1 What is this course about?

Consider the process of building a statistical or machine learning model. We typically first collect training data, then fit a model to that data, and finally use the model to make predictions on new test data.

In theory and in practice, we generally rely on the train and test data coming from the same distribution, or at least being closely related in some way. However, there are several ways this could fail to be the case:

- The data collection process itself could be noisy and thus not reflect the actual underlying signal we wish to learn. For instance, there could be human error in labelling or annotation, or measurement error due to imperfect sensors.
- At test time, inputs could be corrupted due to failures in the input pipeline (a sensor fails in a car, or packets get dropped in a web service) or due to the actions of malicious users (a trespasser trying to fool a face recognition system).
- There could be distributional shift, due to changes in the world over time or because we seek to deploy the model in some new situation (a language model trained on news articles but deployed on twitter).

Robustness concerns what we should do when the train and test distribution are not the same, for any of the reasons above. There are two underlying perspectives influencing the choice of material in this course. First, we are generally interested in worst-case rather than average-case robustness. For instance, when handling data collection errors we will avoid modeling the errors as random noise and instead build procedures that are robust to any errors within some allowed family. We prefer this because average-case robustness requires the errors to satisfy a single, specific distribution for robustness guarantees to be meaningful, while a goal of robustness is to handle unanticipated situations that are difficult to model precisely in advance.

Second, we will study robustness in high-dimensional settings. Many natural approaches to robustness that work in low dimensions fail in high dimensions. For instance, the median is a robust estimate of the mean in one dimension, but the per-coordinate median is a poor robust estimator when the dimension is large (its error grows as $\sqrt{d}$ in $d$ dimensions). We will see that more sophisticated estimators can substantially improve on this first attempt.

We will model robustness with the following general framework: We let $p^*$ denote the true test distribution we wish to estimate, and assume that training data $X_1, \ldots, X_n$ is sampled i.i.d. from some distribution $\tilde{p}$ such that $D(\tilde{p}, p^*) \leq \epsilon$ according to some discrepancy $D$. We also assume that $p^* \in \mathcal{G}$, which encodes the distributional assumptions we make (e.g. that $p^*$ has bounded moments or tails, which is typically necessary for robust estimation to be possible). We benchmark an estimator $\hat{\theta}(X_1, \ldots, X_n)$ according to some cost $L(p^*, \hat{\theta})$ (the test error). The diagram in Figure 1 illustrates this.

This framework captures each of the settings discussed above. However, it will be profitable to think about each case separately due to different emphases:

- For corrupted training data, we think of $\tilde{p}$ as being corrupted and $p^*$ as being nice.
- For corrupted test data, we think of $\tilde{p}$ as being nice and $p^*$ as being corrupted.
• For distributional shift, we think of $\tilde{p}$ and $p^*$ as both being nice (but different).

Additionally, since both $\tilde{p}$ and $p^*$ are nice for distributional shift, we should have greater ambitions and seek to handle larger differences between train and test than in the corruption cases.

Training robustness. Designing robust estimators for training corruptions usually involves reasoning about what the real data “might have” looked like. This could involve operations such as removing outliers, smoothing points away from extremes, etc. Unfortunately, many intuitive algorithms in low dimensions achieve essentially trivial bounds in high dimensions. We will show how to achieve more meaningful bounds, focusing on three aspects:

1. good dependence of the error on the dimension,
2. good finite-sample bounds,
3. computational tractability.

Each of these aspects turns out to require new machinery and we will devote roughly equal space to each.

Test robustness. Here we typically think in terms of a single test input that is perturbed in some way. We essentially want continuity of the estimated model with respect to the perturbation. A major challenge is that continuity is a computationally challenging property to certify, and we will introduce two approaches that give reasonable bounds on the constant of continuity for medium-size problems; these are based on randomized smoothing and on convex relaxations, respectively. Another challenge is formally specifying the norm in which we want to be continuous, since real-world corruptions have complex structure that need not be mathematically convenient. We will discuss some methodological considerations for handling this issue.

Distributional shift. In contrast to test robustness where we focus on single points, distributional shift focuses on large samples or populations of points. This is a crucial difference, as test robustness can do little to leverage statistical knowledge and must instead focus on optimizing the learned model to have good deterministic properties (such as continuity). For distributional shift, we can make use of statistical concepts like model uncertainty to get good out-of-distribution error estimates. Another approach is to seek invariant features or structure that can transfer information from the train to test distributions.

2 Training Time Robustness

We will start our investigation with training time robustness. As in Figure 1, we observe samples $X_1, \ldots, X_n$ from a corrupted training distribution $\tilde{p}$, whose relationship to the true (test) distribution is controlled by the constraint $D(\tilde{p}, p^*) \leq \epsilon$. We additionally constrain $p^* \in G$, which encodes our distributional assumptions.
Table 1: Comparison of different robust settings.

<table>
<thead>
<tr>
<th>Train robustness</th>
<th>Test robustness</th>
<th>Distributional shift</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p^*$ nice</td>
<td>$\tilde{p}$ nice</td>
<td>$p^*$ and $\tilde{p}$ both nice</td>
</tr>
<tr>
<td>statistical properties</td>
<td>deterministic properties</td>
<td>statistical properties</td>
</tr>
<tr>
<td>undo corruptions</td>
<td>robust optimization</td>
<td>invariant features</td>
</tr>
</tbody>
</table>

Figure 2: Possible corruptions to be robust to. Left: data contains outliers. Middle: outputs are perturbed (process noise); Right: inputs are perturbed (measurement error).

Note that this setting corresponds to an oblivious adversary that can only apply corruptions at the population level (changing $p^*$ to $\tilde{p}$); we can also consider a more powerful adaptive adversary that can apply corruptions to the samples themselves. Such an adversary is called adaptive because it is allowed to adapt to the random draw of the samples points $X_1, \ldots, X_n$. Formally defining adaptive adversaries is somewhat technical and we defer this to later.

Figure 2 illustrates several ways in which a training distribution could be corrupted. In the left panel, an $\epsilon$ fraction of real points have been replaced by outliers. This can be modeled by the discrepancy $D(p, q) = TV(p, q)$, where $TV$ is the total variation distance:

$$TV(p, q) \overset{\text{def}}{=} \sup \{|p(E) - q(E)| \mid E \text{ is a measurable event}\}.$$  

(1)

If $p$ and $q$ both have densities then an equivalent characterization is $TV(p, q) = \frac{1}{2} \int |p(x) - q(x)|dx$.

In the middle and right panels of Figure 2, either the inputs or outputs have been moved slightly. Both operations can be modeled using Wasserstein distances (also called earthmover distances), which we will discuss later. For now, however, we will focus on the case of handling outliers. Thus for the next several sections our discrepancy will be the total variation distance $D = TV$.

2.1 Robustness to Outliers in 1 Dimension

First consider mean estimation in one dimension: we observe $n$ data points $x_1, \ldots, x_n \in \mathbb{R}$ drawn from $\tilde{p}$, and our goal is to estimate the mean $\mu = \mathbb{E}_{x \sim p^*}[x]$ of $p^*$. Accordingly our loss is $L(p^*, \theta) = |\theta - \mu(p^*)|$.

The following histogram illustrates a possible dataset, where the height of each bar represents the number of points with a given value:

Are the red points outliers? Or part of the real data? Depending on the conclusion, the estimated mean could vary substantially. Without further assumptions on the data-generating distribution $p^*$, we cannot rule out either case. This brings us to an important principle:
With no assumptions on the distribution \( p^* \), robust estimation is impossible.

In particular, we must make assumptions that are strong enough to reject sufficiently extreme points as outliers, or else even a small fraction of such points can dominate the estimate of the mean. For simplicity, here and in the next several sections we will assume that we directly observe the training distribution \( \tilde{p} \) rather than samples \( x_{1:n} \) from \( p \). This allows us to avoid analyzing finite-sample concentration, which requires introducing additional technical tools that we will turn to in Section 2.5.

**Assumption: bounded variance.** One possible assumption is that \( p^* \) has bounded variance: \( \mathbb{E}_{x \sim p^*}[(x - \mu)^2] \leq \sigma^2 \) for some parameter \( \sigma \). We take \( \mathcal{G} = \mathcal{G}_{\text{cov}}(\sigma) \) to be the set of distributions satisfying this constraint.

Under this assumption, we can estimate \( \mu \) to within error \( O(\sigma \sqrt{\epsilon}) \) under TV-perturbations of size \( \epsilon \). Indeed, consider the following procedure:

**Algorithm 1 TrimmedMean**

1. Remove the upper and lower \((2\epsilon)\)-quantiles from \( \tilde{p} \) (so \( 4\epsilon \) mass is removed in total).
2. Let \( \tilde{p}_2 \) denote the new distribution after re-normalizing, and return the mean of \( \tilde{p}_2 \).

To analyze Algorithm 1, we will make use of a strengthened version of Chebyshev’s inequality, which we recall here (see Section B.1 for a proof):

**Lemma 2.1 (Chebyshev inequality).** Suppose that \( p \) has mean \( \mu \) and variance \( \sigma^2 \). Then, \( \mathbb{P}_{X \sim p}[X \geq \mu + \sigma / \sqrt{3}] \leq \delta \). Moreover, if \( E \) is any event with probability at least \( \delta \), then \( |\mathbb{E}_{X \sim p}[X \mid E] - \mu| \leq \sigma \sqrt{2(1 - \delta)} / \delta \).

The first part, which is the standard Chebyshev inequality, says that it is unlikely for a point to be more than a few standard deviations away from \( \mu \). The second part says that any large population of points must have a mean close to \( \mu \). This second property, which is called resilience, is central to robust estimation, and will be studied in more detail in Section 2.4.

With Lemma 2.1 in hand, we can prove the following fact about Algorithm 1:

**Proposition 2.2.** Assume that \( \text{TV}(\tilde{p}, p^*) \leq \epsilon \leq \frac{1}{8} \). Then the output \( \hat{\mu} \) of Algorithm 1 satisfies \( |\hat{\mu} - \mu| \leq 9\sigma \sqrt{\epsilon} \).

**Proof.** If \( \text{TV}(\tilde{p}, p^*) \leq \epsilon \), then we can get from \( p^* \) to \( \tilde{p} \) by adding an \( \epsilon \)-fraction of points (outliers) and deleting an \( \epsilon \)-fraction of the original points.

First note that all outliers that exceed the \( \epsilon \)-quantile of \( p^* \) are removed by Algorithm 1. Therefore, all non-removed outliers lie within \( \frac{\sigma}{\sqrt{3}} \) of the mean \( \mu \) by Chebyshev’s inequality.

Second, we and the adversary together remove at most a \( 5\epsilon \)-fraction of the mass in \( p^* \). Applying Lemma 2.1 with \( \delta = 1 - 5\epsilon \), the mean of the remaining good points lies within \( \sigma \sqrt{\frac{10\epsilon}{1 - 5\epsilon}} \) of \( \mu \).

Now let \( \epsilon' \) be the fraction of remaining points which are bad, and note that \( \epsilon' \leq \frac{\epsilon}{1 - 4\epsilon} \). The mean of all the remaining points differs from \( \mu \) by at most \( \epsilon' \cdot \sigma \sqrt{\frac{1}{\epsilon} + (1 - \epsilon') \cdot \sigma \sqrt{\frac{10\epsilon}{1 - 5\epsilon}}} \), which is at most \( (1 + \sqrt{10}) \sqrt{\frac{\epsilon}{1 - 4\epsilon}} \cdot \sigma \).

This is in turn at most \( 9\sigma \sqrt{\epsilon} \) assuming that \( \epsilon \leq \frac{1}{8} \).

**Optimality.** The \( O(\sigma \sqrt{\epsilon}) \) dependence is optimal, because the adversary can themselves apply the same trimming procedure we do, and in general this will shift the mean of a bounded covariance distribution by \( O(\sigma \sqrt{\epsilon}) \) while keeping the covariance bounded.

**Alternate assumptions.** The key fact driving the proof of Proposition 2.2 is that any \( (1 - \epsilon) \)-fraction of the good points has mean at most \( O(\sigma \sqrt{\epsilon}) \) away from the true mean due to Chebyshev’s inequality (Lemma 2.1), which makes use of the bound \( \sigma^2 \) on the variance. Any other bound on the deviation from the mean would yield an analogous result. For instance, if \( p^* \) has bounded \( k \)th moment, then the \( O(\sigma \sqrt{\epsilon}) \) in Lemma 2.1 can be improved to \( O(\sigma_k \epsilon^{1/k}) \), where \( (\sigma_k)^k \) is a bound on the \( k \)th moment; in this case Algorithm 1 will estimate \( \mu \) with a correspondingly improved error of \( O(\sigma_k \epsilon^{1/k}) \).
Figure 3: The outliers can lie at distance $\sqrt{d}$ without being detected, skewing the mean by $\epsilon \sqrt{d}$.

2.2 Problems in High Dimensions

In the previous section, we saw how to robustly estimating the mean of a 1-dimensional dataset assuming the true data had bounded variance. Our estimator worked by removing data points that are too far away from the mean, and then returning the mean of the remaining points.

It is tempting to apply this same idea in higher dimensions—for instance, removing points that are far away from the mean in $\ell_2$-distance. Unfortunately, this incurs large error in high dimensions.

To see why, consider the following simplified example. The distribution $p^*$ over the true data is an isotropic Gaussian $\mathcal{N}(\mu, I)$, with unknown mean $\mu$ and independent variance 1 in every coordinate. In this case, the typical distance $\|x_i - \mu\|_2$ of a sample $x_i$ from the mean $\mu$ is roughly $\sqrt{d}$, since there are $d$ coordinates and $x_i$ differs from $\mu$ by roughly 1 in every coordinate. (In fact, $\|x_i - \mu\|_2$ can be shown to concentrate around $\sqrt{d}$ with high probability.) This means that the outliers can lie at a distance $\sqrt{d}$ from $\mu$ without being detected, thus shifting the mean by $\Theta(\epsilon \sqrt{d})$; Figure 3 depicts this. Therefore, filtering based on $\ell_2$ distance incurs an error of at least $\epsilon \sqrt{d}$. This dimension-dependent $\sqrt{d}$ factor often renders bounds meaningless.

In fact, the situation is even worse; not only are the bad points no further from the mean than the good points in $\ell_2$-distance, they actually have the same probability density under the true data-generating distribution $\mathcal{N}(\mu, I)$. There is thus no procedure that measures each point in isolation and can avoid the $\sqrt{d}$ factor in the error.

