[Lecture 5]

0.1 Finite-Sample Analysis

Now that we have developed tools for analyzing statistical concentration, we will use these to analyze the finite-sample behavior of robust estimators. Recall that we previously studied the minimum distance functional defined as

$$\hat{\theta}(\tilde{p}) = \theta^*(q), \text{ where } q = \arg\min_{q \in \mathcal{G}} \mathsf{TV}(q, \tilde{p}). \tag{1}$$

This projects onto the set $\mathcal{G}$ under TV distance and outputs the optimal parameters for the projected distribution.

The problem with the minimum distance functional defined above is that projection under TV usually doesn't make sense for finite samples! For instance, suppose that p is a Gaussian distribution and let p_n and p'_n be the empirical distributions of two different sets of n samples. Then $\mathsf{TV}(p_n, p) = \mathsf{TV}(p_n, p'_n) = 1$ almost surely. This is because samples from a continuous probability distribution will almost surely be distinct, and TV distance doesn't give credit for being "close"—the TV distance between two point masses at 1 and 1.000001 is still 1.[1]

To address this issue, we will consider two solutions. The first solution is to *relax the distance*. Intuitively, the issue is that the TV distance is too strong—it reports a large distance even between a population distribution p and the finite-sample distribution p_n. We will replace the distance TV with a more forgiving distance $\widetilde{\mathsf{TV}}$ and use the minimum distance functional corresponding to this relaxed distance. To show that projection under $\widetilde{\mathsf{TV}}$ still works, we will need to check that the modulus $\mathfrak{m}(\mathcal{G}, \epsilon)$ is still small after we replace TV with $\widetilde{\mathsf{TV}}$, and we will also need to check that the distance $\widetilde{\mathsf{TV}}(p, p_n)$ between p and its empirical distribution is small with high probability. We do this below in Section 0.1.1.

An alternative to relaxing the distance from TV to $\widetilde{\mathsf{TV}}$ is to expand the destination set from $\mathcal{G}$ to some $\mathcal{M} \supset \mathcal{G}$, such that even though p is not close to the empirical distribution p_n, *some* element of $\mathcal{M}$ is close to p_n. Another advantage to expanding the destination set is that projecting onto $\mathcal{G}$ may not be computationally tractable, while projecting onto some larger set $\mathcal{M}$ can sometimes be done efficiently. We will see how to statistically analyze this modified projection algorithm in Section **??**, and study the computational feasibility of projecting onto a set $\mathcal{M}$ starting in Section **??**.

0.1.1 Relaxing the Distance

Here we instantiate the first solution of replacing TV with some $\widetilde{\mathsf{TV}}$ for the projection algorithm. The following lemma shows that properties we need $\widetilde{\mathsf{TV}}$ to satisfy:

Lemma 0.1. *Suppose that $\widetilde{\mathsf{TV}}$ is a (pseudo-)metric such that $\widetilde{\mathsf{TV}}(p, q) \le \mathsf{TV}(p, q)$ for all p, q. If we assume that $p^* \in \mathcal{G}$ and $\mathsf{TV}(p^*, \tilde{p}) \le \epsilon$, then the error of the minimum distance functional (**??**) with $D = \widetilde{\mathsf{TV}}$ is at most $\mathfrak{m}(\mathcal{G}, 2\epsilon', \widetilde{\mathsf{TV}}, L)$, where $\epsilon' = \epsilon + \widetilde{\mathsf{TV}}(\tilde{p}, \tilde{p}_n)$.*

Proof. By Proposition **??** we already know that the error is bounded by $\mathfrak{m}(\mathcal{G}, 2\widetilde{\mathsf{TV}}(p^*, \tilde{p}_n), \widetilde{\mathsf{TV}}, L)$. Since $\widetilde{\mathsf{TV}}$ is a pseudometric, by the triangle inequality we have $\widetilde{\mathsf{TV}}(p^*, \tilde{p}_n) \le \widetilde{\mathsf{TV}}(p^*, \tilde{p}) + \widetilde{\mathsf{TV}}(\tilde{p}, \tilde{p}_n)$. Finally, $\widetilde{\mathsf{TV}}(p^*, \tilde{p}) \le \mathsf{TV}(p^*, \tilde{p})$ by assumption. □

Lemma 0.1 shows that we need $\widetilde{\mathsf{TV}}$ to satisfy two properties: $\widetilde{\mathsf{TV}}(\tilde{p}, \tilde{p}_n)$ should be small, and the modulus $\mathfrak{m}(\mathcal{G}, \epsilon, \widetilde{\mathsf{TV}})$ should not be too much larger than $\mathfrak{m}(\mathcal{G}, \epsilon, \mathsf{TV})$.

