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 \( p^* \in \mathcal{G} \), 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 $D(p, q) = \text{TV}(p, q)$, where $\text{TV}$ is the total variation distance:

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

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

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 $D = \text{TV}$.

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 \( \tilde{p} \). 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 \( \mathcal{O}(\sigma \sqrt{\epsilon}) \) 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 \( \tilde{p}_{2\epsilon} \) denote the new distribution after re-normalizing, and return the mean of \( \tilde{p}_{2\epsilon} \).

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 \( |\mathbb{E}_{X \sim p}[X | E] - \mu| \leq \sigma \sqrt{2(1-\delta)/\delta} \).

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 \( \epsilon' \cdot \sigma \sqrt{\frac{1}{2} + (1-\epsilon') \cdot \sigma \sqrt{\frac{10\epsilon}{1-5\epsilon}}} \), which is at most \( (1 + \sqrt{10}) \frac{\sqrt{\epsilon}}{1-4\epsilon} \sigma \). This is in turn at most \( 9\sigma \sqrt{\epsilon} \) assuming that \( \epsilon \leq \frac{1}{8} \). \( \square \)

**Optimality.** The \( \mathcal{O}(\sigma \sqrt{\epsilon}) \) 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 \( \mathcal{O}(\sigma \sqrt{\epsilon}) \) 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 \( \mathcal{O}(\sigma \sqrt{\epsilon}) \) 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 \( \mathcal{O}(\sigma \sqrt{\epsilon}) \) in Lemma 2.1 can be improved to \( \mathcal{O}(\sigma_k \epsilon^{1-1/k}) \), where \( \sigma_k \) is a bound on the \( k \)th moment; in this case Algorithm 1 will estimate \( \mu \) with a correspondingly improved error of \( \mathcal{O}(\sigma_k \epsilon^{1-1/k}) \).
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.

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

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

(2)

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 $\mathcal{G}$:
Proposition 2.4. Suppose \( D \) is a pseudometric. Then the cost \( L(p^*, \hat{\theta}(\hat{p})) \) of the minimum distance functional is at most the maximum loss between any pair of distributions in \( \mathcal{G} \) of distance at most \( 2\epsilon \):

\[
m(\mathcal{G}, 2\epsilon, D, L) \triangleq \sup_{p, q \in \mathcal{G} : D(p, q) \leq 2\epsilon} L(p, \theta^*(q)).
\] (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 \( \sup_{p, q \in \mathcal{G} : \text{TV}(p, q) \leq 2\epsilon} \|\mu(p) - \mu(q)\|_2 \). 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

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

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

(5)

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

(6)

(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.

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 \( \Sigma \preceq \sigma^2 I \). 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

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

(7)

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

As a consequence, the minimum distance functional robustly estimates the mean bounded covariance distributions with error $O(\sigma \sqrt{\epsilon})$, 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 $G$ 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

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

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

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

Resilient distributions always have bounded modulus:

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

$$m(G_{TV}(\rho, \epsilon), 2\epsilon) \leq 2\rho$$

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 $r = \min(p,q) / \max(1-\text{TV}(p,q))$ that is a deletion of both distributions. This midpoint distribution connects the two distributions, and it follows from the triangle inequality that the modulus of $G_{TV}$ is bounded. We illustrate this idea in Figure 4 and make it precise below.

Recall that

$$m(G_{TV}(\rho, \epsilon), 2\epsilon) = \sup_{p,q \in G_{TV}(\rho, \epsilon) : \text{TV}(p,q) \leq 2\epsilon} \|\mu(p) - \mu(q)\|.$$  

From $\text{TV}(p,q) \leq 2\epsilon$, we know that $r = \min(p,q) / \max(1-\text{TV}(p,q))$ can be obtained from either $p$ and $q$ by conditioning on an event of probability $1-\epsilon$. It then follows from $p,q \in G_{TV}(\rho, \epsilon)$ 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. 