This leads us to an important take-away: in high dimensions, outliers can substantially perturb the mean while individually looking innocuous. To handle this, we will instead need to analyze entire populations of outliers at once. In the next section we will do this using minimum distance functionals, which will allow us to avoid the dimension-dependent error.

[ Lecture 2 ]

2.3 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 Donoho and Liu (1988).

**Definition 2.3** (Minimum distance functional). For a family $\mathcal{G}$ and discrepancy $D$, the minimum distance functional is

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

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 $\epsilon$. More importantly, it satisfies the following cost bound in terms of the modulus of continuity of $\mathcal{G}$:
The equality (i) is a couple lines of algebra, but is easiest to see by drawing a graph of the two Gaussians above. Then

\[ m(\mathcal{G}, 2\epsilon, D, L) \triangleq \sup_{p, q \in \mathcal{G} : D(p, q) \leq 2\epsilon} L(p, \theta^*(q)). \tag{3} \]

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

Specialize again to the case \( D = \text{TV} \) 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 \( \sup_{p, q \in \mathcal{G} : \text{TV}(p, q) \leq \epsilon} \|\mu(p) - \mu(q)\|_2 \). As a concrete example, let \( \mathcal{G} \) be the family of Gaussian distributions with unknown mean \( \mu \) and identity covariance. For this family, the TV distance is essentially linear in the difference in mean:

**Lemma 2.5.** Let \( \mathcal{N}(\mu, I) \) denote a Gaussian distribution with mean \( \mu \) and identity covariance. Then

\[ \min(u/2, 1)/\sqrt{2\pi} \leq \text{TV}(\mathcal{N}(\mu, I), \mathcal{N}(\mu', I)) \leq \min(u/\sqrt{2\pi}, 1), \tag{4} \]

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

\[
\text{TV}(\mathcal{N}(-u/2, 1), \mathcal{N}(u/2, 1)) = \frac{1}{2\sqrt{2\pi}} \int_{-\infty}^{\infty} |e^{-(t+u/2)^2/2} - e^{-(t-u/2)^2/2}| dt
\]

\[
\overset{(i)}{=} \frac{1}{2\sqrt{2\pi}} \int_{-u/2}^{u/2} e^{-t^2/2} dt.
\]

(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 \( e^{-t^2/2} \geq \frac{1}{2} \) if \( |t| \leq 1 \).

For the upper bound, similarly note that \( e^{-t^2/2} \leq 1 \) 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 2.5 allows us to compute the modulus for Gaussians:

**Corollary 2.6.** Let \( \mathcal{G}_{\text{gauss}} \) be the family of isotropic Gaussians, \( D = \text{TV} \), and \( L \) the difference in means as above. Then \( m(\mathcal{G}_{\text{gauss}}, \epsilon, D, L) \leq 2\sqrt{2\pi} \epsilon \) whenever \( \epsilon \leq \frac{1}{2\sqrt{2\pi}} \).

In particular, by Proposition 2.4 the minimum distance functional achieves error \( \mathcal{O}(\epsilon) \) for Gaussian distributions when \( \epsilon \leq \frac{1}{2\sqrt{2\pi}} \). This improves substantially on the \( \epsilon\sqrt{d} \) error of the trimmed mean estimator from Section 2.2. 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 \( m \) is bounded for much more general families. As one example, we can take the distributions with bounded covariance (compare to Proposition 2.2):

**Lemma 2.7.** Let \( \mathcal{G}_{\text{cov}}(\sigma) \) be the family of distributions whose covariance matrix \( \Sigma \) satisfies \( \Sigma \leq \sigma^2 I \). Then \( m(\mathcal{G}_{\text{cov}}(\sigma), \epsilon) = \mathcal{O}(\sigma \sqrt{\epsilon}) \).

**Proof.** Let \( p, q \in \mathcal{G}_{\text{cov}}(\sigma) \) such that \( \text{TV}(p, q) \leq \epsilon \). This means that we can get from \( p \) to \( q \) by first deleting \( \epsilon \) mass from \( p \) and then adding \( \epsilon \) 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 \( \epsilon \) mass (and then renormalizing). In fact, this distribution is exactly

\[
r = \frac{\min(p, q)}{1 - \text{TV}(p, q)}.
\]

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 2.1):
Lemma 2.8 (Chebyshev in $\mathbb{R}^d$). Suppose that $p$ has mean $\mu$ and covariance $\Sigma$, where $\Sigma \preceq \sigma^2 I$. Then, if $E$ is any event with probability at least $\delta$, we have $\|E_{X \sim p}[X | E] - \mu\|_2 \leq \sigma \sqrt{\frac{2(1-\delta)}{\delta}}$.

As a consequence, we have $\|\mu(r) - \mu(p)\|_2 \leq \sigma \sqrt{\frac{2\epsilon}{(1-\epsilon)}}$ and $\|\mu(r) - \mu(q)\|_2 \leq \sigma \sqrt{\frac{2\epsilon}{(1-\epsilon)}}$ (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 2.7, 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.

2.4 Resilience

Here we generalize Lemma 2.7 to prove that the modulus of continuity $m$ is bounded for a general family of distributions containing Gaussians, sub-Gaussians, bounded covariance distributions, and many others. The main observation is that in the proof of Lemma 2.7, all we needed was that the tails of distributions in $\mathcal{G}$ were bounded, in the sense that deleting an $\epsilon$-fraction of the points could not substantially change the mean. This motivates the following definition:

**Definition 2.9.** A distribution $p$ over $\mathbb{R}^d$ is said to be $(\rho, \epsilon)$-resilient (with respect to some norm $\| \cdot \|$) if

$$\|E_{X \sim p}[X | E] - E_{X \sim p}[X]\| \leq \rho$$

for all events $E$ with $p(E) \geq 1 - \epsilon$. (8)

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

We observe that $\mathcal{G}_{cov}(\sigma) \subset \mathcal{G}_{TV}(\sigma \sqrt{2\epsilon/(1-\epsilon)}, \epsilon)$ for all $\epsilon$ by Lemma 2.8; in other words, bounded covariance distributions are resilient. We can also show that $\mathcal{G}_{gauss} \subset \mathcal{G}_{TV}(2\epsilon \sqrt{\log(1/\epsilon)}, \epsilon)$, so Gaussians are resilient as well.

Resilient distributions always have bounded modulus:

**Theorem 2.10.** The modulus of continuity $m(\mathcal{G}_{TV}, 2\epsilon)$ satisfies the bound

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

(9)

whenever $\epsilon < 1/2$.

**Proof.** As in Lemma 2.7, the key idea is that any two distributions $p, q$ that are close in TV have a midpoint distribution $r = \min(p, q)$ (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 $\mathcal{G}_{TV}$ is bounded. We illustrate this idea in Figure 4 and make it precise below.

Recall that

$$m(\mathcal{G}_{TV}(\rho, \epsilon), 2\epsilon) = \sup_{p, q \in \mathcal{G}_{TV}(\rho, \epsilon)} \|\mu(p) - \mu(q)\|_2.$$  

(10)

From $TV(p, q) \leq 2\epsilon$, we know that $r = \frac{\min(p, q)}{1 - TV(p, q)}$ can be obtained from either $p$ and $q$ by conditioning on an event of probability $1 - \epsilon$. It then follows from $p, q \in \mathcal{G}_{TV}(\rho, \epsilon)$ that $\|\mu(p) - \mu(r)\| \leq \epsilon$ and similarly $\|\mu(q) - \mu(r)\| \leq \epsilon$. Thus by the triangle inequality $\|\mu(p) - \mu(q)\| \leq 2\rho$, 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 $G_{\text{cov}}$ that is implied by resilience is optimal ($O(\sigma \sqrt{\epsilon})$), while for $G_{\text{gauss}}$ it is optimal up to log factors ($O(\epsilon \sqrt{\log(1/\epsilon)})$ vs. $O(\epsilon)$).

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

$$\|X\|_{\psi} \triangleq \inf \left\{ t > 0 : \mathbb{E}_{X}[\psi\left(\frac{|X|}{t}\right)] \leq 1 \right\}. \quad (11)$$

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

As special cases, we say that a random variable $X \sim p$ is sub-Gaussian with parameter $\sigma$ if $\|\langle X - \mathbb{E}_p[X], v\rangle\|_{\psi_2} \leq \sigma$ whenever $\|v\|_2 \leq 1$, 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 2.11 applies to distributions on $\mathbb{R}$, but we can generalize this to distributions on $\mathbb{R}^d$ by taking one-dimensional projections:

**Definition 2.12 (Orlicz norm in $\mathbb{R}^d$).** For a random variable $X \in \mathbb{R}^d$ and Orlicz function $\psi$, we define the $d$-dimensional $\psi$-norm as

$$\|X\|_{\psi} \triangleq \inf \{ t > 0 : \|\langle X, v \rangle\|_{\psi} \leq t \text{ whenever } \|v\|_2 \leq 1 \}. \quad (12)$$

We let $G_{\psi}(\sigma)$ denote the distributions with bounded $\psi$-norm as in Definition 2.11.

Thus a distribution has bounded $\psi$-norm if each of its 1-dimensional projections does. As an example, $G_{\text{cov}}(\sigma) = G_{\psi}(\sigma)$ for $\psi(x) = x^2$, so Orlicz norms generalize bounded covariance. It is also possible to generalize Definition 2.12 to norms other than the $\ell_2$-norm, which we will see in an exercise.

Functions with bounded Orlicz norm are resilient:

**Lemma 2.13.** The family $G_{\psi}(\sigma)$ is contained in $G_{\text{TV}}(2\sigma \epsilon \psi^{-1}(1/\epsilon), \epsilon)$ 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

$$
\| E_{X \sim p}[X \mid E] \|_2 \overset{(i)}{=} \frac{\epsilon'}{1 - \epsilon'} \| E_{X \sim p}[X \mid E^c] \|_2 \\
= \frac{\epsilon'}{1 - \epsilon'} \sup_{\| v \|_2 \leq 1} \| E_{X \sim p}[(X, v) \mid E^c] \|_2 \overset{(ii)}{\leq} \frac{\epsilon'}{1 - \epsilon'} \sup_{\| v \|_2 \leq 1} \sigma \psi^{-1}(E_{X \sim p}[\psi(\langle X, v \rangle/\sigma) \mid E^c]) \overset{(iii)}{\leq} \frac{\epsilon'}{1 - \epsilon'} \sup_{\| v \|_2 \leq 1} \sigma \psi^{-1}(E_{X \sim p}[\psi(\langle X, v \rangle/\sigma)])/\epsilon' \overset{(iv)}{\leq} \frac{\epsilon'}{1 - \epsilon'} \sigma \psi^{-1}(1/\epsilon') \leq 2\epsilon \sigma \psi^{-1}(1/\epsilon),$$

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

As a consequence, the modulus $m$ of $G_\psi(\sigma)$ is $O(\sigma \epsilon \psi^{-1}(1/\epsilon))$, and hence the minimum distance functional estimates the mean with error $O(\sigma \epsilon \psi^{-1}(1/\epsilon))$. Note that for $\psi(x) = x^2$ this reproduces our result for bounded covariance. For $\psi(x) = x^k$ we get error $O(\sigma \epsilon \psi^{-1}(1/k))$ when a distribution has $k$th moments bounded by $\sigma^k$. Similarly for sub-Gaussian distributions we get error $O(\sigma \epsilon \sqrt{\log(1/\epsilon)})$. We will show in an exercise that the error bound implied by Lemma 2.13 is optimal for any Orlicz function $\psi$.

**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 2.14.** For a fixed vector $v$, let $\tau_\epsilon(v)$ denote the $\epsilon$-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 $(\rho, \epsilon)$-resilient in a norm $\| \cdot \|$ if and only if the $\epsilon$-tail of $p$ has bounded mean when projected onto any dual unit vector $v$:

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

In particular, the $\epsilon$-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 $\epsilon$-tail of $x - \mu$ must have mean at most $\frac{1 - \epsilon}{\epsilon} \rho$. Lemma 2.14 is proved in Section C.

The intuition for Lemma 2.14 is the following picture, which is helpful to keep in mind more generally:

![Figure 5: The optimal set $T$ discards the smallest $\epsilon|S|$ elements projected onto a dual unit vector $v$.](image)

Specifically, letting $\hat{\mu} = E[X \mid E]$, 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 D for a proof):

**Lemma 2.15.** 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 $2\rho: \mathbb{E}_{x \sim p}[|\langle x - \mu, v \rangle|] \leq 2\rho\|v\|_*$ for all $v \in \mathbb{R}^d$.

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

**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 $2\rho$ for the family of $(\rho, \epsilon)$-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 $G$. Instead, we can always project onto the resilient family $G_{TV}(\rho, \epsilon)$ and be confident that this will typically yield an optimal error bound. The only complication is that projection still depends on the parameters $\rho$ and $\epsilon$; however, we can do without knowledge of either one of the parameters as long as we know the other.

[Lecture 3]

### 2.5 Concentration Inequalities

So far we have only considered the infinite-data limit where we directly observe $\tilde{p}$; but in general we would like to analyze what happens in finite samples where we only observe $X_1, \ldots, X_n$ sampled independently from $\tilde{p}$. In order to do this, we will want to be able to formalize statements such as “if we take the average of a large number of samples, it converges to the population mean”. In order to do this, we will need a set of mathematical tools called concentration inequalities. A proper treatment of concentration could itself occupy an entire course, but we will cover the ideas here that are most relevant for our later analyses. See Boucheron et al. (2003), Boucheron et al. (2013), or Ledoux (2001) for more detailed expositions. Terence Tao also has some well-written lectures notes.

Concentration inequalities usually involve two steps:

1. We establish concentration for a single random variable, in terms of some property of that random variable.

2. We show that the property composes nicely for products of independent random variables.

A prototypical example (covered below) is showing that (1) a random variable has at most a $1/t^2$ probability of being $t$ standard deviations from its mean; and (2) the standard deviation of a sum of $n$ i.i.d. random variables is $\sqrt{n}$ times the standard deviation of a single variable.

The simplest concentration inequality is **Markov’s inequality**. Consider the following question:

A slot machine has an expected pay-out of $5 (and its payout is always non-negative). What can we say about the probability that it pays out at least $100? We observe that the probability must be at most 0.05, since a 0.05 chance of a $100 payout would by itself already contribute $5 to the expected value. Moreover, this bound is achievable by taking a slot machine that pays $0 with probability 0.95 and $100 with probability 0.05. Markov’s inequality is the generalization of this observation:

**Theorem 2.16 (Markov’s inequality).** Let $X$ be a non-negative random variable with mean $\mu$. Then, $\mathbb{P}[X \geq t \cdot \mu] \leq \frac{1}{t}$. 

