[Lecture 2]

0.1 Minimum Distance Functionals

In the previous section we saw that simple approaches to handling outliers in high-dimensional data, such as the trimmed mean, incur a $\sqrt{d}$ error. We will avoid this error using *minimum distance functionals*, an idea which seems to have first appeared in **?**.

Definition 0.1 (Minimum distance functional)**.** For a family $\mathcal{G}$ and discrepancy D, the minimum distance functional is

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

In other words, $\hat{\theta}$ is the parameters obtained by first projecting $\tilde{p}$ onto $\mathcal{G}$ under D, and then outputting the optimal parameters for the resulting distribution.

An attractive property of the minimum-distance functional is that it does not depend on the perturbation level ϵ. More importantly, it satisfies the following cost bound in terms of the *modulus of continuity* of $\mathcal{G}$:

Proposition 0.2. *Suppose D is a pseudometric. Then the cost $L(p^*, \hat{\theta}(\tilde{p}))$ of the minimum distance functional is at most the maximum loss between any pair of distributions in $\mathcal{G}$ of distance at most 2ϵ:*

$$\mathfrak{m}(\mathcal{G}, 2\epsilon, D, L) \triangleq \sup_{p,q \in \mathcal{G} : D(p,q) \le 2\epsilon} L(p, \theta^*(q)). \tag{2}$$

The quantity $\mathfrak{m}$ is called the modulus of continuity because, if we think of $L(p, \theta^*(q))$ as a discrepancy between distributions, then $\mathfrak{m}$ is the constant of continuity between L and D when restricted to pairs of nearby distributions in $\mathcal{G}$.

Specialize again to the case $D = \mathsf{TV}$ and $L(p^*, \theta) = \|\theta - \mu(p^*)\|_2$ (here we allow p^* to be a distribution over $\mathbb{R}^d$ rather than just $\mathbb{R}$). Then the modulus is $\sup_{p,q \in \mathcal{G} : \mathsf{TV}(p,q) \le 2\epsilon} \|\mu(p) - \mu(q)\|_2$. As a concrete example, let $\mathcal{G}$ be the family of Gaussian distributions with unknown mean μ and identity covariance. For this family, the TV distance is essentially linear in the difference in mean:

Lemma 0.3. *Let $\mathcal{N}(\mu, I)$ denote a Gaussian distribution with mean μ and identity covariance. Then*

$$\min(u/2, 1)/\sqrt{2\pi} \le \mathsf{TV}(\mathcal{N}(\mu, I), \mathcal{N}(\mu', I)) \le \min(u/\sqrt{2\pi}, 1), \tag{3}$$

where $u = \|\mu - \mu'\|_2$.

Proof. By rotational and translational symmetry, it suffices to consider the case of one-dimensional Gaussians $\mathcal{N}(-u/2, 1)$ and $\mathcal{N}(u/2, 1)$. Then we have that

$$\mathsf{TV}(\mathcal{N}(-u/2, 1), \mathcal{N}(u/2, 1)) = \frac{1}{2\sqrt{2\pi}} \int_{-\infty}^{\infty} |e^{-(t+u/2)^2/2} - e^{-(t-u/2)^2/2}| dt \tag{4}$$

$$\stackrel{(i)}{=} \frac{1}{\sqrt{2\pi}} \int_{-u/2}^{u/2} e^{-t^2/2} dt. \tag{5}$$

(The equality (i) is a couple lines of algebra, but is easiest to see by drawing a graph of the two Gaussians and cancelling out most of the probability mass.)

For the lower bound, note that $e^{-t^2/2} \ge \frac{1}{2}$ if $|t| \le 1$.

For the upper bound, similarly note that $e^{-t^2/2} \le 1$ for all $t \in \mathbb{R}$, and also that the entire integral must be at most 1 because it is the probability density of a Gaussian. $\square$

Lemma 0.3 allows us to compute the modulus for Gaussians:

Corollary 0.4. *Let $\mathcal{G}_{\mathsf{gauss}}$ be the family of isotropic Gaussians, $D = \mathsf{TV}$, and L the difference in means as above. Then $\mathfrak{m}(\mathcal{G}_{\mathsf{gauss}}, \epsilon, D, L) \le 2\sqrt{2\pi}\epsilon$ whenever $\epsilon \le \frac{1}{2\sqrt{2\pi}}$.*

In particular, by Proposition 0.2 the minimum distance functional achieves error $\mathcal{O}(\epsilon)$ for Gaussian distributions when $\epsilon \le \frac{1}{2\sqrt{2\pi}}$. This improves substantially on the $\epsilon\sqrt{d}$ error of the trimmed mean estimator from Section **??**. We have achieved our goal at least for Gaussians.