For mean estimation (where recall $L(p, \theta) = \|\theta - \mu(p)\|_2$), we will use the following $\widetilde{\mathsf{TV}}$:

$$\widetilde{\mathsf{TV}}_{\mathcal{H}}(p, q) \stackrel{\text{def}}{=} \sup_{f \in \mathcal{H}, \tau \in \mathbb{R}} |\mathbb{P}_{X \sim p}[f(X) \ge \tau] - \mathbb{P}_{X \sim q}[f(X) \ge \tau]|. \tag{2}$$

(Note the similarity to the distance in Proposition **??**; we will make use of this later.) We will make the particular choice $\mathcal{H} = \mathcal{H}_{\mathsf{lin}}$, where $\mathcal{H}_{\mathsf{lin}} \stackrel{\text{def}}{=} \{x \mapsto \langle v, x \rangle \mid v \in \mathbb{R}^d\}$.

[1] We will later study the W_1 distance, which does give credit for being close.

First note that $\widetilde{\mathsf{TV}}_{\mathcal{H}}$ is indeed upper-bounded by TV, since $\mathsf{TV}(p,q) = \sup_E |p(E) - q(E)|$ is the supremum over all events E, and (2) takes a supremum over a subset of events. The intuition for taking the particular family $\mathcal{H}$ is that linear projections of our data contain all information needed to recover the mean, so perhaps it is enough for distributions to be close only under these projections.

Bounding the modulus. To formalize this intuition, we prove the following *mean crossing lemma*:

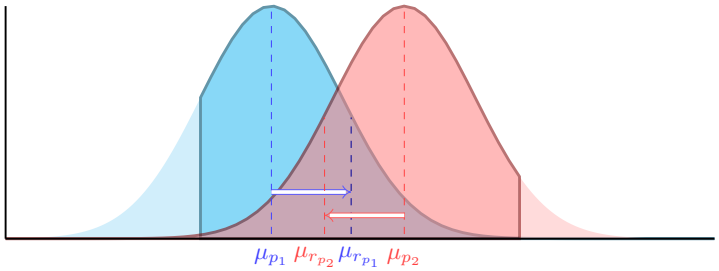

Figure 1: Illustration of mean cross lemma. For any distributions p_1, p_2 that are close under $\widetilde{\mathsf{TV}}$, we can truncate the ϵ-tails of each distribution to make their means cross.

Lemma 0.2. *Suppose that p and q are two distributions such that $\widetilde{\mathsf{TV}}_{\mathcal{H}}(p,q) \le \epsilon$. Then for any $f \in \mathcal{H}$, there are distributions $r_p \le \frac{p}{1-\epsilon}$ and $r_q \le \frac{q}{1-\epsilon}$ such that $\mathbb{E}_{X\sim r_p}[f(X)] \ge \mathbb{E}_{Y\sim r_q}[f(Y)]$.*

Proof. We will prove the stronger statement that $f(X)$ under r_p *stochastically dominates* $f(Y)$ under r_q. Starting from p, q, we delete ϵ probability mass corresponding to the smallest points of $f(X)$ in p to get r_p, and delete ϵ probability mass corresponding to the largest points $f(Y)$ in q to get r_q. Since $\widetilde{\mathsf{TV}}_{\mathcal{H}}(p,q) \le \epsilon$ we have

$$\sup_{\tau\in\mathbb{R}} |\mathbb{P}_{X\sim p}(f(X) \ge \tau) - \mathbb{P}_{Y\sim q}(f(Y) \ge \tau)| \le \epsilon, \tag{3}$$

which implies that $\mathbb{P}_{r_p}(f(X) \ge \tau) \ge \mathbb{P}_{r_q}(f(Y) \ge \tau)$ for all $t \in \mathbb{R}$. Hence, r_p stochastically dominates r_q and $\mathbb{E}_{r_p}[f(X)] \ge \mathbb{E}_{r_q}[(Y)]$. □

Mean crossing lemmas such as Lemma 0.2 help us bound the modulus of relaxed distances for the family of resilient distributions. In this case we have the following corollary:

Corollary 0.3. *For the family $\mathcal{G}_{\mathsf{TV}}(\rho,\epsilon)$ of (ρ,ϵ)-resilient distributions and $L(p,\theta) = \|\theta - \mu(p)\|_2$, we have*

$$\mathfrak{m}(\mathcal{G}_{\mathsf{TV}}(\rho,\epsilon), \epsilon, \widetilde{\mathsf{TV}}_{\mathcal{H}_{\mathsf{lin}}}) \le 2\rho. \tag{4}$$

Compare to Theorem **??** where we showed that $\mathfrak{m}(\mathcal{G}_{\mathsf{TV}}, \epsilon, \mathsf{TV}) \le \rho$. Thus as long as Theorem **??** is tight, relaxing from TV to $\widetilde{\mathsf{TV}}_{\mathcal{H}_{\mathsf{lin}}}$ doesn't increase the modulus at all!

Proof of Corollary 0.3. Let $p, q \in \mathcal{G}_{\mathsf{TV}}$ such that $\widetilde{\mathsf{TV}}(p,q) \le \epsilon$. Take $v = \arg\max_{\|v\|_2=1} v^\top(\mathbb{E}_p[X] - \mathbb{E}_q[X])$, hence $\mathbb{E}_p[v^\top X] - \mathbb{E}_q[v^\top X] = \|\mathbb{E}_p[X] - \mathbb{E}_q[X]\|_2$. It follows from Lemma 0.2 that there exist $r_p \le \frac{p}{1-\epsilon}, r_q \le \frac{q}{1-\epsilon}$ such that

$$\mathbb{E}_{r_p}[v^\top X] \le \mathbb{E}_{r_q}[v^\top X]. \tag{5}$$

Furthermore, from $p, q \in \mathcal{G}_{\mathsf{TV}}(\rho,\epsilon)$, we have

$$\mathbb{E}_p[v^\top X] - \mathbb{E}_{r_p}[v^\top X] \le \rho, \tag{6}$$
$$\mathbb{E}_{r_q}[v^\top X] - \mathbb{E}_q[v^\top X] \le \rho. \tag{7}$$

Then,

$$\|\mathbb{E}_p[X] - \mathbb{E}_q[X]\|_2 = \mathbb{E}_p[v^\top X] - \mathbb{E}_q[v^\top X] \tag{8}$$

$$= \underbrace{\mathbb{E}_p[v^\top X] - \mathbb{E}_{r_p}[v^\top X]}_{\leq \rho} + \underbrace{\mathbb{E}_{r_p}[v^\top X] - \mathbb{E}_{r_q}[v^\top X]}_{\leq 0} + \underbrace{\mathbb{E}_{r_q}[v^\top X] - \mathbb{E}_q[v^\top X]}_{\leq \rho} \tag{9}$$

$$\leq 2\rho, \tag{10}$$

which shows the modulus is small as claimed. □

Bounding the distance to the empirical distribution. Now that we have bounded the modulus, it remains to bound the distance $\widetilde{\mathsf{TV}}(\tilde{p}, \tilde{p}_n)$. Note that $\widetilde{\mathsf{TV}}(\tilde{p}, \tilde{p}_n)$ is exactly the quantity bounded in equation (**??**) of Proposition **??**; we thus have that $\widetilde{\mathsf{TV}}_{\mathcal{H}}(\tilde{p}, \tilde{p}_n) \leq \mathcal{O}\big(\sqrt{\frac{\mathsf{vc}(\mathcal{H}) + \log(1/\delta)}{n}}\big)$ with probability $1 - \delta$. Here $\mathsf{vc}(\mathcal{H})$ is the VC dimension of the family of threshold functions $\{x \mapsto \mathbb{I}[f(x) \geq \tau] \mid f \in \mathcal{H}, \tau \in \mathbb{R}\}$. So, for $\mathcal{H} = \mathcal{H}_{\mathsf{lin}}$ all we need to do is bound the VC dimension of the family of halfspace functions on $\mathbb{R}^d$.

We claimed earlier that this VC dimension is $d + 1$, but we prove it here for completeness. We will show that no set of points $x_1, \ldots, x_{d+2} \in \mathbb{R}^d$ cannot be shattered into all 2^{d+2} possible subsets using halfspaces. For any such points we can find multipliers $a_1, \ldots, a_{d+2} \in \mathbb{R}$ such that

$$\sum_{i=1}^{d+2} a_i x_i = 0, \sum_{i=1}^{d+2} a_i = 0. \tag{11}$$

Let $S_+ = \{i \mid a_i > 0\}$ and $S_- = \{i \mid a_i < 0\}$. We will show that the convex hulls of S_+ and S_- intersect. Consequently, there is no vector v and threshold τ such that $\langle x_i, v\rangle \geq \tau$ iff $i \in S_+$. (This is because both a halfspace and its complement are convex, so if we let $H_{v,\tau}$ denote the half-space, it is impossible to have $S_+ \subset H_{v,\tau}$, $S_- \subset H_{v,\tau}^c$, and $\mathrm{conv}(S_+) \cap \mathrm{conv}(S_-) \neq \emptyset$.)