10
Markov’s inequality accomplishes our first goal of establishing concentration for a single random variable, but it has two issues: first, the $\frac{1}{2}$ tail bound decays too slowly in many cases (we instead would like exponentially decaying tails); second, Markov’s inequality doesn’t compose well and so doesn’t accomplish our second goal.

We can address both issues by applying Markov’s inequality to some transformed random variable. For instance, applying Markov’s inequality to the random variable $Z = (X - \mu)^2$ yields the stronger Chebyshev inequality:

**Theorem 2.17** (Chebyshev’s inequality). Let $X$ be a real-valued random variable with mean $\mu$ and variance $\sigma^2$. Then, $\Pr[|X - \mu| \geq \tau \cdot \sigma] \leq \frac{1}{\tau^2}$.

**Proof.** Since $Z = (X - \mu)^2$ is non-negative, we have that $\Pr[Z \geq \tau^2 \cdot \sigma^2] \leq \frac{1}{\tau^2}$ by Markov’s inequality. Taking the square-root gives $\Pr[|X - \mu| \geq \tau \cdot \sigma] \leq \frac{1}{\tau}$, as was to be shown.

Chebyshev’s inequality improves the $1/t$ dependence to $1/t^2$. But more importantly, it gives a bound in terms of a quantity (the variance $\sigma^2$) that composes nicely:

**Lemma 2.18** (Additivity of variance). Let $X_1, \ldots, X_n$ be pairwise independent random variables, and let $\text{Var}[X]$ denote the variance of $X$. Then,

$$\text{Var}[X_1 + \cdots + X_n] = \text{Var}[X_1] + \cdots + \text{Var}[X_n].$$

**Proof.** It suffices by induction to prove this for two random variables. Without loss of generality assume that both variables have mean zero. Then we have $\text{Var}[X + Y] = \mathbb{E}[(X + Y)^2] = \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\mathbb{E}[XY] = \text{Var}[X] + \text{Var}[Y] + 2\mathbb{E}[X]\mathbb{E}[Y] = \text{Var}[X] + \text{Var}[Y]$, where the second-to-last step uses pairwise independence.

Chebyshev’s inequality together with Lemma 2.18 together allow us to show that an average of i.i.d. random variables converges to its mean at a $1/\sqrt{n}$ rate:

**Corollary 2.19.** Suppose $X_1, \ldots, X_n$ are drawn i.i.d. from $p$, where $p$ has mean $\mu$ and variance $\sigma^2$. Also let $S = \frac{1}{n}(X_1 + \cdots + X_n)$. Then, $\Pr[|S - \mu|/\sigma \geq \sqrt{n}/\sqrt{t}] \leq 1/t^2$.

**Proof.** Lemma 2.18 implies that $\text{Var}[S] = \sigma^2/n$, from which the result follows by Chebyshev’s inequality.

**Higher moments.** Chebyshev’s inequality gives bounds in terms of the second moment of $X - \mu$. Can we do better by considering higher moments such as the 4th moment? Supposing that $\mathbb{E}[(X - \mu)^4] \leq \tau^4$, we do get the analogous bound $\Pr[|X - \mu| \geq \tau \cdot \sigma] \leq 1/t^4$. However, the 4th moment doesn’t compose as nicely as the variance; if $X$ and $Y$ are two independent mean-zero random variables, then we have

$$\mathbb{E}[(X + Y)^4] = \mathbb{E}[X^4] + \mathbb{E}[Y^4] + 6\mathbb{E}[X^2]\mathbb{E}[Y^2],$$

where the $\mathbb{E}[X^2]\mathbb{E}[Y^2]$ can’t be easily dealt with. It is possible to bound higher moments under composition, for instance using the Rosenthal inequality which states that

$$\mathbb{E}[\|\sum X_i\|^p] \leq O(p)^p \sum \mathbb{E}[\|X_i\|^p] + O(\sqrt{p})^p(\sum \mathbb{E}[X_i^2])^{p/2}$$

for independent random variables $X_i$. Note that the first term on the right-hand-side typically grows as $n \cdot O(p)^p$ while the second term typically grows as $O(\sqrt{p})^p$.

We will typically not take the Rosenthal approach and instead work with an alternative, nicer object called the moment generating function:

$$m_X(\lambda) \overset{\text{def}}{=} \mathbb{E}[\exp(\lambda(X - \mu))].$$

For independent random variables, the moment generating function composes via the identity $m_{X_1 + \cdots + X_n}(\lambda) = \prod_{i=1}^n m_{X_i}(\lambda)$. Applying Markov’s inequality to the moment generating function yields the Chernoff bound:

**Theorem 2.20** (Chernoff bound). For a random variable $X$ with moment generating $m_X(\lambda)$, we have

$$\Pr[X - \mu \geq t] \leq \inf_{\lambda \geq 0} m_X(\lambda)e^{-\lambda t}.$$ 

**Proof.** By Markov’s inequality, $\Pr[X - \mu \geq t] = \Pr[\exp(\lambda(X - \mu)) \geq \exp(\lambda t)] \leq \mathbb{E}[\exp(\lambda(X - \mu))]/\exp(\lambda t)$, which is equal to $m_X(\lambda)e^{-\lambda t}$ by the definition of $m_X$. Taking inf over $\lambda$ yields the claimed bound.
Sub-exponential and sub-Gaussian distributions. An important special case is sub-exponential random variables; recall these are random variables satisfying $\mathbb{E}[\exp(|X - \mu|/\sigma)] \leq 2$. For these, applying the Chernoff bound with $\lambda = 1/\sigma$ yields $\mathbb{P}[X - \mu \geq t] \leq 2e^{-t/\sigma}$.

Another special case is sub-Gaussian random variables (those satisfying $\mathbb{E}[\exp((X - \mu)^2/\sigma^2)] \leq 2$). In this case, using the inequality $ab \leq a^2/4 + b^2$, we have

$$m_X(\lambda) = \mathbb{E}[\exp(\lambda(X - \mu))] \leq \mathbb{E}[\exp(\lambda^2\sigma^2/4 + (X - \mu)^2/\sigma^2)] \leq 2 \exp(\lambda^2\sigma^2/4).$$

The factor of 2 is pesky and actually we can get the more convenient bound $m_X(\lambda) \leq \exp(3\lambda^2\sigma^2/2)$ (Rivasplata, 2012). Plugging this into the Chernoff bound yields $\mathbb{P}[X - \mu \geq t] \leq \exp(3\lambda^2\sigma^2/2 - \lambda t)$; minimizing over $\lambda$ gives the optimized bound $\mathbb{P}[X - \mu \geq t] \leq \exp(-t^2/6\sigma^2)$.

Sub-Gaussians are particularly convenient because the bound $m_X(\lambda) \leq \exp(3\lambda^2\sigma^2/2)$ composes well. Let $X_1, \ldots, X_n$ be independent sub-Gaussians with constants $\sigma_1, \ldots, \sigma_n$. Then we have $m_{X_1 + \cdots + X_n}(\lambda) \leq \exp(3\lambda^2(\sigma_1^2 + \cdots + \sigma_n^2)/2)$. We will use this to bound the behavior of sums of bounded random variables using Höftding’s inequality:

**Theorem 2.21** (Hoeffding’s inequality). Let $X_1, \ldots, X_n$ be zero-mean random variables lying in $[-M, M]$, and let $S = \frac{1}{n}(X_1 + \cdots + X_n)$. Then, $\mathbb{P}[S \geq t] \leq \exp(-\ln(2)nt^2/6M^2) \leq \exp(-nt^2/9M^2)$.

**Proof.** First, note that each $X_i$ is sub-Gaussian with parameter $\sigma = M/\sqrt{\ln 2}$, since $\mathbb{E}[\exp(X_i^2/\sigma^2)] \leq \exp(M^2/\sigma^2) = \exp(\ln(2)) = 2$. We thus have $m_{X_i}(\lambda) \leq \exp(3\lambda^2M^2/2\ln 2)$, and so by the multiplicativity of moment generating functions we obtain $m_S(\lambda) \leq \exp(3\lambda^2M^2/(2n\ln 2))$. Plugging into Chernoff’s bound and optimizing $\lambda$ as before yields $\mathbb{P}[S \geq t] \leq \exp(-\ln(2)nt^2/6M^2)$ as claimed.

Hoeffding’s inequality shows that a sum of independent random variables converges to its mean at a $1/\sqrt{n}$ rate, with tails that decay as fast as a Gaussian as long as each of the individual variables is bounded. Compare this to the $1/t^2$ decay that we obtained earlier through Chebyshev’s inequality.

**Cumulants.** The moment generating function is a convenient tool because it multiplies over independent random variables. However, its existence requires that $X$ already have thin tails, since $\mathbb{E}[\exp(\lambda X)]$ must be finite. For heavy-tailed distributions a (laborious) alternative is to use cumulants.

The cumulant function is defined as

$$K_X(\lambda) \overset{\text{def}}{=} \log \mathbb{E}[\exp(\lambda X)].$$

Note this is the log of the moment-generating function. Taking the log is convenient because now we have additivity: $K_{X+Y}(\lambda) = K_X(\lambda) + K_Y(\lambda)$ for independent $X, Y$. Cumulants are obtained by writing $K_X(\lambda)$ as a power series:

$$K_X(\lambda) = 1 + \sum_{n=1}^{\infty} \frac{\kappa_n(X)}{n!} \lambda^n.$$

When $\mathbb{E}[X] = 0$, the first few values of $\kappa_n$ are:

$$\kappa_1(X) = 0,$$
$$\kappa_2(X) = \mathbb{E}[X^2],$$
$$\kappa_3(X) = \mathbb{E}[X^3],$$
$$\kappa_4(X) = \mathbb{E}[X^4] - 3\mathbb{E}[X^2]^2,$$
$$\kappa_5(X) = \mathbb{E}[X^5] - 10\mathbb{E}[X^3]\mathbb{E}[X^2],$$
$$\kappa_6(X) = \mathbb{E}[X^6] - 16\mathbb{E}[X^4]\mathbb{E}[X^2] - 10\mathbb{E}[X^3]^2 + 30\mathbb{E}[X^2]^3.$$

Since $K$ is additive, each of the $\kappa_n$ are as well. Thus while we ran into the issue that $\mathbb{E}[(X + Y)^4] \neq \mathbb{E}[X^4] + \mathbb{E}[Y^4]$, it is the case that $\kappa_4(X + Y) = \kappa_4(X) + \kappa_4(Y)$ as long as $X$ and $Y$ are independent. By going back and forth between moments and cumulants it is possible to obtain tail bounds even if only some of the moments exist. However, this can be arduous and Rosenthal’s inequality is probably the better route in such cases.

---

1 Most of the constants presented here are suboptimal; we have focused on giving simpler proofs at the expense of sharp constants.
2.5.1 Applications of concentration inequalities

Having developed the machinery above, we next apply it to a few concrete problems to give a sense of how to use it. A key lemma which we will use repeatedly is the union bound, which states that if \( E_1, \ldots, E_n \) are events with probabilities \( \pi_1, \ldots, \pi_n \), then the probability of \( E_1 \cup \cdots \cup E_n \) is at most \( \pi_1 + \cdots + \pi_n \). A corollary is that if \( n \) events each have probability \( \ll 1/n \), then there is a large probability that none of the events occur.

Maximum of sub-Gaussians. Suppose that \( X_1, \ldots, X_n \) are mean-zero sub-Gaussian with parameter \( \sigma \), and let \( Y = \max_{i=1}^n X_i \). How large is \( Y \)? We will show the following:

Lemma 2.22. The random variable \( Y \) is \( O(\sigma \sqrt{\log(n/\delta)}) \) with probability \( 1 - \delta \).

Proof. By the Chernoff bound for sub-Gaussians, we have that \( \mathbb{P}[X_i \geq \sigma \sqrt{6 \log(n/\delta)}] \leq \exp(-\log(n/\delta)) = \delta/n \). Thus by the union bound, the probability that any of the \( X_i \) exceed \( \sigma \sqrt{6 \log(n/\delta)} \) is at most \( \delta \). Thus with probability at least \( 1 - \delta \) we have \( Y \leq \sigma \sqrt{6 \log(n/\delta)} \), as claimed.

Lemma 2.22 illustrates a typical proof strategy: We first decompose the event we care about as a union of simpler events, then show that each individual event holds with high probability by exploiting independence.

Eigenvalue of random matrix. Let \( X_1, \ldots, X_n \) be independent zero-mean sub-Gaussian variables in \( \mathbb{R}^d \) with parameter \( \sigma \), and let \( M = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \). How large is \( \|M\| \), the maximum eigenvalue of \( M \)? We will show:

Lemma 2.23. The maximum eigenvalue \( \|M\| \) is \( O(\sigma^2 \cdot (1 + d/n + \log(1/\delta)/n)) \) with probability \( 1 - \delta \).

Proof. The maximum eigenvalue can be expressed as

\[
\|M\| = \sup_{\|v\|_2 \leq 1} v^\top M v = \sup_{\|v\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^n |\langle X_i, v \rangle|^2.
\]

The quantity inside the sup is attractive to analyze because it is an average of independent random variables. Indeed, we have

\[
\mathbb{E}[\exp(\frac{n}{\sigma^2} v^\top M v)] = \mathbb{E}[\exp(\sum_{i=1}^n |\langle X_i, v \rangle|^2 / \sigma^2)]
\]

\[
= \prod_{i=1}^n \mathbb{E}[\exp(|\langle X_i, v \rangle|^2 / \sigma^2)] \leq 2^n \exp(-nt/\sigma^2).
\]

where the last step follows by sub-Gaussianity if \( \langle X_i, v \rangle \). The Chernoff bound then gives \( \mathbb{P}[v^\top M v \geq t] \leq 2^n \exp(-nt/\sigma^2) \).