More general families. Taking $\mathcal{G}$ to be Gaussians is restrictive, as it assumes that p^* has a specific parametric form—counter to our goal of being robust! However, the modulus $\mathfrak{m}$ is bounded for much more general families. As one example, we can take the distributions with bounded covariance (compare to Proposition **??**):

Lemma 0.5. *Let $\mathcal{G}_{\mathsf{cov}}(\sigma)$ be the family of distributions whose covariance matrix Σ satisfies $\Sigma \preceq \sigma^2 I$. Then $\mathfrak{m}(\mathcal{G}_{\mathsf{cov}}(\sigma), \epsilon) = \mathcal{O}(\sigma\sqrt{\epsilon})$.*

Proof. Let $p, q \in \mathcal{G}_{\mathsf{cov}}(\sigma)$ such that $\mathsf{TV}(p, q) \leq \epsilon$. This means that we can get from p to q by first deleting ϵ mass from p and then adding ϵ new points to end up at q. Put another way, there is a distribution r that can be reached from both p and q by deleting ϵ mass (and then renormalizing). In fact, this distribution is exactly

$$r = \frac{\min(p, q)}{1 - \mathsf{TV}(p, q)}. \tag{6}$$

Since r can be obtained from both p and q by deletions, we can make use of the following multi-dimensional analogue of Chebyshev's inequality (Lemma **??**):

Lemma 0.6 (Chebyshev in $\mathbb{R}^d$)**.** *Suppose that p has mean μ and covariance Σ, where $\Sigma \preceq \sigma^2 I$. Then, if E is any event with probability at least δ, we have $\|\mathbb{E}_{X\sim p}[X \mid E] - \mu\|_2 \leq \sigma\sqrt{\frac{2(1-\delta)}{\delta}}$.*

As a consequence, we have $\|\mu(r) - \mu(p)\|_2 \leq \sigma\sqrt{2\epsilon/(1-\epsilon)}$ and $\|\mu(r) - \mu(q)\|_2 \leq \sigma\sqrt{2\epsilon/(1-\epsilon)}$ (since r can be obtained from either p or q by conditioning on an event of probability $1 - \epsilon$). By triangle inequality and assuming $\epsilon \leq \frac{1}{2}$, we have $\|\mu(p) - \mu(q)\|_2 \leq 4\sigma\sqrt{\epsilon}$, as claimed. □

As a consequence, the minimum distance functional robustly estimates the mean bounded covariance distributions with error $\mathcal{O}(\sigma\sqrt{\epsilon})$, generalizing the 1-dimensional bound obtained by the trimmed mean.

In Lemma 0.5, the two key properties we needed were:

- The *midpoint property* of TV distance (i.e., that there existed an r that was a deletion of p and q).
- The *bounded tails* guaranteed by Chebyshev's inequality.

If we replace bounded covariance distributions with any other family that has tails bounded in a similar way, then the minimum distance functional will similarly yield good bounds. A general family of distributions satisfying this property are *resilience distributions*, which we turn to next.

0.2 Resilience

Here we generalize Lemma 0.5 to prove that the modulus of continuity $\mathfrak{m}$ is bounded for a general family of distributions containins Gaussians, sub-Gaussians, bounded covariance distributions, and many others. The main observation is that in the proof of Lemma 0.5, all we needed was that the tails of distributions in $\mathcal{G}$ were bounded, in the sense that deleting an ϵ-fraction of the points could not substantially change the mean. This motivates the following definition:

Definition 0.7. A distribution p over $\mathbb{R}^d$ is said to be (ρ, ϵ)-resilient (with respect to some norm $\|\cdot\|$) if

$$\|\mathbb{E}_{X\sim p}[X \mid E] - \mathbb{E}_{X\sim p}[X]\| \leq \rho \text{ for all events } E \text{ with } p(E) \geq 1 - \epsilon. \tag{7}$$

We let $\mathcal{G}_{\mathsf{TV}}(\rho, \epsilon)$ denote the family of (ρ, ϵ)-resilient distributions.

We observe that $\mathcal{G}_{\mathsf{cov}}(\sigma) \subset \mathcal{G}_{\mathsf{TV}}(\sigma\sqrt{2\epsilon/(1-\epsilon)}, \epsilon)$ for all ϵ by Lemma 0.6; in other words, bounded covariance distributions are resilient. We can also show that $\mathcal{G}_{\mathsf{gauss}} \subset \mathcal{G}_{\mathsf{TV}}(2\epsilon\sqrt{\log(1/\epsilon)}, \epsilon)$, so Gaussians are resilient as well.