7
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 $\psi : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ is an Orlicz function if $\psi$ is convex, non-decreasing, and satisfies $\psi(0) = 0$, $\psi(x) \rightarrow \infty$ as $x \rightarrow \infty$. For an Orlicz function $\psi$, the Orlicz norm or $\psi$-norm of a random variable $X$ is defined as

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

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

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 $O(\sigma \epsilon^{1-1/k})$ 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 $\langle x - \mu, v \rangle$: $P_{x\sim p}[\langle x - \mu, v \rangle \geq \tau_\epsilon(v)] = \epsilon$. 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 $v$:

\[
E_p[\langle x - \mu, v \rangle \mid \langle x - \mu, v \rangle \geq \tau_\epsilon(v)] \leq \frac{1 - \epsilon}{\epsilon} \rho \text{ whenever } \|v\|_* \leq 1. \tag{18}
\]

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 \(\mathcal{G}\). Instead, we can always project onto the resilient family \(\mathcal{G}_{TV}(\rho, \epsilon)\) 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 \(\hat{p}\); 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 \(\hat{p}\). 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, \(\mathbb{P}[X \geq t \cdot \mu] \leq \frac{1}{t}\).
Markov’s inequality accomplishes our first goal of establishing concentration for a single random variable, but it has two issues: first, the $\frac{1}{t}$ 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, $\Pr[|X - \mu| \geq t \cdot \sigma] \leq \frac{1}{t^2}$.

**Proof.** Since $Z = (X - \mu)^2$ is non-negative, we have that $\Pr[Z \geq t^2 \cdot \sigma^2] \leq \frac{1}{t^2}$ by Markov’s inequality. Taking the square-root gives $\Pr[|X - \mu| \geq t \cdot \sigma] \leq \frac{1}{t^2}$, 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, $\Pr[|S - \mu|/\sigma \geq t/\sqrt{n}] \leq 1/t^2$.

**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 get the analogous bound $\Pr[|X - \mu| \geq t \cdot \tau] \leq 1/t^4$. 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

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

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 $O(\sqrt{n})^p$.

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 $E[\exp(|X - \mu|/\sigma)] \leq 2$. For these, applying the Chernoff bound with $\lambda = 1/\sigma$ yields $P[X - \mu \geq t] \leq 2e^{-t/\sigma}$.

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

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

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 $P[X - \mu \geq t] \leq \exp(3\lambda^2\sigma^2/2 - \lambda t)$; minimizing over $\lambda$ gives the optimized bound $P[X - \mu \geq t] \leq \exp(-t^2/6\sigma^2)$.

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 $m_{X_1 + \ldots + X_n}(\lambda) \leq \exp(3\lambda^2(\sigma_1^2 + \cdots + \sigma_n^2)/2)$. We will use this to bound the behavior of sums of bounded random variables using Hoeffding’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, $P[S \geq t] \leq \exp(-\ln(2)nt^2/6M^2) \leq \exp(-nt^2/9M^2)$.

**Proof.** First, note that each $X_i$ is sub-Gaussian with parameter $\sigma = M/\sqrt{\ln 2}$, since $E[\exp(X_i^2/\sigma^2)] \leq \exp(M^2/\sigma^2) = \exp(\ln(2)) = 2$. 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 $P[S \geq t] \leq \exp(-\ln(2)nt^2/6M^2)$ 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 $E[\exp(\lambda X)]$ must be finite. For heavy-tailed distributions a (laborious) alternative is to use cumulants.

The cumulant function is defined as

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

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

$$\kappa_1(X) = 0,$$

$$\kappa_2(X) = E[X^2],$$

$$\kappa_3(X) = E[X^3],$$

$$\kappa_4(X) = E[X^4] - 3E[X^2]^2,$$

$$\kappa_5(X) = E[X^5] - 10E[X^3]E[X^2],$$


Since $K$ is additive, each of the $\kappa_n$ are as well. Thus while we ran into the issue that $E[(X + Y)^4] \neq E[X^4] + E[Y^4]$, 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 $P[X_i \geq \sigma \sqrt{6 \log(n/\delta)}] \leq \exp(-\log(n/\delta)) = \delta/n$. 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. \qed