If we were to follow the same strategy as Lemma 2.22, the next step would be to union bound over \( v \). Unfortunately, there are infinitely many \( v \) so we cannot do this directly. Fortunately, we can get by with only considering a large but finite number of \( v \); we will construct a finite subset \( N_{1/4} \) of the unit ball such that

\[
\sup_{v \in N_{1/4}} v^\top M v \geq \frac{1}{2} \sup_{\|v\|_2 \leq 1} v^\top M v.
\]

Our construction follows Section 5.2.2 of Vershynin (2010). Let \( N_{1/4} \) be a maximal set of points in the unit ball such that \( \|x - y\|_2 \geq 1/4 \) for all distinct \( x, y \in N_{1/4} \). We observe that \( |N_{1/4}| \leq 9^d \); this is because the balls of radius 1/8 around each point in \( N_{1/4} \) are disjoint and contained in a ball of radius 9/8.
To establish (36), let \( v \) maximize \( v^\top Mv \) over \( \|v\|_2 \leq 1 \) and let \( u \) maximize \( v^\top Mv \) over \( \mathcal{N}_{1/4} \). Then
\[
|v^\top Mv - u^\top Mu| = |v^\top M(v - u) + u^\top M(v - u)| \leq (\|v\|_2 + \|u\|_2)\|M\|\|v - u\|_2 \leq 2 \cdot \|M\| \cdot (1/4) = \|M\|/2.
\]
Since \( v^\top Mv = \|M\| \), we obtain \( \|M\| - u^\top Mu \leq \|M\|/2 \), whence \( u^\top Mu \geq \|M\|/2 \), which establishes (36). We are now ready to apply the union bound: Recall that from the Chernoff bound on \( v^\top Mv \), we had
\[
\mathbb{P}[v^\top Mv \geq t] \leq 2^t \exp(-nt/\sigma^2),
\]
so
\[
\mathbb{P}[\sup_{v \in \mathcal{N}_{1/4}} v^\top Mv \geq t] \leq 9^d 2^n \exp(-nt/\sigma^2).
\]
Solving for this quantity to equal \( \delta \), we obtain
\[
t = \frac{\sigma^2}{n} \cdot (n \log(2) + d \log(9) + \log(1/\delta)) = O(\sigma^2 \cdot (1 + d/n + \log(1/\delta)/n)),
\]
as was to be shown.

**VC dimension.** Our final example will be important in the following section: it concerns how quickly a family of events with certain geometric structure converges to its expectation. Let \( \mathcal{H} \) be a collection of functions \( f : \mathcal{X} \to \{0, 1\} \), and define the VC dimension \( \text{vc}(\mathcal{H}) \) to be the maximum \( d \) for which there are points \( x_1, \ldots, x_d \) such that \( (f(x_1), \ldots, f(x_d)) \) can take on all \( 2^d \) possible values. For instance:

- If \( \mathcal{X} = \mathbb{R} \) and \( \mathcal{H} = \{1[x \geq \tau] \mid \tau \in \mathbb{R}\} \) is the family of threshold functions, then \( \text{vc}(\mathcal{H}) = 1 \).
- If \( \mathcal{X} = \mathbb{R}^d \) and \( \mathcal{H} = \{1[\langle x, v \rangle \geq \tau] \mid v \in \mathbb{R}^d, \tau \in \mathbb{R}\} \) is the family of half-spaces, then \( \text{vc}(\mathcal{H}) = d + 1 \).

Additionally, for a point set \( S = \{x_1, \ldots, x_n\} \), let \( V_\mathcal{H}(S) \) denote the number of distinct values of \( (f(x_1), \ldots, f(x_n)) \) and \( V_\mathcal{H}(n) = \max\{V_\mathcal{H}(S) \mid |S| = n\} \). Thus the VC dimension is exactly the maximum \( n \) such that \( V_\mathcal{H}(n) = 2^n \).

We will show the following:

**Proposition 2.24.** Let \( \mathcal{H} \) be a family of functions with \( \text{vc}(\mathcal{H}) = d \), and let \( X_1, \ldots, X_n \sim p \) be i.i.d. random variables over \( \mathcal{X} \). For \( f : \mathcal{X} \to \{0, 1\} \), let \( \nu_n(f) = \#\{i \mid f(X_i) = 1\} \) and let \( \nu(f) = p(f(X) = 1) \). Then
\[
\sup_{f \in \mathcal{H}} |\nu_n(f) - \nu(f)| \leq O\left(\sqrt{\frac{d + \log(1/\delta)}{n}}\right)
\]
with probability \( 1 - \delta \).

We will prove a weaker result that has a \( d \log(n) \) factor instead of \( d \), and which bounds the expected value rather than giving a probability \( 1 - \delta \) bound. The \( \log(1/\delta) \) tail bound follows from McDiarmid’s inequality, which is a standard result in a probability course but requires tools that would take us too far afield. Removing the \( \log(n) \) factor is slightly more involved and uses a tool called chaining.

**Proof of Proposition 2.24.** The importance of the VC dimension for our purposes lies in the Sauer-Shelah lemma:

**Lemma 2.25** (Sauer-Shelah). Let \( d = \text{vc}(\mathcal{H}) \). Then \( V_\mathcal{H}(n) \leq \sum_{k=0}^d \binom{n}{k} \leq 2n^d \).

It is tempting to union bound over the at most \( V_\mathcal{H}(n) \) distinct values of \( (f(X_1), \ldots, f(X_n)) \); however, this doesn’t work because revealing \( X_1, \ldots, X_n \) uses up all of the randomness in the problem and we have no randomness left from which to get a concentration inequality! We will instead have to introduce some new randomness using a technique called symmetrization.
Regarding the expectation, let \( X_1', \ldots, X_n' \) be independent copies of \( X_1, \ldots, X_n \) and let \( \nu'_n(f) \) denote the version of \( \nu_n(f) \) computed with the \( X_i' \). Then we have
\[
\mathbb{E}_X [\sup_{f \in \mathcal{H}} |\nu_n(f) - \nu(f)|] \leq \mathbb{E}_X, X' [\sup_{f \in \mathcal{H}} |\nu_n(f) - \nu'_n(f)|]
\]
\[
= \frac{1}{n} \mathbb{E}_{X, X'} [\sup_{f \in \mathcal{H}} \sum_{i=1}^n f(X_i) - f(X'_i)].
\]

We can create our new randomness by noting that since \( X_i \) and \( X'_i \) are identically distributed, \( f(X_i) - f(X'_i) \) has the same distribution as \( s_i(f(X_i) - f(X'_i)) \), where \( s_i \) is a random sign variable that is \( \pm 1 \) with equal probability. Introducing these variables and continuing the inequality, we thus have
\[
\frac{1}{n} \mathbb{E}_{X, X'} [\sup_{f \in \mathcal{H}} \sum_{i=1}^n f(X_i) - f(X'_i)] = \frac{1}{n} \mathbb{E}_{X, X', s} [\sup_{f \in \mathcal{H}} \sum_{i=1}^n s_i(f(X_i) - f(X'_i))].
\]

We now have enough randomness to exploit the Sauer-Shelah lemma. If we fix \( X \) and \( X' \), note that the quantities \( f(X_i) - f(X'_i) \) take values in \([-1, 1]\) and collectively can take on at most \( V_{\mathcal{H}}(n)^2 = \mathcal{O}(n^{2d}) \) values. But for fixed \( X, X' \), the quantities \( s_i(f(X_i) - f(X'_i)) \) are independent, zero-mean, bounded random variables and hence for fixed \( f \) we have \( \mathbb{P}[\sum_i s_i(f(X_i) - f(X'_i)) \geq t] \leq \exp(-t^2/9n) \) by Hoeffding’s inequality. Union bounding over the \( \mathcal{O}(n^{2d}) \) effectively distinct \( f \), we obtain
\[
\mathbb{P}[\sup_{f \in \mathcal{H}} \sum_{i=1}^n s_i(f(X_i) - f(X'_i)) \geq t] \leq \mathcal{O}(n^{2d}) \exp(-t^2/9n).
\]

This is small as long as \( t \gg \sqrt{d \log n} \), so (45) is \( \mathcal{O}(\sqrt{d \log n}/n) \), as claimed.

A particular consequence of Proposition 2.24 is the Dvoretzky-Kiefer-Wolfowitz inequality:

**Proposition 2.26 (DKW inequality).** For a distribution \( p \) on \( \mathbb{R} \) and i.i.d. samples \( X_1, \ldots, X_n \sim p \), define the empirical cumulative density function as \( F_n(x) = \frac{1}{n} \sum_{i=1}^n 1[X_i \leq x] \), and the population cumulative density function as \( F(x) = p(X \leq x) \). Then \( \mathbb{P}[\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \geq t] \leq 2e^{-2nt^2} \).

This follows from applying Proposition 2.24 to the family of threshold functions.

[ Lecture 5 ]

### 2.6 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
\[
\theta(\hat{p}) = \theta^*(q), \text{ where } q = \arg \min_{q \in \mathcal{G}} TV(q, \hat{p}).
\]

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 \( TV(p_n, p) = TV(p_n, p'_n) = 1 \) 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.

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

---

\(^2\)We will later study the \( W_1 \) distance, which does give credit for being close.
distance \( \tilde{TV} \) and use the minimum distance functional corresponding to this relaxed distance. To show that projection under \( TV \) still works, we will need to check that the modulus \( m(\mathcal{G}, \epsilon) \) is still small after we replace \( TV \) with \( \tilde{TV} \), and we will also need to check that the distance \( \tilde{TV}(p, p_n) \) between \( p \) and its empirical distribution is small with high probability. We do this below in Section 2.6.1.

An alternative to relaxing the distance from \( TV \) to \( \tilde{TV} \) 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 2.6.2, and study the computational feasibility of projecting onto a set \( \mathcal{M} \) starting in Section 2.7.

### 2.6.1 Relaxing the Distance

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

**Lemma 2.27.** Suppose that \( \tilde{TV} \) is a (pseudo-)metric such that \( \tilde{TV}(p, q) \leq TV(p, q) \) for all \( p, q \). If we assume that \( p^* \in \mathcal{G} \) and \( TV(p^*, \tilde{p}) \leq \epsilon \), then the error of the minimum distance functional (2) with \( D = \tilde{TV} \) is at most \( m(\mathcal{G}, 2\epsilon', \tilde{TV}, L) \), where \( \epsilon' = \epsilon + TV(\tilde{p}, \tilde{p}_n) \).

**Proof.** By Proposition 2.4 we already know that the error is bounded by \( m(\mathcal{G}, 2\tilde{TV}(p^*, \tilde{p}_n), \tilde{TV}, L) \). Since \( \tilde{TV} \) is a pseudometric, by the triangle inequality we have \( \tilde{TV}(p^*, \tilde{p}_n) \leq \tilde{TV}(p^*, \tilde{p}) + TV(\tilde{p}, \tilde{p}_n) \). Finally, \( TV(p^*, \tilde{p}) \leq TV(p^*, \tilde{p}) \) by assumption.

**Lemma 2.27** shows that we need \( \tilde{TV} \) to satisfy two properties: \( \tilde{TV}(\tilde{p}, \tilde{p}_n) \) should be small, and the modulus \( m(\mathcal{G}, \epsilon, \tilde{TV}) \) should not be too much larger than \( m(\mathcal{G}, \epsilon, TV) \).

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

\[
\tilde{TV}_{\mathcal{H}}(p, q) \overset{\text{def}}{=} \sup_{f \in \mathcal{H}, \tau \in \mathbb{R}} |P_{X \sim p}[f(X) \geq \tau] - P_{X \sim q}[f(X) \geq \tau]|. \tag{48}
\]

(Note the similarity to the distance in Proposition 2.24; we will make use of this later.) We will make the particular choice \( \mathcal{H} = \mathcal{H}_{\text{lin}} \), where \( \mathcal{H}_{\text{lin}} \overset{\text{def}}{=} \{ x \mapsto \langle v, x \rangle \mid v \in \mathbb{R}^d \} \).

First note that \( \tilde{TV}_{\mathcal{H}} \) is indeed upper-bounded by \( TV \), since \( TV(p, q) = \sup_E |p(E) - q(E)| \) is the supremum over all events \( E \), and (48) 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:

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{mean_crossing.png}
\caption{Illustration of mean cross lemma. For any distributions \( p_1, p_2 \) that are close under \( \tilde{TV} \), we can truncate the \( \epsilon \)-tails of each distribution to make their means cross.}
\end{figure}
Lemma 2.28. Suppose that $p$ and $q$ are two distributions such that $\tilde{\text{TV}}_H(p, q) \leq \epsilon$. Then for any $f \in H$, there are distributions $r_p \leq \frac{p}{1-\epsilon}$ and $r_q \leq \frac{q}{1-\epsilon}$ such that $E_{X \sim r_p}[f(X)] \geq E_{Y \sim r_q}[f(Y)]$.

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 $\epsilon$ probability mass corresponding to the largest points of $f(X)$ in $p$ to get $r_p$, and delete $\epsilon$ probability mass corresponding to the smallest points $f(Y)$ in $q$ to get $r_q$. Since $\tilde{\text{TV}}_H(p, q)$ we have

$$\sup_{\tau \in \mathbb{R}} |P_{X \sim r_p}(f(X) \geq \tau) - P_{Y \sim r_q}(f(Y) \geq \tau)| \leq \epsilon,$$

which implies that $P_{r_p}(f(X) \geq \tau) \leq P_{r_q}(f(Y) \geq \tau)$ for all $\tau \in \mathbb{R}$. Hence, $r_q$ stochastically dominates $r_p$ and $E_{r_p}[f(X)] \leq E_{r_q}[f(Y)]$. □

Mean crossing lemmas such as Lemma 2.28 help us bound the modulus of relaxed distances for the family of resilient distributions. In this case we have the following corollary:

Corollary 2.29. For the family $\mathcal{G}_{TV}(\rho, \epsilon)$ of $(\rho, \epsilon)$-resilient distributions and $L(p, \theta) = ||\theta - \mu(p)||_2$, we have

$$m(\mathcal{G}_{TV}(\rho, \epsilon), \epsilon, \tilde{\text{TV}}_{H_{lin}}) \leq 2\rho.$$ (50)

Compare to Theorem 2.10 where we showed that $m(\mathcal{G}_{TV}, \epsilon, \text{TV}) \leq \rho$. Thus as long as Theorem 2.10 is tight, relaxing from TV to $\tilde{\text{TV}}_{H_{lin}}$ doesn’t increase the modulus at all!