Resilient distributions always have bounded modulus:

Theorem 0.8. *The modulus of continuity $\mathfrak{m}(\mathcal{G}_{\mathsf{TV}}, 2\epsilon)$ satisfies the bound*

$$\mathfrak{m}(\mathcal{G}_{\mathsf{TV}}(\rho, \epsilon), 2\epsilon) \leq 2\rho \tag{8}$$

whenever $\epsilon < 1/2$.

Proof. As in Lemma 0.5, the key idea is that any two distributions p, q that are close in TV have a *midpoint* distribution $r = \frac{\min(p,q)}{1-\mathsf{TV}(p,q)}$ that is a deletion of both distributions). This midpoint distribution connects the two distributions, and it follows from the triangle inequality that the modulus of $\mathcal{G}_{\mathsf{TV}}$. is bounded. We illustrate this idea in Figure 1 and make it precise below.

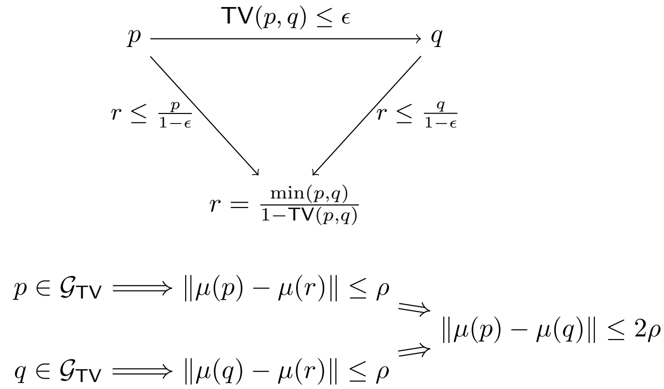

Figure 1: Midpoint distribution r helps bound the modulus for $\mathcal{G}_{\mathsf{TV}}$.

Recall that

$$\mathfrak{m}(\mathcal{G}_{\mathsf{TV}}(\rho,\epsilon), 2\epsilon) = \sup_{p,q \in \mathcal{G}_{\mathsf{TV}}(\rho,\epsilon): \mathsf{TV}(p,q) \le 2\epsilon} \|\mu(p) - \mu(q)\|. \tag{9}$$

From $\mathsf{TV}(p,q) \le 2\epsilon$, we know that $r = \frac{\min(p,q)}{1-\mathsf{TV}(p,q)}$ can be obtained from either p and q by conditioning on an event of probability $1-\epsilon$. It then follows from $p, q \in \mathcal{G}_{\mathsf{TV}}(\rho,\epsilon)$ that $\|\mu(p) - \mu(r)\| \le \epsilon$ and similarly $\|\mu(q) - \mu(r)\| \le \epsilon$. Thus by the triangle inequality $\|\mu(p) - \mu(q)\| \le 2\rho$, which yields the desired result. □

We have seen so far that resilient distributions have bounded modulus, and that both Gaussian and bounded covariance distributions are resilient. The bound on the modulus for $\mathcal{G}_{\mathsf{cov}}$ that is implied by resilience is optimal ($\mathcal{O}(\sigma\sqrt{\epsilon})$), while for $\mathcal{G}_{\mathsf{gauss}}$ it is optimal up to log factors ($\mathcal{O}(\epsilon\sqrt{\log(1/\epsilon)})$ vs. $\mathcal{O}(\epsilon)$). In fact, Gaussians are a special case and resilience yields an essentially optimal bound at least for most non-parametric families of distributions. As one family of examples, consider distributions with bounded *Orlicz norm*:

Definition 0.9 (Orlicz norm)**.** A function $\psi : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ is an *Orlicz function* if ψ is convex, non-decreasing, and satisfies $\psi(0) = 0$, $\psi(x) \to \infty$ as $x \to \infty$. For an Orlicz function ψ, the Orlicz norm or ψ-norm of a random variable X is defined as

$$\|X\|_\psi \triangleq \inf\left\{t > 0 : \mathbb{E}_p\left[\psi\left(\frac{|X|}{t}\right)\right] \le 1\right\}. \tag{10}$$

We let $\mathcal{G}_\psi(\sigma)$ denote the family of distributions with $\|X - \mathbb{E}[X]\|_\psi \le \sigma$.