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. As long as the “failure probability” of a single event is much smaller than the inverse of the number of events, we obtain a meaningful bound. In fact, this strategy can be employed even for an infinite number of events by discretizing to an “$\epsilon$-net”, as we will see below:

**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

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

(33)

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

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

(34)

$$
\leq 2^n, 
$$

(35)

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

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

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

(36)

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 \( v^\top M v \) over \( \|v\|_2 \leq 1 \) and let \( u \) maximize \( v^\top M v \) over \( \mathcal{N}_{1/4} \). Then
\[
|v^\top M v - u^\top M u| = |v^\top M (v - u) + u^\top M (v - u)| \\
\leq (\|v\|_2 + \|u\|_2) \|M\| \|v - u\|_2 \\
\leq 2 \cdot \|M\| \cdot (1/4) = \|M\|/2.
\]
Since \( v^\top M v = \|M\| \), we obtain \( \|M\| - u^\top M u \leq \|M\|/2 \), whence \( v^\top M u \geq \|M\|/2 \), which establishes (36). We are now ready to apply the union bound: Recall that from the Chernoff bound on \( v^\top M v \), we had \( \mathbb{P}[v^\top M v \geq t] \leq 2^n \exp(-nt/\sigma^2) \), so
\[
\mathbb{P}\left[ \sup_{v \in \mathcal{N}_{1/4}} v^\top M v \geq t \right] \leq 9^d 2^n \exp(-nt/\sigma^2).
\]
Solving for this quantity to equal \( \delta \), we obtain
\[
t = \frac{\sigma^2}{n} \cdot (n \log(2) + d \log(9) + \log(1/\delta)) = \mathcal{O}(\sigma^2 \cdot (1 + d/n + \log(1/\delta)/n)),
\]
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
\[
\sup_{f \in \mathcal{H}} |\nu_n(f) - \nu(f)| \leq \mathcal{O}\left(\sqrt{\frac{d + \log(1/\delta)}{n}}\right)
\]
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

$$E_X[\sup_{f \in \mathcal{H}} |\nu_n(f) - \nu(f)|| \leq E_{X,X'}[\sup_{f \in \mathcal{H}} |\nu_n(f) - \nu'_n(f)||]$$

$$= \frac{1}{n}E_{X,X'}[\sup_{f \in \mathcal{H}} \sum_{i=1}^n f(X_i) - f(X'_i)].$$

(43)

(44)

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

$$\frac{1}{n}E_{X,X'}[\sup_{f \in \mathcal{H}} \sum_{i=1}^n f(X_i) - f(X'_i)] = \frac{1}{n}E_{X,X',s}[\sup_{f \in \mathcal{H}} \sum_{i=1}^n s_i(f(X_i) - f(X'_i))].$$

(45)

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 $V_n(n)^2 = O(n^{2d})$ 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 $P[\sum_i s_i(f(X_i) - f(X'_i)) \geq t] \leq \exp(-t^2/9n)$ by Hoeffding’s inequality. Union bounding over the $O(n^{2d})$ effectively distinct $f$, we obtain

$$P[\sup_{f \in \mathcal{H}} |\sum_i s_i(f(X_i) - f(X'_i))| \geq t] \leq O(n^{2d}) \exp(-t^2/9n).$$

(46)

This is small as long as $t \gg \sqrt{d \log n}$, so (45) is $O(\sqrt{d \log n}/n)$, 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 $F_n(x) = \frac{1}{n}\sum_{i=1}^n I[X_i \leq x]$, and the population cumulative density function as $F(x) = p(X \leq x)$. Then $P[\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \geq t] \leq 2e^{-2nt^2}$.

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

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

(47)

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.$^2$

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 $m(G, \epsilon)$ 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 $G$ to some $M \supset G$, such that even though $p$ is not close to the empirical distribution $p_n$, some element of $M$ is close to $p_n$. Another advantage to expanding the destination set is that projecting onto $\tilde{TV}$ may not be computationally tractable, while projecting onto some larger set $M$ 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 $M$ 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 $p^* \in G$ and $TV(p^*, \tilde{p}) \leq \epsilon$, then the error of the minimum distance functional (2) with $D = TV$ is at most $m(G, 2\epsilon', \tilde{TV}, L)$, where $\epsilon' = \epsilon + TV(\tilde{p}, \tilde{p}_n)$.