Proof of Corollary 2.29. Let $p, q \in \mathcal{G}_{TV}$ such that $\tilde{\text{TV}}(p, q) \leq \epsilon$. Take $v = \arg \max ||v||_2 = 1 v^T(E_p[X] - E_q[X])$, hence $E_p[v^T X] - E_q[v^T X] = ||E_p[X] - E_q[X]||_2$. It follows from Lemma 2.28 that there exist $r_p \leq \frac{p}{1-\epsilon}, r_q \leq \frac{q}{1-\epsilon}$ such that

$$E_{r_p}[v^T X] \leq E_{r_q}[v^T X].$$ (51)

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

$$E_p[v^T X] - E_{r_p}[v^T X] \leq \rho,$$ (52)

$$E_{r_q}[v^T X] - E_q[v^T X] \leq \rho.$$ (53)

Then,

$$||E_p[X] - E_q[X]||_2 = E_p[v^T X] - E_{r_p}[v^T X] \leq E_p[v^T X] - E_q[v^T X] \leq E_{r_p}[v^T X] + E_{r_q}[v^T X] - E_q[v^T X] \leq 2\rho,$$ (56)

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 $\tilde{\text{TV}}(\hat{p}, \hat{p}_n)$. Note that $\tilde{\text{TV}}(\hat{p}, \hat{p}_n)$ is exactly the quantity bounded in equation (42) of Proposition 2.24; we thus have that $\tilde{\text{TV}}_H(\hat{p}, \hat{p}_n) \leq O(\sqrt{\text{vc}(\mathcal{H}) + \log(1/\delta)})$ with probability $1 - \delta$. Here $\text{vc}(\mathcal{H})$ is the VC dimension of the family of threshold functions \( \{x \mapsto \text{I}[f(x) \geq \tau] \mid f \in \mathcal{H}, \tau \in \mathbb{R} \} \). So, for $\mathcal{H} = H_{lin}$ 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

$$\sum_{i=1}^{d+2} a_i x_i = 0, \sum_{i=1}^{d+2} a_i = 0.$$ (57)
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 $\tau$ such that $(x_i, v) \geq \tau$ 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 $\text{conv}(S_+) \cap \text{conv}(S_-) \neq \emptyset$.)

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

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

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

This shows that $x_1, \ldots, x_{d+2}$ cannot be shattered, so $\text{vc}(H_{\text{lin}}) \leq d+1$. Combining this with Proposition 2.24, we obtain:

**Proposition 2.30.** With probability $1 - \delta$, we have $\widetilde{TV}_{H_{\text{lin}}} (\tilde{p}, \tilde{p}_n) \leq O\left(\sqrt{\frac{d + \log(1/\delta)}{n}}\right)$.

Combining this with Corollary 2.29 and Lemma 2.27, we see that projecting onto $G_{\text{TV}} (\rho, 2\epsilon')$ under $\widetilde{TV}_{H_{\text{lin}}}$ performs well in finite samples, for $\epsilon' = \epsilon + O(\sqrt{d/n})$. For instance, if $G$ has bounded covariance we achieve error $O(\sqrt{\epsilon + \sqrt{d/n}})$; if $G$ is sub-Gaussian we achieve error $O(\epsilon + \sqrt{d/n})$; and in general if $G$ has bounded $\psi$-norm we achieve error $O\left((\epsilon + \sqrt{d/n}) \psi^{-1}\left(\frac{1}{\epsilon + \sqrt{d/n}}\right)\right) \leq O\left(\epsilon \psi^{-1}(1/\epsilon)\right)$.

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 $\epsilon$ 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

$$
\min_{\mu} \max_{v \in \mathbb{R}^d} \mathbb{P}_{X \sim \tilde{p}_n} \left[ (X, v) \geq (\mu, v) \right] - \frac{1}{2}.
$$

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

$$
\min_{q \in G} \max_{v \in \mathbb{R}^d, \tau \in \mathbb{R}} \mathbb{P}_{X \sim \tilde{p}_n} \left[ (X, v) \geq \tau \right] - \mathbb{P}_{X \sim q} \left[ (X, v) \geq \tau \right].
$$

The differences are: (1) the Tukey median only minimizes over the mean rather than the full distribution $q$; (2) it only considers the threshold $(\mu, v)$ rather than all thresholds $\tau$; it assumes that the median of any one-dimensional projection $(X, v)$ is equal to its mean (which is why we subtract $\frac{1}{2}$ in (59)). 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.

[Lectures 6-7]

### 2.6.2 Expanding the Set

In Section 2.6.1 we saw how to resolve the issue with $\text{TV}$ projection by relaxing to a weaker distance $\widetilde{TV}$. We will now study an alternate approach, based on expanding the destination set $G$ to a larger set $\mathcal{M}$. For this approach we will need to reference the “true empirical distribution” $p_n^\ast$. What we mean by this is the following: Whenever $\text{TV}(p^\ast, \tilde{p}) \leq \epsilon$, we know that $p^\ast$ and $\tilde{p}$ are identical except for some event $E$ of probability $\epsilon$. Therefore we can sample from $\tilde{p}$ as follows:

1. Draw a sample from $X \sim p^\ast$.
2. Check if $E$ holds; if it does, replace $X$ with a sample from the conditional distribution $\tilde{p}_{|E}$. 
3. Otherwise leave $X$ as-is.

Thus we can interpret a sample from $\hat{p}$ as having a $1 - \epsilon$ chance of being “from” $p^\ast$. More generally, we can construct the empirical distribution $\hat{p}_n$ by first constructing the empirical distribution $p_n^\ast$ coming from $p^\ast$, then replacing $\text{Binom}(n, \epsilon)$ of the points with samples from $\hat{p}_E$. Formally, we have created a coupling between the random variables $p_n^\ast$ and $\hat{p}_n$ such that $\text{TV}(p_n^\ast, \hat{p}_n)$ is distributed as $\frac{1}{n}\text{Binom}(n, \epsilon)$.

Let us return to expanding the set from $G$ to $M$. For this to work, we need three properties to hold:

- $M$ is large enough: $\min_{q \in M} \text{TV}(q, p_n^\ast)$ is small with high probability.
- The empirical loss $L(p_n^\ast, \theta)$ is a good approximation to the population loss $L(p^\ast, \theta)$.
- The modulus is still bounded: $\min_{p, q \in M: \text{TV}(p, q) \leq 2\epsilon} L(p, p^\ast(q))$ is small.

In fact, it suffices for $M$ to satisfy a weaker property; we only need the “generalized modulus” to be small relative to some $G^\prime \subset M$:

**Proposition 2.31.** For a set $G^\prime \subset M$, define the generalized modulus of continuity as

$$m(G^\prime, M, 2\epsilon) \overset{\text{def}}{=} \min_{p \in G^\prime, q \in M: \text{TV}(p, q) \leq 2\epsilon} L(p, p^\ast(q)).$$

Assume that the true empirical distribution $p_n^\ast$ lies in $G^\prime$ with probability $1 - \delta$. Then the minimum distance functional projecting under $\text{TV}$ onto $M$ has empirical error $L(p_n^\ast, \hat{\theta})$ at most $m(G^\prime, M, 2\epsilon)$ with probability at least $1 - \delta - \Pr[\text{Binom}(\epsilon, n) \geq \epsilon^\prime n]$.

**Proof.** Let $\epsilon^\prime = \text{TV}(p_n^\ast, \hat{p}_n)$, which is $\text{Binom}(\epsilon, n)$-distributed. If $p_n^\ast$ lies in $G^\prime$, then since $G^\prime \subset M$ we know that $\hat{p}_n$ has distance at most $\epsilon^\prime$ from $M$, and so the projected distribution $q$ satisfies $\text{TV}(q, p_n^\ast) \leq \epsilon^\prime$ and hence $\text{TV}(q, p_n^\ast) \leq 2\epsilon^\prime$. It follows from the definition that $L(p_n^\ast, \hat{\theta}) = L(p_n^\ast, p^\ast(q)) \leq m(G^\prime, M, 2\epsilon^\prime)$. \qed

A useful bound on the binomial tail is that $\Pr[\text{Binom}(\epsilon, n) \geq 2\epsilon n] \leq \exp(-\epsilon n / 3)$. In particular the empirical error is at most $m(G^\prime, M, 4\epsilon)$ with probability at least $1 - \delta - \exp(-\epsilon n / 3)$.

**Application: bounded $k$th moments.** First suppose that the distribution $p^\ast$ has bounded $k$th moments, i.e. $G_{\text{mom}, k}(\sigma) = \{ p : \| p \| \psi \leq \sigma \}$, where $\psi(x) = x^k$. When $k > 2$, the empirical distribution $p_n^\ast$ will not have bounded $k$th moments until $n \geq \Omega(d^{k/2})$. This is because if we take a single sample $x \sim p$ and let $v$ be a unit vector in the direction of $x_1 - \mu$, then $\mathbb{E}_{x \sim p_n^\ast} [(x - \mu, v)^k] \geq \frac{1}{n} \| x_1 - \mu \|^k \geq (d^{k/2} / n)$, since the norm of $\| x - \mu \|^2$ is typically $\sqrt{d}$.

Consequently, it is necessary to expand the set and we will choose $G' = M = G_{\text{TV}}(\rho, \epsilon)$ to be the set of resilience distributions with appropriate parameters $\rho$ and $\epsilon$. We already know that the modulus of $M$ is bounded by $O(\sigma \epsilon^{1-1/k})$, so the hard part is showing that the empirical distribution $p_n^\ast$ lies in $M$ with high probability.

As noted above, we cannot hope to prove that $p_n^\ast$ has bounded moments except when $n = \Omega(d^{k/2})$, which is too large. We will instead show that certain truncated moments of $p_n^\ast$ are bounded as soon as $n = \Omega(d)$, and that these truncated moments suffice to show resilience. Specifically, if $\psi(x) = x^k$ is the Orlicz function for the $k$th moments, we will define the truncated function

$$\tilde{\psi}(x) = \begin{cases} x^k & : x \leq x_0 \\ kx_0^{k-1}(x - x_0) + x_0^k & : x > x_0 \end{cases}$$

In other words, $\tilde{\psi}$ is equal to $\psi$ for $x \leq x_0$, and is the best linear lower bound to $\psi$ for $x > x_0$. Note that $\tilde{\psi}$ is $L$-Lipschitz for $L = kx_0^{k-1}$. We will eventually take $x_0 = (1/\epsilon)^{1/k}$ and hence $L = k(1/\epsilon)^{k/(k-1)}$. Using a symmetrization argument, we will bound the truncated supremum $\sup_{\| v \| \leq \sigma} \mathbb{E}_{p_n^\ast}[\tilde{\psi}((x - \mu, v) / \sigma)]$.

**Proposition 2.32.** Let $X_1, \ldots, X_n \sim p^\ast$, where $p^\ast \in G_{\text{mom}, k}(\sigma)$. Then,

$$\mathbb{E}_{X_1, \ldots, X_n \sim p^\ast} \left[ \sup_{\| v \| \leq \sigma} \frac{1}{n} \sum_{i=1}^n \tilde{\psi} \left( \frac{|(X_i - \mu, v)|}{\sigma} \right) \right] \leq 1 + O \left( 2L \sqrt{\frac{dk}{n}} \right),$$

where $L = kx_0^{k-1}$. 

19
Before proving Proposition 2.32, let us interpret its significance. Take $x_0 = (1/\epsilon)^{1/k}$ and hence $L = k\epsilon^{1-1/k}$. Take $n$ large enough so that the second term in the right-hand-side of (63) is at most 1, which requires $n \geq \Omega(k^3 d/\epsilon^2 - 2/k)$. We then obtain a bound on the $\psi$-norm of $p_n^*$, which implies that $p_n^*$ is resilient with parameter $\rho = \sigma \psi^{-1}(2/\epsilon) = 2\sigma \epsilon^{1-1/k}$. This matches the population-bound of $O(\sigma \epsilon^{1-1/k})$, and only requires $d/\epsilon^2 - 2/k$ samples, in contrast to the $d/\epsilon^2$ samples required before. Indeed, this sample complexity dependence is optimal; the only drawback is that we do not get exponential tails (we can show tails of $\delta^{-1/k}$ through more careful analysis, but this is worse than the $\sqrt{\log(1/\delta)}$ from before).

We would like to use the fact that $\tilde{\psi}$ is $L$-Lipschitz to replace the expression $\psi(|(X - \mu, v)|/\sigma)$ in (81) with the simpler expression $L(X - \mu, v)/\sigma$. We can do so with the following proposition:

**Theorem 2.33 (Ledoux-Talagrand Contraction).** Let $\phi : \mathbb{R} \to \mathbb{R}$ be an $L$-Lipschitz function such that $\phi(0) = 0$. Then for any convex, increasing function $g$ and Rademacher variables $\epsilon_{1,n} \sim \{\pm 1\}$, we have

$$\mathbb{E}_{\epsilon_{1,n}}[g(\sup_{t \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \phi(t_i))] \leq \mathbb{E}_{\epsilon_{1,n}}[g(L \sup_{t \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i t_i)].$$

Let us interpret this result. We should think of the $t_i$ as a quantity such as $(x_i - \mu, v)$, where abstracting to $t_i$ yields generality and notational simplicity. Theorem 2.33 says that if we let $Y = \sup_{t \in \mathcal{T}} \sum_{i=1}^{n} \epsilon_i \phi(t_i)$ and $Z = L \sup_{t \in \mathcal{T}} \sum_{i=1}^{n} \epsilon_i t_i$, then $\mathbb{E}[g(Y)] \leq \mathbb{E}[g(Z)]$ for all convex increasing functions $g$. When this holds we say that $Y$ stochastically dominates $Z$ in second order; intuitively, it is equivalent to saying that $Z$ has larger mean than $Y$ and greater variation around its mean. For distributions supported on just two points, we can formalize this as follows:

**Lemma 2.34 (Two-point stochastic dominance).** Let $Y$ take values $y_1$ and $y_2$ with probability $\frac{1}{2}$, and $Z$ take values $z_1$ and $z_2$ with probability $\frac{1}{2}$. Then $Z$ stochastically dominates $Y$ (in second order) if and only if

$$\frac{z_1 + z_2}{2} \geq \frac{y_1 + y_2}{2} \text{ and } \max(z_1, z_2) \geq \max(y_1, y_2).$$

**Proof.** Without loss of generality assume $z_2 \geq z_1$ and $y_2 \geq y_1$. We want to show that $\mathbb{E}[g(Y)] \leq \mathbb{E}[g(Z)]$ for all convex increasing $g$ if and only if (65) holds. We first establish necessity of (65). Take $g(x) = x$, then we require $\mathbb{E}[Y] \leq \mathbb{E}[Z]$, which is the first condition in (65). Taking $g(x) = \max(x - z_2, 0)$ yields $\mathbb{E}[g(Y)] = 0$ and $\mathbb{E}[g(Y)] \geq \frac{1}{2} \max(y_2 - z_2, 0)$, so $\mathbb{E}[g(Y)] \leq \mathbb{E}[g(Z)]$ implies that $y_2 \leq z_2$, which is the second condition in (65).