As special cases, we say that a random variable $X \sim p$ is *sub-Gaussian* with parameter σ if $\|\langle X - \mathbb{E}_p[X], v\rangle\|_{\psi_2} \le \sigma$ whenever $\|v\|_2 \le 1$, where $\psi_2(x) = e^{x^2} - 1$. We define a *sub-exponential* random variable similarly for the function $\psi_1(x) = e^x - 1$.

Definition 0.9 applies to distributions on $\mathbb{R}$, but we can generalize this to distributions on $\mathbb{R}^d$ by taking one-dimensional projections:

Definition 0.10 (Orlicz norm in $\mathbb{R}^d$)**.** For a random variable $X \in \mathbb{R}^d$ and Orlicz function ψ, we define the d-dimensional ψ-norm as

$$\|X\|_\psi \triangleq \inf\{t > 0 : \|\langle X, v\rangle\|_\psi \le t \text{ whenever } \|v\|_2 \le 1\}. \tag{11}$$

We let $\mathcal{G}_\psi(\sigma)$ denote the distributions with bounded ψ-norm as in Definition 0.9.

Thus a distribution has bounded ψ-norm if each of its 1-dimensional projections does. As an example, $\mathcal{G}_{\mathsf{cov}}(\sigma) = \mathcal{G}_{\psi}(\sigma)$ for $\psi(x) = x^2$, so Orlicz norms generalize bounded covariance. It is also possible to generalize Definition 0.10 to norms other than the ℓ_2-norm, which we will see in an exercise.

Functions with bounded Orlicz norm are resilient:

Lemma 0.11. *The family $\mathcal{G}_{\psi}(\sigma)$ is contained in $\mathcal{G}_{\mathsf{TV}}(2\sigma\epsilon\psi^{-1}(1/\epsilon), \epsilon)$ for all $0 < \epsilon < 1/2$.*

Proof. Without loss of generality assume $\mathbb{E}[X] = 0$. For any event E with $p(E) = 1 - \epsilon' \geq 1 - \epsilon$, denote its complement as E^c. We then have

$$\|\mathbb{E}_{X\sim p}[X \mid E]\|_2 \overset{(i)}{=} \frac{\epsilon'}{1-\epsilon'}\|\mathbb{E}_{X\sim p}[X \mid E^c]\|_2 \tag{12}$$

$$= \frac{\epsilon'}{1-\epsilon'} \sup_{\|v\|_2\leq 1} \mathbb{E}_{X\sim p}[\langle X, v\rangle \mid E^c] \tag{13}$$

$$\overset{(ii)}{\leq} \frac{\epsilon'}{1-\epsilon'} \sup_{\|v\|_2\leq 1} \sigma\psi^{-1}(\mathbb{E}_{X\sim p}[\psi(|\langle X, v\rangle|/\sigma) \mid E^c]) \tag{14}$$

$$\overset{(iii)}{\leq} \frac{\epsilon'}{1-\epsilon'} \sup_{\|v\|_2\leq 1} \sigma\psi^{-1}(\mathbb{E}_{X\sim p}[\psi(|\langle X, v\rangle|/\sigma)]/\epsilon') \tag{15}$$

$$\overset{(iv)}{\leq} \frac{\epsilon'}{1-\epsilon'}\sigma\psi^{-1}(1/\epsilon') \leq 2\epsilon\sigma\psi^{-1}(1/\epsilon), \tag{16}$$

as was to be shown. Here (i) is because $(1-\epsilon')\mathbb{E}[X \mid E] + \epsilon'\mathbb{E}[X \mid E^c] = 0$. Meanwhile (ii) is by convexity of ψ, (iii) is by non-negativity of ψ, and (iv) is the assumed ψ-norm bound. ☐

As a consequence, the modulus $\mathfrak{m}$ of $\mathcal{G}_{\psi}(\sigma)$ is $\mathcal{O}(\sigma\epsilon\psi^{-1}(1/\epsilon))$, and hence the minimum distance functional estimates the mean with error $\mathcal{O}(\sigma\epsilon\psi^{-1}(1/\epsilon))$. Note that for $\psi(x) = x^2$ this reproduces our result for bounded covariance. For $\psi(x) = x^k$ we get error $\mathcal{O}(\sigma\epsilon^{1-1/k})$ when a distribution has kth moments bounded by σ^k. Similarly for sub-Gaussian distributions we get error $\mathcal{O}(\sigma\epsilon\sqrt{\log(1/\epsilon)})$. We will show in an exercise that the error bound implied by Lemma 0.11 is optimal for any Orlicz function ψ.