Proof. By Proposition 2.4 we already know that the error is bounded by $m(G, 2\tilde{TV}(p^*, \tilde{p}_n), TV, L)$. Since $\tilde{TV}$ is a pseudometric, by the triangle inequality we have $\tilde{TV}(p^*, \tilde{p}) \leq \tilde{TV}(p^*, \tilde{p}_n) + TV(\tilde{p}, \tilde{p}_n)$. 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 $m(G, \epsilon, \tilde{TV})$ should not be too much larger than $m(G, \epsilon, TV)$.

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

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

(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 $\mathcal{H}_{\text{lin}} \overset{\text{def}}{=} \{ v \mapsto \langle v, x \rangle \mid v \in \mathbb{R}^d \}$.

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:

Figure 6: 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.
Lemma 2.28. Suppose that \( p \) and \( q \) are two distributions such that \( \TV_{\mathcal{H}}(p,q) \leq \epsilon \). Then for any \( f \in \mathcal{H} \), there are distributions \( r_p \leq \frac{p}{1-\epsilon} \) and \( r_q \leq \frac{q}{1-\epsilon} \) such that \( \mathbb{E}_{X \sim r_p}[f(X)] \geq \mathbb{E}_{Y \sim r_q}[f(Y)] \).

Proof. We will prove the stronger statement that \( f(X) \) under \( r_p \) stochastically dominates \( f(Y) \) under \( r_q \). Starting from \( p, q \), we delete \( \epsilon \) 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 \( \TV_{\mathcal{H}}(p,q) \) we have

\[
\sup_{\tau \in \mathbb{R}} |\mathbb{P}_{X \sim r_p}(f(X) \geq \tau) - \mathbb{P}_{Y \sim r_q}(f(Y) \geq \tau)| \leq \epsilon, \tag{49}
\]

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

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

\[
m(\mathcal{G}_{TV}(\rho, \epsilon), \epsilon, \TV_{\mathcal{H}_{lin}}) \leq 2\rho. \tag{50}
\]

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

Proof of Corollary 2.29. Let \( p, q \in \mathcal{G}_{TV} \) such that \( \TV(p,q) \leq \epsilon \). Take \( v = \arg \max_{\|v\|_2 = 1} v^T(\mathbb{E}_p[X] - \mathbb{E}_q[X]) \), hence \( \mathbb{E}_p[v^T X] - \mathbb{E}_q[v^T X] = \|\mathbb{E}_p[X] - \mathbb{E}_q[X]\|_2 \). 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

\[
\mathbb{E}_{r_p}[v^T X] \leq \mathbb{E}_{r_q}[v^T X]. \tag{51}
\]

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

\[
\mathbb{E}_p[v^T X] - \mathbb{E}_{r_p}[v^T X] \leq \rho, \tag{52}
\]

\[
\mathbb{E}_{r_q}[v^T X] - \mathbb{E}_q[v^T X] \leq \rho. \tag{53}
\]

Then,

\[
\|\mathbb{E}_p[X] - \mathbb{E}_q[X]\|_2 = \mathbb{E}_p[v^T X] - \mathbb{E}_q[v^T X] \leq \mathbb{E}_p[v^T X] - \mathbb{E}_{r_p}[v^T X] + \mathbb{E}_{r_q}[v^T X] - \mathbb{E}_q[v^T X] \leq 2\rho, \tag{54}
\]

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

\[
\sum_{i=1}^{d+2} a_i x_i = 0, \quad \sum_{i=1}^{d+2} a_i = 0. \tag{57}
\]
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 half-space 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 $x_1, \ldots, x_d$ 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 $\text{TV}_{H_{\text{lin}}}(\tilde{p}, \tilde{p}_n) \leq O\left(\sqrt{\frac{d + \log(1/\delta)}{n}}\right)$.

