P_1)$. That is,

$$\phi = \mathbb{1}\{\mathsf{TV}(P, P_0) > \mathsf{TV}(P, P_1)\}. \tag{1}$$

A natural question to ask is - is the extension to the finite sample case $\phi = \mathbb{1}\{\mathsf{TV}(\hat{P}_n, P_0) > \mathsf{TV}(\hat{P}_n, P_1)\}$, where $\hat{P}_n$ is the empirical distribution of X_1^n the right thing to do? The answer to this question is a resounding no, since the total variation distance between any discrete distribution ($\hat{P}_n$) and any continuous distribution (P_0 and / or P_1 can be continuous) is 1.

The fact that total variation distance is too strong to control the distance between an empirical distribution and population distribution suggests that the right idea instead might be to relax it to some other distance $\widetilde{\mathsf{TV}}(\cdot, \cdot)$. To begin with, we require $\widetilde{\mathsf{TV}}$ to satisfy the following properties:

(a) $\widetilde{\mathsf{TV}}(\cdot,\cdot)$ should satisfy the triangle inequality,

(b) For any two distributions, $\widetilde{\mathsf{TV}}(p,q) \leq \mathsf{TV}(p,q)$

(c) For any $p,q \in \mathcal{G} \triangleq \{P_0, P_1\}$, $\widetilde{\mathsf{TV}}(p,q) = \mathsf{TV}(p,q)$.

Recall that $\mathsf{TV}(P_0,P_1) = \sup_A P_0(A) - P_1(A)$ where A is a measurable set. The set that optimizes this expression is easy to characterize explicitly (upto measure 0 equivalence), and is equal to $A_{\text{sup}} \triangleq \{x : p_0(x) > p_1(x)\}$ where p_0 and p_1 are the densities of P_0 and P_1. This is because of the fact that for any set A, $P_0(A) - P_1(A) = \int_A p_0(x) - p_1(x)\mathrm{d}x$ and hence the set supported on the set where the integrand is pointwise positive is optimal. This motivates defining $\widetilde{\mathsf{TV}}(p,q)$ as

$$\widetilde{\mathsf{TV}}(p,q) \triangleq |p(A_{\text{sup}}) - q(A_{\text{sup}})|$$

To avoid confusion, we reiterate that A_{sup} is the set that achieves the TV distance for P_0 and P_1 and not the particular p and q.
We verify the 3 posited properties that $\widetilde{\mathsf{TV}}(\cdot,\cdot)$ must satisfy.

(a) $\widetilde{\mathsf{TV}}(\cdot,\cdot)$ satisfies the triangle inequality. In particular,

$$\widetilde{\mathsf{TV}}(p,q) = |p(A_{\text{sup}} - q(A_{\text{sup}})| \leq |p(A_{\text{sup}} - r(A_{\text{sup}})| + |r(A_{\text{sup}} - q(A_{\text{sup}})| = \widetilde{\mathsf{TV}}(p,r) + \widetilde{\mathsf{TV}}(r,q)$$

(b) For any two distributions, $\widetilde{\mathsf{TV}}(p,q) \leq \mathsf{TV}(p,q)$. In particular,

$$\widetilde{\mathsf{TV}}(p,q) = |p(A_{\text{sup}} - q(A_{\text{sup}})| \leq \sup_A |p(A - q(A)| = \mathsf{TV}(p,q)$$

(c) For any $p,q \in \mathcal{G} \triangleq \{P_0, P_1\}$, $\widetilde{\mathsf{TV}}(p,q) = \mathsf{TV}(p,q)$. For $p = q$, $\widetilde{\mathsf{TV}}(p,q) = \mathsf{TV}(p,q) = 0$. For $p = P_0$ and $q = P_1$, by definition, $\widetilde{\mathsf{TV}}(p,q) = \mathsf{TV}(p,q)$.

Having established these properties, we return to the analysis of the proposed hypothesis test, $\phi = \mathbb{1}\{\widetilde{\mathsf{TV}}(\hat{P}_n, P_0) > \widetilde{\mathsf{TV}}(\hat{P}_n, P_1)\}$. We show the following result,

Theorem 4. *Both type-I (α) and type-II (β) errors of the test ϕ are upper bounded by $e^{-\frac{n}{2}(\mathsf{TV}(P_0,P_1)-2\varepsilon)^2}$.*

Proof We prove the result for α with the understanding that the result for β proceeds analogously. We abbreviate probabilities under $\mathcal{H}_0 : X_1^n \sim P \in P_0$ as P_0. Using properties (a) - (c) of $\widetilde{\mathsf{TV}}(\cdot,\cdot)$, we have the following

$$\begin{aligned}
\alpha &= P_0(\widetilde{\mathsf{TV}}(\hat{P}_n, P_0) > \widetilde{\mathsf{TV}}(\hat{P}_n, P_1)) \\
&\stackrel{(a)}{\leq} P_0(\widetilde{\mathsf{TV}}(\hat{P}_n, P_0) > \widetilde{\mathsf{TV}}(P_0, P_1) - \widetilde{\mathsf{TV}}(\hat{P}_n, P_0)) \\
&\stackrel{(c)}{=} P_0(2\widetilde{\mathsf{TV}}(\hat{P}_n, P_0) > \mathsf{TV}(P_0, P_1)) \\
&\stackrel{(a)}{\leq} P_0(2\widetilde{\mathsf{TV}}(\hat{P}_n, P) > \mathsf{TV}(P_0, P_1) - 2\widetilde{\mathsf{TV}}(P_0, P)) \\
&\stackrel{(b)}{\leq} P_0(2\widetilde{\mathsf{TV}}(\hat{P}_n, P) > \mathsf{TV}(P_0, P_1) - 2\mathsf{TV}(P_0, P)) \\
&\leq P_0(2\widetilde{\mathsf{TV}}(\hat{P}_n, P) > \mathsf{TV}(P_0, P_1) - 2\varepsilon)
\end{aligned}$$

Now, observe that $\widetilde{\mathsf{TV}}(\hat{P}_n, P) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i \in A_{\text{sup}}\} - P(X_i \in A_{\text{sup}})$. Observe that each term $\mathbb{1}\{X_i \in A_{\text{sup}}\} - P(X_i \in A_{\text{sup}})$ is an i.i.d. mean 0 random variable lying in an interval of width 1. Therefore we can use Hoeffding's inequality to get,

$$\alpha \leq e^{-\frac{n}{2}(\mathsf{TV}(P_0,P_1)-2\varepsilon)^2}$$

□

References

[CGR16] Mengjie Chen, Chao Gao, and Zhao Ren. A general decision theory for huber's ϵ-contamination model. *Electron. J. Statist.*, 10(2):3752–3774, 2016.

[Hub64] Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

[Yat85] Yannis G. Yatracos. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *Ann. Statist.*, 13(2):768–774, 06 1985.