Further properties and dual norm perspective. Having seen several examples of resilient distributions, we now collect some basic properties of resilience, as well as a dual perspective that is often fruitful. First, we can make the connection between resilience and tails even more precise with the following lemma:

Lemma 0.12. *For a fixed vector v, let $\tau_\epsilon(v)$ denote the ϵ-quantile of $\langle x - \mu, v\rangle$: $\mathbb{P}_{x\sim p}[\langle x - \mu, v\rangle \geq \tau_\epsilon(v)] = \epsilon$. Then, p is (ρ, ϵ)-resilient in a norm $\|\cdot\|$ if and only if the ϵ-tail of p has bounded mean when projected onto any dual unit vector v:*

$$\mathbb{E}_p[\langle x - \mu, v\rangle \mid \langle x - \mu, v\rangle \geq \tau_\epsilon(v)] \leq \frac{1-\epsilon}{\epsilon}\rho \text{ whenever } \|v\|_* \leq 1. \tag{17}$$

In particular, the ϵ-quantile satisfies $\tau_\epsilon(v) \leq \frac{1-\epsilon}{\epsilon}\rho$.

In other words, if we project onto any unit vector v in the dual norm, the ϵ-tail of $x - \mu$ must have mean at most $\frac{1-\epsilon}{\epsilon}\rho$. Lemma 0.12 is proved in Section **??**.

The intuition for Lemma 0.12 is the following picture, which is helpful to keep in mind more generally:

Specifically, letting $\hat{\mu} = \mathbb{E}[X \mid E]$, if we have $\|\hat{\mu} - \mu\| = \rho$, then there must be some dual norm unit vector v such that $\langle \hat{\mu} - \mu, v\rangle = \rho$ and $\|v\|_* = 1$. Moreover, for such a v, $\langle \hat{\mu} - \mu, v\rangle$ will be largest when E consists of the $(1-\epsilon)$-fraction of points for which $\langle X - \mu, v\rangle$ is largest. Therefore, resilience reduces to a 1-dimensional problem along each of the dual unit vectors v.

A related result establishes that for $\epsilon = \frac{1}{2}$, resilience in a norm is equivalent to having bounded first moments in the dual norm (see Section **??** for a proof):

Lemma 0.13. *Suppose that p is $(\rho, \frac{1}{2})$-resilient in a norm $\|\cdot\|$, and let $\|\cdot\|_*$ be the dual norm. Then p has 1st moments bounded by 2ρ: $\mathbb{E}_{x\sim p}[|\langle x - \mu, v\rangle|] \leq 2\rho\|v\|_*$ for all $v \in \mathbb{R}^d$.*

Conversely, if p has 1st moments bounded by ρ, it is $(2\rho, \frac{1}{2})$-resilient.

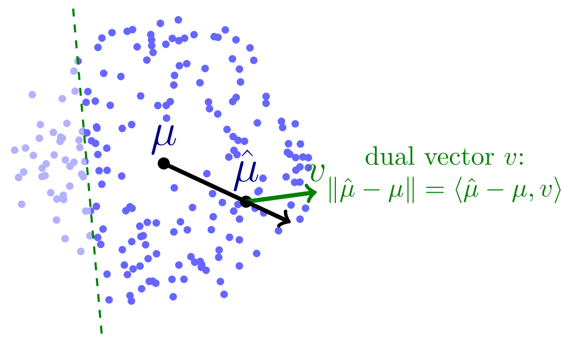

Figure 2: The optimal set T discards the smallest $\epsilon|S|$ elements projected onto a dual unit vector v.

Recap. We saw that the error of the trimmed mean grew as $\sqrt{d}$ in d dimensions, and introduced an alternative estimator–the minimum distance functional–that achieves better error. Specifically, it achieves error 2ρ for the family of (ρ, ϵ)-resilient distributions, which includes all distributions with bounded Orlicz norm (including bounded covariance, bounded moments, and sub-Gaussians).

The definition of resilience is important not just as an analysis tool. Without it, we would need a different estimator for each of the cases of bounded covariance, sub-Gaussian, etc., since the minimum distance functional depends on the family $\mathcal{G}$. Instead, we can always project onto the resilient family $\mathcal{G}_{\mathsf{TV}}(\rho, \epsilon)$ and be confident that this will typically yield an optimal error bound. The only complication is that projection still depends on the parameters ρ and ϵ; however, we can do without knowledge of either one of the parameters as long as we know the other.