We next establish sufficiency, by conjuring up appropriate weights for Jensen’s inequality. We have

$$\frac{y_2 - z_1}{z_2 - z_1} g(z_2) + \frac{z_2 - y_2}{z_2 - z_1} g(z_1) \geq g\left(\frac{z_2(y_2 - z_1) + z_1(y_2 - y_1)}{z_2 - z_1}\right) = g(y_2),$$

$$\frac{z_2 - y_2}{z_2 - z_1} g(z_2) + \frac{y_2 - z_1}{z_2 - z_1} g(z_1) \geq g\left(\frac{z_2(z_2 - y_2) + z_1(z_2 - z_1)}{z_2 - z_1}\right) = g(z_1) + z_2 - y_2 \geq g(y_1).$$

Here the first two inequalities are Jensen while the last is by the first condition in (??) together with the monotonicity of $g$. Adding these together yields $g(z_2) + g(z_1) \geq g(y_2) + g(y_1)$, or $\mathbb{E}[g(Z)] \geq \mathbb{E}[g(Y)]$, as desired. We need only check that the weights $\frac{y_2 - z_1}{z_2 - z_1}$ and $\frac{z_2 - y_2}{z_2 - z_1}$ are positive. The second weight is positive by the assumption $z_2 \geq y_2$. The first weight could be negative if $y_2 < z_1$, meaning that both $y_1$ and $y_2$ are smaller than both $z_1$ and $z_2$. But in this case, the inequality $\mathbb{E}[g(Y)] \leq \mathbb{E}[g(Z)]$ trivially holds by monotonicity of $g$. This completes the proof.

We are now ready to prove Theorem 2.33.

**Proof of Theorem 2.33.** Without loss of generality we may take $L = 1$. Our strategy will be to iteratively apply an inequality for a single $\epsilon_i$ to replace all the $\phi(t_i)$ with $t_i$ one-by-one. The inequality for a single $\epsilon_i$ is the following:
Lemma 2.35. For any 1-Lipschitz function $\phi$ with $\phi(0) = 0$, any collection $T$ of ordered pairs $(a,b)$, and any convex increasing function $g$, we have

$$E_{\epsilon \sim \{-1,+1\}}[g(\sup_{(a,b) \in T} a + \epsilon \phi(b))] \leq E_{\epsilon \sim \{-1,+1\}}[g(\sup_{(a,b) \in T} a + \epsilon b)].$$

To prove this, let $(a_+, b_+)$ attain the sup of $a + \epsilon \phi(b)$ for $\epsilon = +1$, and $(a_-, b_-)$ attain the sup for $\epsilon = -1$. We will check the conditions of Lemma 2.34 for

$$y_1 = a_+ - \phi(b_-),$$
$$y_2 = a_+ + \phi(b_+),$$
$$z_1 = \max(a_- - b_-, a_+ - b_+),$$
$$z_2 = \max(a_- + b_-, a_+ + b_+).$$

(Note that $z_1$ and $z_2$ are lower-bounds on the right-hand-side sup for $\epsilon = -1, +1$ respectively.)

First we need $\max(y_1, y_2) \leq \max(z_1, z_2)$. But $\max(z_1, z_2) = \max(a_- + |b_-, a_+ + |b_-|) \geq \max(a_- - \phi(b_-), a_+ + \phi(b_+)) = \max(y_1, y_2)$. Here the inequality follows since $\phi(b) \leq |b|$ since $\phi$ is Lipschitz and $\phi(0) = 0$.

Second we need $\frac{y_1+y_2}{2} \leq \frac{z_1+z_2}{2}$. We have $z_1 + z_2 \geq \max((a_- - b_-) + (a_+ + b_+), (a_- + b_-) + (a_+ - b_-)) = a_+ + a_- + |b_+ - b_-|$, so it suffices to show that $\frac{a_+ + a_- + |b_+ - b_-|}{2} \leq \frac{a_+ + a_- + \phi(b_+) - \phi(b_-)}{2}$. This exactly reduces to $\phi(b_+) - \phi(b_-) \leq |b_+ - b_-|$, which again follows since $\phi$ is Lipschitz. This completes the proof of the lemma.

Now to prove the general proposition we observe that if $g(x)$ is convex in $x$, so is $g(x+t)$ for any $t$. We then proceed by iteratively applying Lemma 2.35:

$$E_{\epsilon_1,n}[g(\sup_{t \in T} \sum_{i=1}^n \epsilon_i \phi(t_i))] = E_{\epsilon_1,n-1}[E_{\epsilon_n}[g(\sup_{t \in T} \sum_{i=1}^{n-1} \epsilon_i \phi(t_i) + \epsilon_n \phi(t_n)) | \epsilon_1:n-1]]$$

$$\leq E_{\epsilon_1,n-1}[E_{\epsilon_n}[g(\sup_{t \in T} \sum_{i=1}^{n-1} \epsilon_i \phi(t_i) + \epsilon_n t_n) | \epsilon_1:n-1]]$$

$$= E_{\epsilon_1,n}[g(\sup_{t \in T} \sum_{i=1}^{n-1} \epsilon_i \phi(t_i) + \epsilon_n t_n)]$$

$$\vdots$$

$$\leq E_{\epsilon_1,n}[g(\sup_{t \in T} \sum_{i=2}^n \epsilon_i t_i)]$$

$$\leq E_{\epsilon_1,n}[g(\sup_{t \in T} \sum_{i=1}^n \epsilon_i t_i)],$$

which completes the proof.

Let us return now to bounding the truncated moments in Proposition 2.32.

Proof of Proposition 2.32. We start with a symmetrization argument. Let $\mu_\psi = E_{X \sim p^\ast}[\tilde{\psi}(\langle |X - \mu|/\sigma \rangle)]$, and note that $\mu_\psi \leq \mu_\tilde{\psi} \leq 1$. Now, by symmetrization we have

$$E_{X_1,\ldots,X_n \sim p^\ast} \left[ \sup_{\|v\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^n \tilde{\psi}\left(\frac{|X_i - \mu|}{\sigma}\right) - \mu_\tilde{\psi}\right]^k \leq E_{X,X' \sim p^\ast} \left[ \sup_{\|v\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^n \epsilon_i \tilde{\psi}\left(\frac{|X_i - \mu|}{\sigma}\right) - \mu_\tilde{\psi}\left(\frac{|X_i - \mu|}{\sigma}\right)\right]^k \leq 2^k E_{X \sim p, \epsilon} \left[ \sup_{\|v\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^n \epsilon_i \tilde{\psi}\left(\frac{|X_i - \mu|}{\sigma}\right)\right]^k.$$


Here the first inequality adds and subtracts the mean, the second applies symmetrization, while the third uses the fact that optimizing a single \( v \) for both \( X \) and \( X' \) is smaller than optimizing \( v \) separately for each (and that the expectations of the expressions with \( X' \) are equal to each other in that case).

We now apply Ledoux-Talagrand contraction. Invoking Theorem 2.33 with \( g(x) = |x|^k \), \( \phi(x) = \tilde{\psi}(|x|) \) and \( t_i = \langle X_i - \mu, v \rangle / \sigma \), we obtain

\[
E_{X \sim p, \epsilon} \left[ \sup_{||v|| \leq 1} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \tilde{\psi} \left( \frac{\langle X_i - \mu, v \rangle}{\sigma} \right) \right]^k \leq \left( \frac{L}{\sigma} \right)^k E_{X \sim p, \epsilon} \left[ \sup_{||v|| \leq 1} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \langle X_i - \mu, v \rangle \right]^k \leq \left( \frac{L}{\sigma} \right)^k E_{X \sim p, \epsilon} \left[ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \langle X_i - \mu \rangle \right]^k,
\]

(82)

We are thus finally left to bound \( E_{X \sim p, \epsilon} [\| \sum_{i=1}^{n} \epsilon_i (X_i - \mu) \|^k] \). Here we will use Khintchine’s inequality, which says that

\[
A_k \| z \|_2 \leq E_{\epsilon} [\| \sum_i \epsilon_i z_i \|^{k/2}] \leq B_k \| z \|_2,
\]

(84)

where \( A_k \) is \( \Theta(1) \) and \( B_k \) is \( \Theta(\sqrt{k}) \) for \( k \geq 1 \). Applying this in our case, we obtain

\[
E_{X, \epsilon} [\| \sum_{i=1}^{n} \epsilon_i (X_i - \mu) \|^k] \leq O(1)^k E_{X, \epsilon, \epsilon'} [\| \sum_{i=1}^{n} \epsilon_i (X_i - \mu, \epsilon') \|^k] \leq O(\sqrt{k})^k E_{X, \epsilon} [\| \sum_{i=1}^{n} (X_i - \mu, \epsilon')^2 \|^k/2].
\]

(85)

(86)

Here the last inequality applies Khintchine to the \( \epsilon_i \) to replace them with the norm of the vector with ith entry \( \langle X_i - \mu, \epsilon' \rangle \). Now, assuming \( k \) is even, the term \( \| \sum_{i=1} \langle X_i - \mu, \epsilon' \rangle^2 \|^{k/2} \) expands out to \( n^k/2 \) distinct terms, each of which is of the form \( E_{X, \epsilon} [\prod_{i=1}^{k/2} (X_i - \mu, \epsilon')] \). Conditioning on \( \epsilon' \) and taking expectation over the \( X_i \), each term is bounded by \( \sigma^k \| \epsilon' \|_2^k \), since the \( X_i \) each have bounded \( k \)th moment and are independent. Additionally, we have \( \| \epsilon' \|_2^2 = d^{k/2} \) since \( \epsilon' \) is a sign vector. Putting this together and plugging back into the inequalities above yields

\[
E_{X, \epsilon} [\| \sum_{i=1}^{n} \epsilon_i (X_i - \mu) \|^k] \leq O(\sqrt{k})^k \cdot \sigma^k n^{k/2} d^{k/2}.
\]

(87)

Plugging back into (??) bounds the symmetrized truncated moments by \( O(\sqrt{kd/nL})^k \), and plugging back into (81) completes the proof.

**Application: isotropic Gaussians.** Next take \( G_{\text{gauss}} \) to be the family of isotropic Gaussians \( N(\mu, I) \). We saw earlier that the modulus \( m(G_{\text{gauss}}, \epsilon) \) was \( O(\epsilon) \) for the mean estimation loss \( L(p, \theta) = \| \theta - \mu(p) \|_2 \). Thus projecting onto \( G_{\text{gauss}} \) yields error \( O(\epsilon) \) for mean estimation in the limit of infinite samples, but doesn’t work for finite samples since the \( TV \) distance to \( G_{\text{gauss}} \) will always be 1.

Instead we will project onto the set \( G_{\text{cov}}(\sigma) = \{ p \mid \| E[ (X - \mu)(X - \mu)^\top) ] \| \leq \sigma^2 \} \), for \( \sigma^2 = O(1 + d/n + \log(1/\delta)/n) \). We already saw in Lemma 2.23 that when \( p^* \) is (sub-)Gaussian the empirical distribution \( p_n^* \) lies within this set. But the modulus of \( G_{\text{cov}} \) only decays as \( O(\sqrt{\epsilon}) \), which is worse than the \( O(\epsilon) \) dependence that we had in infinite samples! How can we resolve this issue?

We will let \( G_{\text{iso}} \) be the family of distributions whose covariance is not only bounded, but close to the identity, and where moreover this holds for all \( (1 - \epsilon) \)-subsets:

\[
G_{\text{iso}}(\sigma_1, \sigma_2) \overset{\text{def}}{=} \{ p \mid \| E_r [X - \mu] \|_2 \leq \sigma_1 \text{ and } \| E_r [(X - \mu)(X - \mu)^\top - I] \| \leq (\sigma_2)^2, \text{ whenever } r \leq \frac{p}{1 - \epsilon} \}.
\]

(88)

The following improvement on Lemma 2.23 implies that \( p_n^* \in G_{\text{iso}}(\sigma_1, \sigma_2) \) for \( \sigma_1 = O(\epsilon \sqrt{\log(1/\epsilon)}) \) and \( \sigma_2 = O(\sqrt{\epsilon \log(1/\epsilon)}) \).
Lemma 2.36. Suppose that $X_1, \ldots, X_n$ are drawn independently from a sub-Gaussian distribution with sub-Gaussian parameter $\sigma$, mean 0, and identity covariance. Then, with probability $1 - \delta$ we have

$$
\left\| \frac{1}{|S|} \sum_{i \in S} X_i X_i^\top - I \right\| \leq O\left( \sigma^2 \left( \epsilon \log(1/\epsilon) + \frac{d + \log(1/\delta)}{n} \right) \right),
$$

and

$$
\left\| \frac{1}{|S|} \sum_{i \in S} X_i \right\|_2 \leq O\left( \sigma \left( \epsilon \sqrt{\log(1/\epsilon)} + \sqrt{\frac{d + \log(1/\delta)}{n}} \right) \right)
$$

for all subsets $S \subseteq \{1, \ldots, n\}$ with $|S| \geq (1 - \epsilon)n$. In particular, if $n \gg d/(\epsilon^2 \log(1/\epsilon))$ then $\delta \leq \exp(-c n \log(1/\epsilon))$ for some constant $c$.

We will return to the proof of Lemma 2.36 later. For now, note that this means that $p_n^* \in G'$ for $G' = G_{\text{iso}}(O(\epsilon \sqrt{\log(1/\epsilon)}), O(\sqrt{\epsilon \log(1/\epsilon)}))$, at least for large enough $n$. Furthermore, $G' \subseteq M$ for $M = G_{\text{cov}}(1 + O(\epsilon \log(1/\epsilon)))$.

Now we bound the generalized modulus of continuity:

Lemma 2.37. Suppose that $p \in G_{\text{iso}}(\sigma_1, \sigma_2)$ and $q \in G_{\text{cov}}(\sqrt{1 + \sigma_2^2})$, and furthermore $\text{TV}(p, q) \leq \epsilon$. Then $\|\mu(p) - \mu(q)\|_2 \leq O(\sigma_1 + \sigma_2 \sqrt{\epsilon} + \epsilon)$.