To prove that the convex hulls intersect, note that we have

$$\frac{1}{A}\sum_{i \in S_+} a_i x_i = \frac{1}{A}\sum_{i \in S_-} (-a_i) x_i, \tag{12}$$

where $A = \sum_{i \in S_+} a_i = \sum_{i \in S_-} (-a_i)$. But the left-hand-side lies in $\mathrm{conv}(S_+)$ while the right-hand-side lies in $\mathrm{conv}(S_-)$, so the convex hulls do indeed intersect.

This shows that $x_1, \ldots, x_{d+2}$ cannot be shattered, so $\mathsf{vc}(\mathcal{H}_{\mathsf{lin}}) \leq d+1$. Combining this with Proposition **??**, we obtain:

Proposition 0.4. *With probability $1 - \delta$, we have $\widetilde{\mathsf{TV}}_{\mathcal{H}_{\mathsf{lin}}}(\tilde{p}, \tilde{p}_n) \leq \mathcal{O}\big(\sqrt{\frac{d + \log(1/\delta)}{n}}\big)$.*

Combining this with Corollary 0.3 and Lemma 0.1, we see that projecting onto $\mathcal{G}_{\mathsf{TV}}(\rho, 2\epsilon')$ under $\widetilde{\mathsf{TV}}_{\mathcal{H}_{\mathsf{lin}}}$ performs well in finite samples, for $\epsilon' = \epsilon + \mathcal{O}(\sqrt{d/n})$. For instance, if $\mathcal{G}$ has bounded covariance we achieve error $\mathcal{O}(\sqrt{\epsilon + \sqrt{d/n}})$; if $\mathcal{G}$ is sub-Gaussian we achieve error $\tilde{\mathcal{O}}(\epsilon + \sqrt{d/n})$; and in general if $\mathcal{G}$ has bounded ψ-norm we achieve error $\mathcal{O}\big((\epsilon + \sqrt{d/n})\psi^{-1}\big(\frac{1}{\epsilon + \sqrt{d/n}}\big)\big) \leq \mathcal{O}((\epsilon + \sqrt{d/n})\psi^{-1}(1/\epsilon))$.

This analysis is slightly sub-optimal as the best lower bound we are aware of is $\Omega(\epsilon\psi^{-1}(1/\epsilon) + \sqrt{d/n})$, i.e. the $\psi^{-1}(1/\epsilon)$ coefficient in the dependence on n shouldn't be there. However, it is accurate as long as ϵ is large compared to $\sqrt{d/n}$.

Connection to Tukey median. A classical robust estimator for the mean is the *Tukey median*, which solves the problem

$$\min_{\mu} \max_{v \in \mathbb{R}^d} |\mathbb{P}_{X \sim \tilde{p}_n}[\langle X, v\rangle \geq \langle \mu, v\rangle] - \tfrac{1}{2}| \tag{13}$$

[Note: this definition is slightly wrong as it does not behave gracefully when there is a point mass at μ.]

It is instructive to compare this to projection under $\widetilde{\mathsf{TV}}$, which corresponds to

$$\min_{q \in \mathcal{G}} \max_{v \in \mathbb{R}^d, \tau \in \mathbb{R}} |\mathbb{P}_{X \sim \tilde{p}_n}[\langle X, v \rangle \geq \tau] - \mathbb{P}_{X \sim q}[\langle X, v \rangle \geq \tau]|. \tag{14}$$

The differences are: (1) the Tukey median only minimizes over the mean rather than the full distribution q; (2) it only considers the threshold $\langle \mu, v \rangle$ rather than all thresholds τ; it assumes that the median of any one-dimensional projection $\langle X, v \rangle$ is equal to its mean (which is why we subtract $\frac{1}{2}$ in (13)). Distributions satisfying this final property are said to be *unskewed*.

For unskewed distributions with "sufficient probability mass" near the mean, the Tukey median yields a robust estimator. In fact, it can be robust even if the true distribution has heavy tails (and hence is not resilient), by virtue of leveraging the unskewed property. We will explore this in an exercise.