Combining this with Corollary 2.29 and Lemma 2.27, we see that projecting onto $G_{\text{TV}}(\rho, 2\epsilon')$ under $\text{TV}_{H_{\text{lin}}}$ 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 $O\left(\epsilon + \sqrt{d/n} \psi^{-1}\left(\frac{1}{\epsilon + \sqrt{d/n}}\right)\right) \leq O\left(\epsilon + \sqrt{d/n} \psi^{-1}(1/\epsilon)\right)$.

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

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

(59)

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

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

(60)

The differences are: (1) the Tukey median only minimizes over the mean rather than the full distribution $q$; (2) it only considers the threshold $\langle \mu, v \rangle$ rather than all thresholds $\tau$; it assumes that the median of any one-dimensional projection $\langle X, v \rangle$ is equal to its mean (which is why we subtract $\frac{1}{2}$ in (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 TV projection by relaxing to a weaker distance $\text{TV}$. We will now study an alternate approach, based on expanding the destination set $G$ to a larger set $M$. For this approach we will need to reference the “true empirical distribution” $p_n$. What we mean by this is the following: Whenever $\text{TV}(p^*, \tilde{p}) \leq \epsilon$, we know that $p^*$ 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 $X \sim p^*$.
2. Check if $E$ holds; if it does, replace $X$ with a sample from the conditional distribution $\tilde{p}|_E$. 

18
Assume that the true empirical distribution \( p \) where least

Let \( \psi \) is too large. We will instead show that certain

\[ \text{Proposition 2.32.} \]

\[ \text{Let symmetrization argument, we will bound the truncated sup} \]

\[ \| G \|_{\text{mom}} \] is bounded by \( O(q, p^*) \), so the hard part is showing that the empirical distribution \( \psi \) has bounded moments except when \( n \geq 1 \) and hence \( \min_{q, p} \text{TV}(q, p) \leq 2\epsilon' \). It follows from the definition that \( L(p^*; \hat{\theta}) = L(p^*; \theta^*(q)) \leq \mathbf{m}(G', \mathcal{M}, 2\epsilon') \).

\[ \text{Application: bounded kth moments.} \]

First suppose that the distribution \( p^* \) has bounded \( k \)th moments, i.e. \( \mathcal{G}_{\text{mom}, k}(\sigma) = \{ p \mid \| p \|_\psi \leq \sigma \} \), where \( \psi(x) = x^k \). When \( k > 2 \), the empirical distribution \( p^*_n \) 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 \( \mathbb{E}_{x \sim p^*_n} [ f(x - \mu, v)^k ] \geq \frac{1}{n} \| x_1 - \mu \|_2^k \geq d^{k/2}/n \), since the norm of \( \| x_1 - \mu \|_2 \) is typically \( \sqrt{d} \).

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

As noted above, we cannot hope to prove that \( p^*_n \) has bounded moments except when \( n = \Omega(d^{k/2}) \), which is too large. We will instead show that certain \textit{truncated} moments of \( p^*_n \) 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

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

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 \( L = k x_0^{k-1} \). 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 sup \( \sup_{\| v \|_2 \leq 1} \mathbb{E}_{x \sim p^*_n} [ \tilde{\psi}( | (x - \mu, v) | / \sigma ) ] \).

\[ \text{Proposition 2.32.} \]

\[ \text{Let } X_1, \ldots, X_n \sim p^*, \text{ where } p^* \in \mathcal{G}_{\text{mom}, k}(\sigma). \text{ Then,} \]

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

where \( L = k x_0^{k-1} \).
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 \( n \geq \Omega(k^3d/\epsilon^{2-2/k}) \). We then obtain a bound on the \( \psi \)-norm of \( p_n^* \), which implies that \( p_n^* \) is resilient with parameter \( \rho = \sigma \epsilon^{-1} (2/\epsilon) = 2 \sigma \epsilon^{1-1/k} \). This matches the population-bound of \( O(\sigma \epsilon^{1-1/k}) \), and only requires \( d/\epsilon^2 \) 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 \( \psi(\langle X - \mu, v \rangle/\sigma) \) in (81) with the simpler expression \( L \langle X - \mu, v \rangle/\sigma \). 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

\[
E_{\epsilon_{1:n}}[g(\sup_{t \in T} \sum_{i=1}^{n} \epsilon_i \phi(t_i))] \leq E_{\epsilon_{1:n}}[g(L \sup_{t \in T} \sum_{i=1}^{n} \epsilon_i t_i)].
\]