Proof. Take the midpoint distribution $r = \frac{\min(p,q)}{1+\epsilon}$, and write $q = (1-\epsilon)r + \epsilon q'$. We will bound $\|\mu(r) - \mu(q)\|_2$ (note that $\|\mu(r) - \mu(q)\|_2$ is already bounded since $p \in G_{\text{iso}}$). We have that

$$
\text{Cov}_q[X] = (1-\epsilon)\text{E}_r[(X - \mu_q)(X - \mu_q)^\top] + \epsilon \text{E}_{q'}[(X - \mu_q)(X - \mu_q)^\top]
$$

$$
= (1-\epsilon)(\text{Cov}_r[X] + (\mu_q - \mu_r)(\mu_q - \mu_r)^\top) + \epsilon \text{E}_{q'}[(X - \mu_q)(X - \mu_q)^\top]
$$

$$
\geq (1-\epsilon)(\text{Cov}_r[X] + (\mu_q - \mu_r)(\mu_q - \mu_r)^\top) + \epsilon (\mu_q - \mu_{q'})(\mu_q - \mu_{q'})^\top.
$$

A computation yields

$$
\mu_q - \mu_{q'} = \frac{(1-\epsilon)^2}{\epsilon} (\mu_q - \mu_r).
$$

Plugging this into (93) and simplifying, we obtain that

$$
\text{Cov}_q[X] \geq (1-\epsilon)(\text{Cov}_r[X] + (1/\epsilon)(\mu_q - \mu_r)(\mu_q - \mu_r)^\top).
$$

Now since $\text{Cov}_r[X] \geq (1 - \sigma_2^2) I$, we have $\|\text{Cov}_q[X]\| \geq (1-\epsilon)(1 - \sigma_2^2) + (1/\epsilon)\|\mu_q - \mu_r\|_2^2$. But by assumption $\|\text{Cov}_q[X]\| \leq 1 + \sigma_2^2$. Combining these yields that $\|\mu_r - \mu_q\|_2^2 \leq \epsilon(2\sigma_2^2 + \epsilon + c\sigma_2^2)$, and so $\|\mu_r - \mu_q\|_2 \leq O(\epsilon + \sigma_2 \sqrt{\epsilon})$, which gives the desired result.

In conclusion, projecting onto $G_{\text{cov}}(1 + O(\epsilon \log(1/\epsilon)))$ under TV distance gives a robust mean estimator for isotropic Gaussians, which achieves error $O(\epsilon \sqrt{\log(1/\epsilon)})$. This is slightly worse than the optimal $O(\epsilon)$ bound but improves over the naïve analysis that only gave $O(\sqrt{\epsilon})$.

Another advantage of projecting onto $G_{\text{cov}}$ is that, as we will see in Section 2.7, this projection can be done computationally efficiently.

Proof of Lemma 2.36. TBD

[ Lecture 8 ]

2.7 Efficient Algorithms

We now turn our attention to efficient algorithms. Recall that previously we considered minimum distance functionals projecting onto sets $\mathcal{G}$ and $\mathcal{M}$ under distances $\text{TV}$ and $\text{TV}$. Here we will consider how to approximately project onto the set $\mathcal{G}_{\text{cov}}(\sigma_1, \sigma_2)$, the family of bounded covariance distributions, under $\text{TV}$ distance. The basic idea is that if the true distribution $p^*$ has bounded covariance, and $\tilde{p}$ does not, the largest eigenvector of $\text{Cov}_p[X]$ must be well-aligned with the mean of the bad points, and thus we can use this to remove the bad points. If on the other hand $\tilde{p}$ has bounded covariance, then its mean must be close to $p^*$ by our previous modulus bounds and so we are already done.
To study efficient computation we need a way of representing the distributions $\hat{p}$ and $p^*$. To do this we will suppose that $\hat{p}$ is the empirical distribution over $n$ points $x_1, \ldots, x_n$, while $p^*$ is the empirical distribution over some subset $S$ of these points with $|S| \geq (1 - \epsilon)n$. Thus in particular $p^*$ is an $\epsilon$-deletion of $\hat{p}$.

Before we assumed that $\text{TV}(p^*, \hat{p}) \leq \epsilon$, but taking $p' = \min(p^*, \hat{p})$, we have $p' \leq \frac{\hat{p}}{1-\epsilon}$ and $\|\text{Cov}_{p'}[X]\| \leq \frac{\sigma^2}{\epsilon^2} \leq 2\sigma^2$ whenever $\|\text{Cov}_{\hat{p}}[X]\| \leq \sigma^2$. Therefore, taking $p^* \leq \frac{\hat{p}}{1-\epsilon}$ is equivalent to the TV corruption model from before for our present purposes.

We will construct an efficient algorithm that, given $\hat{p}$, outputs a distribution $q$ such that $\text{TV}(q, p^*) \leq O(\epsilon)$ and $\|\text{Cov}_q[X]\|_2 \leq O(\sigma^2)$. This is similar to the minimum distance functional, in that it finds a distribution close to $p^*$ with bounded covariance; the main difference is that $q$ need not be the projection of $\hat{p}$ onto $G_{\text{cov}}$, and also the covariance of $q$ is bounded by $O(\sigma^2)$ instead of $\sigma^2$. However, the modulus of continuity bound from before says that any distribution $q$ that is near $p^*$ and has bounded covariance will approximate the mean of $p^*$. Specifically, we have

$$\| \mu(q) - \mu(p^*) \|^2_2 \leq O(\max(\|\text{Cov}_q[X]\|, \|\text{Cov}_{p^*}[X]\|)) \cdot \text{TV}(p^*, q) = O(\sigma^2 \epsilon).$$

(95)

We will show the following:

**Proposition 2.38.** Suppose $\hat{p}$ and $p^*$ are empirical distributions as above with $p^* \leq \hat{p}/(1 - \epsilon)$, and further suppose that $\|\text{Cov}_{p^*}[X]\| \leq \sigma^2$. Then given $\hat{p}$ (but not $p^*$), there is an algorithm with runtime $\text{poly}(n, d)$ that outputs a $q$ with $\text{TV}(p^*, q) \leq \epsilon$ and $\|\text{Cov}_q[X]\| \leq O(\sigma^2)$. In particular, $\|\mu(p^*) - \mu(q)\|_2 = O(\sigma \sqrt{\epsilon})$.

Note that the conclusion $\|\mu(p^*) - \mu(q)\|_2 \leq O(\sigma \sqrt{\epsilon})$ follows from the modulus bound on $G_{\text{cov}}(\sigma)$ together with the property $\text{TV}(p^*, q) \leq \epsilon$.

The algorithm, FilterL2, underlying Proposition 2.38 is given below; it maintains a weighted distribution $q(c)$, which places weight $c_i/\sum_{j=1}^n c_j$ on point $x_i$. It then computes the weighted mean and covariance, projects onto the top eigenvector, and downweights points with large projection.

### Algorithm 2 FilterL2

1. Input: $x_1, \ldots, x_n \in \mathbb{R}^d$.
2. Initialize weights $c_1, \ldots, c_n = 1$.
3. Compute the empirical mean $\hat{\mu}_c$ of the data, $\hat{\mu}_c \overset{\text{def}}{=} (\sum_{i=1}^n c_i x_i)/(\sum_{i=1}^n c_i)$.
4. Compute the empirical covariance $\hat{\Sigma}_c \overset{\text{def}}{=} \sum_{i=1}^n c_i (x_i - \hat{\mu}_c) (x_i - \hat{\mu}_c)^\top / \sum_{i=1}^n c_i$.
5. Let $v$ be the maximum eigenvector of $\hat{\Sigma}_c$, and let $\hat{\sigma}^2_c = v^\top \hat{\Sigma}_c v$.
6. If $\hat{\sigma}^2_c \leq 2\sigma^2$, output $q(c)$.
7. Otherwise, let $\tau_i = \langle x_i - \hat{\mu}_c, v \rangle^2$, and update $c_i \leftarrow c_i \cdot (1 - \tau_i/\tau_{\text{max}})$, where $\tau_{\text{max}} = \max_i \tau_i$.
8. Go back to line 3.

The factor $\tau_{\text{max}}$ in the update $c_i \leftarrow c_i \cdot (1 - \tau_i/\tau_{\text{max}})$ is so that the weights remain positive; the specific factor is unimportant and the main property required is that each point is downweighted proportionally to $\tau_i$. Note also that Algorithm 2 must eventually terminate because one additional weight $c_i$ is set to zero in every iteration of the algorithm.

The intuition behind Algorithm 2 is as follows: if the empirical variance $\hat{\sigma}^2_c$ is much larger than the variance $\sigma^2$ of the good data, then the bad points must on average be very far away from the empirical mean (i.e., $\tau_i$ must be large on average for the bad points).

More specifically, note that $\tau_i = \langle x_i - \hat{\mu}_c, v^* \rangle^2$. Let $\bar{\tau}_i = \langle x_i - \mu, v^* \rangle^2$, and imagine for now that $\tau_i \approx \bar{\tau}_i$. We know that the average of $\bar{\tau}$ over the good points is at most $\sigma^2$, since $\bar{\tau}$ is the variance along the projection $v^*$ and $\|\text{Cov}_{p^*}[X]\| \leq \sigma^2$. Thus if the overall average of the $\tau_i$ is large (say $20\sigma^2$), it must be on account of the bad points. But since there are not that many bad points, their average must be quite large—on the order of $\sigma^2/\epsilon$. Thus they should be easy to separate from the good points. This is depicted in Figure 7.

This is the basic idea behind the proof, but there are a couple issues with this:

- The assumption that $\bar{\tau}_i \approx \bar{\tau}_i$ is basically an assumption that $\mu \approx \hat{\mu}_c$ (which is what we are trying to show in the first place!).

---

24
Figure 7: Intuition behind Algorithm 2. Because there is only an \( \epsilon \)-fraction of bad data, it must lie far away to increase the variance by a constant factor.

- The bad points are not deterministically larger than the good points; they are only separated in expected value.
- There are many fewer bad points than good points, so they are harder to find.

We will deal with the first issue by showing that \( \mu \) is close enough to \( \hat{\mu}_c \) for the algorithm to make progress. The second issue is why we need to do soft downweighting rather than picking a hard threshold and removing all points with \( \tau_i \) above the threshold. We will resolve the third issue by showing that we always remove more mass \( c_i \) from the bad points than from the good points when we update \( c_i \). Intuitively, while there are only \( \epsilon \) times as many bad points as good points, this is balanced against the fact that the mean of the bad points is \( 1/\epsilon \) times as large as the mean of the good points.

We next put this intuition together into a formal proof.

**Proof of Proposition 2.38.** As above, for weights \( c_i \in [0, 1] \), let \( q(c) \) be the distribution that assigns weight \( c_i / \sum_j c_j \) to point \( x_i \). Thus when \( c_i = 1 \) for all \( i \), we have \( q(c) = \tilde{p} \). Our hope is that as the algorithm progresses \( q(c) \) approaches \( p^* \) or at least has small covariance. We will establish the following invariant:

\[
\text{TV}(q(c), p^*) \leq \frac{\epsilon}{1 - \epsilon}
\]

for all weight vectors \( c \) used during the execution of Algorithm 2. \((I_1)\)

We will do this by proving the following more complex invariant, which we will later show implies \((I_1)\):

\[
\sum_{i \in S} (1 - c_i) \leq \sum_{i \notin S} (1 - c_i)
\]

\((I_2)\)

The invariant \((I_2)\) says that the total probability mass removed from the good points is less than the total probability mass removed from the bad points. A key lemma relates \((I_2)\) to the \( \tau_i \):

**Lemma 2.39.** If \((I_2)\) and \( \sum_{i \in S} c_i \tau_i \leq \sum_{i \notin S} c_i \tau_i \), then it continues to hold after the update \( c_i' = c_i (1 - \tau_i / \tau_{\text{max}}) \).

**Proof.** For any set \( T \), we have

\[
\sum_{i \in T} 1 - c_i' = \sum_{i \in T} (1 - c_i) + \sum_{i \in T} (c_i - c_i') = \sum_{i \in T} (1 - c_i) + \frac{1}{\tau_{\text{max}}} \sum_{i \in T} c_i \tau_i.
\]

Applying this for \( T = S \) and \( T = [n] \setminus S \) yields the lemma. \(\square\)
Thus our main job is to show that $\sum_{i \in S} c_i \tau_i \leq \sum_{i \in S} c_i \tau_i$. Equivalently, we wish to show that $\sum_{i \in S} c_i \tau_i \leq \frac{1}{2} \sum_{i = 1}^n c_i \tau_i$. For this, the following bound is helpful:

$$\sum_{i \in S} c_i \tau_i = \sum_{i \in S} c_i (x_i - \hat{\mu}_c, v^*)^2 \leq \sum_{i \in S} (x_i - \hat{\mu}_c, v^*)^2$$

$$= (1 - \epsilon) n E_p [ (x_i - \hat{\mu}_c, v^*)^2 ]$$

$$= (1 - \epsilon) n \cdot (v^*)^\top (\text{Cov}_p [X] + (\mu - \hat{\mu}_c)(\mu - \hat{\mu}_c)^\top)(v^*)$$

$$\leq (1 - \epsilon) n \cdot (\| \text{Cov}_p [X] \| + \| \mu - \hat{\mu}_c \|^2_2).$$

Here the second-to-last step uses the fact that for any $\theta$, $E[(X - \theta)(X - \theta)^\top] = \text{Cov}[X] + (\theta - \mu)(\theta - \mu)^\top$. Next note that $\| \text{Cov}_p \| \leq \sigma^2$ while $\| \mu - \hat{\mu}_c \|^2_2 \leq \frac{4\epsilon}{1 - 2\epsilon} \sigma^2$ by the modulus of continuity bound combined with the fact that $p^*, q(c) \in \mathcal{G}_{\text{cov}}(\hat{\sigma})$ and $\text{TV}(p^*, q(c)) \leq \frac{1}{1 - \epsilon}$. Therefore, we have

$$\sum_{i \in S} c_i \tau_i \leq (1 - \epsilon) \sigma^2 n + \frac{8\epsilon (1 - \epsilon)}{1 - 2\epsilon} \sigma^2 n.$$  

On the other hand, we have

$$\sum_{i = 1}^n c_i \tau_i = (\sum_{i = 1}^n c_i) \| \text{Cov}_{q(c)} [X] \| = (\sum_{i = 1}^n c_i) \hat{\sigma}_c^2 \geq (1 - 2\epsilon) \hat{\sigma}_c^2 n,$$

where the final inequality uses the fact that we have so far removed more mass from bad points than good points and hence at most $2\epsilon$ mass in total. Recalling that we wish to show that (102) is at most half of (103), we require that

$$(1 - 2\epsilon) \hat{\sigma}_c^2 \geq 2(1 - \epsilon) \sigma^2 + \frac{16\epsilon (1 - \epsilon)}{1 - 2\epsilon} \sigma^2,$$

which upon re-arrangement yields

$$\hat{\sigma}_c^2 \geq \frac{2(1 - \epsilon)(1 - 2\epsilon)}{1 - 12\epsilon + 12\epsilon^2} \sigma^2.$$ 

Since $\hat{\sigma}_c^2 \geq 20\sigma^2$ whenever the algorithm does not terminate, this holds as long as $\epsilon \leq \frac{1}{12}$ (then the constant in front of $\sigma^2$ is $\frac{5}{3} < 20$). This shows that $(I_2)$ holds throughout the algorithm.

The one remaining detail is to prove that $(I_2)$ implies $(I_1)$. We wish to show that $\text{TV}(p^*, q(c)) \leq \frac{1}{1 - \epsilon}$. We use the following formula for $\text{TV}$: $\text{TV}(p, q) = \int \max(q(x) - p(x), 0) dx$. Let $\beta$ be such that $\sum_{i = 1}^n c_i = (1 - \beta)n$. Then we have

$$\text{TV}(p^*, q(c)) = \sum_{i \in S} \max \left( \frac{c_i}{(1 - \beta)n} - \frac{1}{(1 - \epsilon)n}, 0 \right) + \sum_{i \notin S} \frac{c_i}{(1 - \beta)n}.$$ 

If $\beta \leq \epsilon$, then the first sum is zero while the second sum is at most $\frac{\epsilon}{1 - \beta} \leq \frac{\epsilon}{1 - \epsilon}$. If on the other hand $\beta > \epsilon$, we will instead use the equality obtained by swapping $p$ and $q$, which yields

$$\text{TV}(p^*, q(c)) = \sum_{i \in S} \max \left( \frac{1}{(1 - \epsilon)n} - \frac{c_i}{(1 - \beta)n}, 0 \right)$$

$$= \frac{1}{(1 - \epsilon)(1 - \beta)n} \sum_{i \in S} \max ((1 - \beta)(1 - c_i) + (\epsilon - \epsilon)c_i, 0).$$

Since $(\epsilon - \epsilon)c_i \leq 0$ and $\sum_{i \in S}(1 - c_i) \leq \epsilon n$, this yields a bound of $\frac{(1 - \beta)e}{(1 - \epsilon)(1 - \beta)} = \frac{\epsilon}{1 - \epsilon}$. We thus obtain the desired bound no matter the value of $\beta$, so $\text{TV}(p^*, q(c)) \leq \frac{1}{1 - \epsilon}$ whenever $(I_2)$ holds. This completes the proof. \qed
2.7.1 Approximate Eigenvectors in Other Norms

Algorithm 2 is specific to the \( \ell_2 \)-norm. Let us suppose that we care about recovering an estimate \( \hat{\mu} \) such that \( \| \mu - \hat{\mu} \| \) is small in some norm other than \( \ell_2 \) (such as the \( \ell_1 \)-norm, which may be more appropriate for some combinatorial problems). It turns out that an analog of bounded covariance is sufficient to enable estimation with the typical \( O(\sigma \sqrt{\epsilon}) \) error, as long as we can approximately solve the analogous eigenvector problem. To formalize this, we will make use of the dual norm:

Definition 2.40. Given a norm \( \| \cdot \| \), the dual norm \( \| \cdot \|_* \) is defined as

\[
\|u\|_* = \sup_{\|v\|_2 \leq 1} \langle u, v \rangle. \tag{109}
\]

As some examples, the dual of the \( \ell_2 \)-norm is itself, the dual of the \( \ell_1 \)-norm is the \( \ell_\infty \)-norm, and the dual of the \( \ell_\infty \)-norm is the \( \ell_1 \)-norm. An important property (we omit the proof) is that the dual of the dual is the original norm:

Proposition 2.41. If \( \| \cdot \| \) is a norm on a finite-dimensional vector space, then \( \| \cdot \|_* = \| \cdot \| \).

For a more complex example: let \( \|v\|_{(k)} \) be the sum of the \( k \) largest coordinates of \( v \) (in absolute value). Then the dual of \( \| \cdot \|_{(k)} \) is \( \max(\|u\|_\infty, \|u\|_1/k) \). This can be seen by noting that the vertices of the constraint set \( \{ u \mid \|u\|_\infty \leq 1, \|u\|_1 \leq k \} \) are exactly the \( k \)-sparse \( \{-1,0,+1\} \)-vectors.

Let \( \mathcal{G}_{\text{cov}}(\sigma, \| \cdot \|) \) denote the family of distributions satisfying \( \max_{\|v\|_\leq 1} v^\top \text{Cov}_p[X]v \leq \sigma^2 \). Then \( \mathcal{G}_{\text{cov}} \) is resilient exactly analogously to the \( \ell_2 \)-case:

Proposition 2.42. If \( p \in \mathcal{G}_{\text{cov}}(\sigma, \| \cdot \|) \) and \( r \leq \frac{p}{1+\epsilon} \), then \( \|\mu(r) - \mu(p)\| \leq \sqrt{\frac{2 \epsilon}{1+\epsilon}} \sigma \). In other words, all distributions in \( \mathcal{G}_{\text{cov}}(\sigma, \| \cdot \|) \) are \( (\epsilon, O(\sigma \sqrt{\epsilon})) \)-resilient.

Proof. We have that \( \|\mu(r) - \mu(p)\| = \langle \mu(r) - \mu(p), v \rangle \) for some vector \( v \) with \( \|v\|_* = 1 \). The result then follows by resilience for the one-dimensional distribution \( \langle X, v \rangle \) for \( X \sim p \). \( \square \)

When \( p^* \in \mathcal{G}_{\text{cov}}(\sigma, \| \cdot \|) \), we will design efficient algorithms analogous to Algorithm 2. The main difficulty is that in norms other than \( \ell_2 \), it is generally not possible to exactly solve the optimization problem \( \max_{\|v\|_\leq 1} v^\top \Sigma v \) that is used in Algorithm 2. We instead make use of a \( \kappa \)-approximate oracle:

Definition 2.43. A function \( A(\Sigma) \) is a \( \kappa \)-approximate oracle if for all \( \Sigma, M = A(\Sigma) \) is a positive semidefinite matrix satisfying

\[
\langle M, \Sigma \rangle \geq \sup_{\|v\|_\leq 1} v^\top \Sigma v, \text{ and } \langle M, \Sigma' \rangle \leq \kappa \sup_{\|v\|_* \leq 1} v^\top \Sigma' v \text{ for all } \Sigma' \geq 0. \tag{110}
\]

Thus a \( \kappa \)-approximate oracle over-approximates \( \langle vv^\top, \Sigma \rangle \) for the maximizing vector \( v \) on \( \Sigma \), and it underapproximates \( \langle vv^\top, \Sigma' \rangle \) within a factor of \( \kappa \) for all \( \Sigma' \neq \Sigma \). Given such an oracle, we have the following analog to Algorithm 2:

Algorithm 3 FilterNorm

1: Initialize weights \( c_1, \ldots, c_n = 1 \).
2: Compute the empirical mean \( \hat{\mu}_c \) of the data, \( \hat{\mu}_c \overset{\text{def}}{=} (\sum_{i=1}^n c_i x_i)/(\sum_{i=1}^n c_i) \).
3: Compute the empirical covariance \( \Sigma_c \overset{\text{def}}{=} \sum_{i=1}^n c_i (x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^\top / \sum_{i=1}^n c_i \).
4: Let \( M = A(\Sigma_c) \) be the output of a \( \kappa \)-approximate oracle.
5: If \( \langle M, \Sigma_c \rangle \leq 20\kappa \sigma^2 \), output \( q(c) \).
6: Otherwise, let \( \tau_i = (x_i - \hat{\mu}_c)^\top M(x_i - \hat{\mu}_c) \), and update \( c_i \leftarrow c_i \cdot (1 - \tau_i/\tau_{\text{max}}) \), where \( \tau_{\text{max}} = \max_i \tau_i \).
7: Go back to line 2.

Algorithm 3 outputs an estimate of the mean with error \( O(\sigma \sqrt{\kappa \epsilon}) \). The proof is almost exactly the same as Algorithm 2; the main difference is that we need to ensure that \( \langle \Sigma, M \rangle \), the inner product of \( M \) with the true covariance, is not too large. This is where we use the \( \kappa \)-approximation property. We leave the detailed proof as an exercise, and focus on how to construct a \( \kappa \)-approximate oracle \( A \).
A Properties of Statistical Discrepancies

A.1 Total variation distance

A.2 Wasserstein distance

B Concentration Inequalities

B.1 Proof of Chebyshev’s inequality (Lemma 2.1)

Let $\mathbb{I}[E]$ denote the indicator that $E$ occurs. Then we have

$$\mathbb{E}_{X \sim p}[X | E] - \mu = |\mathbb{E}_{X \sim p}[X - \mu]| / \mathbb{P}[E] = 1 - P(\mathbb{E}[E])$$

$$\leq \sqrt{\mathbb{E}_{X \sim p}[(X - \mu)^2] \cdot \mathbb{E}_{X \sim p}[|\mathbb{I}[E]|^2] / \mathbb{P}[E]} \leq \sqrt{\sigma^2 \cdot \mathbb{E}[E] / \mathbb{P}[E]} = \sigma / \sqrt{\mathbb{P}[E]}.$$ (111)

In particular, if we let $E_0$ be the event that $X \geq \mu + \sigma / \sqrt{\delta}$, we get that $\sigma / \sqrt{\delta} \leq \sigma / \sqrt{\mathbb{P}[E_0]}$, and hence $\mathbb{P}[E_0] \leq \delta$, which proves the first part of the lemma.

For the second part, if $\mathbb{P}[E] \leq \frac{1}{2}$ then (113) already implies the desired result since $\sigma / \sqrt{\delta} \leq \sigma \sqrt{2(1 - \delta) / \delta}$ when $\delta \leq \frac{1}{2}$. If $\mathbb{P}[E] \geq \frac{1}{2}$, then consider the same argument applied to $\neg E$ (the event that $E$ does not occur). We get

$$\mathbb{E}_{X \sim p}[X | E] - \mu = \frac{1 - \mathbb{P}[E]}{\mathbb{P}[E]} |\mathbb{E}_{X \sim p}[X | \neg E] - \mu| \leq \frac{1 - \mathbb{P}[E]}{\mathbb{P}[E]} \cdot \sigma / \sqrt{1 - \mathbb{P}[E]}.$$ (114)

Again the result follows since $\sigma \sqrt{1 - \delta} / \delta \leq \sigma \sqrt{2(1 - \delta) / \delta}$ when $\delta \geq \frac{1}{2}$.

B.2 Proof of $d$-dimensional Chebyshev’s inequality (Lemma 2.8)

C Proof of Lemma 2.14

Since $(\rho, \epsilon)$-resilience is equivalent to $(\frac{1 - \epsilon}{\epsilon} \rho, 1 - \epsilon)$-resilience, it suffices to show that $(1 - \epsilon, \frac{1 - \epsilon}{\epsilon} \rho)$-resilience is equivalent to (18). Suppose that $E$ is an event with probability $\epsilon$, and let $v$ be such that $\|v\|_* = 1$ and
\[ \langle E[X - \mu \mid E], v \rangle = \| E[X - \mu \mid E] \|. \] Then we have

\[
\begin{align*}
\| E[X - \mu \mid E] \| &= \langle E[X - \mu \mid E], v \rangle \\
&= \langle E[(X - \mu, v) \mid E] \rangle \\
&\leq \mathbb{E}[\langle X - \mu, v \mid (X - \mu, v) \rangle \geq \tau_\epsilon(v)] \\
&\leq \frac{1 - \epsilon}{\epsilon} \rho.
\end{align*}
\] (118)

Here (i) is because \( \langle X - \mu, v \rangle \) is at least as large for the \( \epsilon \)-quantile as for any other event \( E \) of probability \( \epsilon \).

This shows that (18) implies \( (1 - \epsilon, \frac{1 - \epsilon}{\epsilon} \rho) \)-resilience. For the other direction, given any \( v \) let \( E_v \) denote the event that \( \langle X - \mu, v \rangle \geq \tau_\epsilon(v) \). Then \( E_v \) has probability \( \epsilon \) and hence

\[
\begin{align*}
\mathbb{E}[\langle X - \mu, v \mid (X - \mu, v) \rangle \geq \tau_\epsilon(v)] &= \mathbb{E}[\langle X - \mu, v \rangle \mid E_v] \\
&= \langle \mathbb{E}[X - \mu \mid E_v], v \rangle \\
&\leq \| \mathbb{E}[X - \mu \mid E_v] \| \\
&\leq \frac{1 - \epsilon}{\epsilon} \rho,
\end{align*}
\] (123)

where (ii) is Hölder’s inequality and (iii) invokes resilience. Therefore, resilience implies (18), so the two properties are equivalent, as claimed.

**D Proof of Lemma 2.15**

Let \( E_+ \) be the event that \( (x_i - \mu, v) \) is positive, and \( E_- \) the event that it is non-negative. Then \( \mathbb{P}[E_+] + \mathbb{P}[E_-] = 1 \), so at least one of \( E_+ \) and \( E_- \) has probability at least \( \frac{1}{2} \). Without loss of generality assume it is \( E_+ \). Then we have

\[
\mathbb{E}_{x \sim p}[\langle (x - \mu, v) \rangle] = 2\mathbb{E}_{x \sim p}[\max(\langle x - \mu, v \rangle, 0)]
\] (124)

\[
= 2\mathbb{P}[E_+] \mathbb{E}_{x \sim p}[\langle x - \mu, v \rangle \mid E_+]
\] (125)

\[
\leq 2\mathbb{P}[E_+] \| \mathbb{E}_{x \sim p}[x - \mu \mid E_+] \| \leq 2\rho,
\] (126)

where the last step invokes resilience applies to \( E_+ \) together with \( \mathbb{P}[E_+] \leq 1 \). Conversely, if \( p \) has bounded 1st moments then

\[
\mathbb{E}[\langle X - \mu, v \rangle \mid (X - \mu, v) \geq \tau_1/2(v)] \leq \mathbb{E}[\| (X - \mu, v) \| / \mathbb{P}[ (X - \mu, v) \geq \tau_1/2(v)]]
\] (127)

\[
= 2\mathbb{E}[\| (X - \mu, v) \|] \leq 2\rho,
\] (128)

so \( p \) is \( (2\rho, \frac{1}{2}) \)-resilient by Lemma 2.14.