(64)

Let us interpret this result. We should think of the \( t_i \) as a quantity such as \( \langle x_i - \mu, v \rangle \), where abstracting to \( t_i \) yields generality and notational simplicity. **Theorem 2.33** says that if we let \( Y = \sup_{t \in T} \sum_{i} \epsilon_i \phi(t_i) \) and \( Z = L \sup_{t \in T} \sum_{i} \epsilon_i t_i \), then \( E[g(Y)] \leq E[g(Z)] \) 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

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

(65)

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

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

(66)

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

(67)

Here the first two inequalities are Jensen while the last is by the first condition in (65) together with the monotonicity of \( g \). Adding these together yields \( g(z_2) + g(z_1) \geq g(y_2) + g(y_1) \), or \( E[g(Z)] \geq E[g(Y)] \), 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 \( E[g(Y)] \leq E[g(Z)] \) trivially holds by monotonicity of \( g \). This completes the proof.

**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:

\[
20
\]
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

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

(68)

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

$$
y_1 = a_- - \phi(b_-),
$$

$$
y_2 = a_+ + \phi(b_+),
$$

$$
z_1 = \max(a_- - b_- + a_+ - b_+),
$$

$$
z_2 = \max(a_- + b_-, a_+ + b_+).
$$

(69) (70) (71) (72)

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

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

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

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

(73) (74) (75)

$$
\vdots
$$

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

$$
\leq \mathbb{E}_{\epsilon, n} [g(\sup_{t \in T} \sum_{i=1}^{n} \epsilon_i t_i)],
$$

(76) (77) (78)

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 $\mu_\psi = \mathbb{E}_{X \sim p^*} [\psi(|X - \mu|/\sigma)]$, and note that $\mu_\psi \leq \mu_{\psi} \leq 1$. Now, by symmetrization we have

$$
\mathbb{E}_{X_1, \ldots, X_n \sim p^*} \left[ \left\| \frac{1}{n} \sum_{i=1}^{n} \psi \left( \frac{|X_i - \mu|}{\sigma} \right) - \mu_\psi \right\|^k \right]
$$

$$
\leq \mathbb{E}_{X, X' \sim p, \epsilon} \left[ \left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \psi \left( \frac{|X_i - \mu|}{\sigma} \right) - \psi \left( \frac{|X_i - \mu|}{\sigma} \right) \right\|^k \right]
$$

$$
\leq 2^k \mathbb{E}_{X \sim p, \epsilon} \left[ \left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \psi \left( \frac{|X_i - \mu|}{\sigma} \right) \right\|^k \right].
$$

(79) (80) (81)
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 \( E \) is the same).

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

\[
E_{X \sim p, \epsilon} \left[ \sup_{\|v\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \tilde{\psi} \left( \frac{|X_i - \mu, v|}{\sigma} \right) \right]^k \leq \left( \frac{L}{\sigma} \right)^k E_{X \sim p, \epsilon} \left[ \sup_{\|v\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \langle X_i - \mu, v \rangle \right]^k \]

\[
= \left( \frac{L}{\sigma} \right)^k E_{X \sim p, \epsilon} \left[ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \langle X_i - \mu \rangle \right]^k, \quad \text{(82)}
\]

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

\[
A_k \|z\|_2 \leq E_{n} \left[ \sum_i \epsilon_i z_i \right]^{1/k} \leq B_k \|z\|_2,
\]

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

\[
E_{X, \epsilon} [\| \sum_{i=1}^{n} \epsilon_i (X_i - \mu) \|_2^k] \leq O(1)^k E_{X, \epsilon, \epsilon'} [\| \sum_{i=1}^{n} \epsilon_i (X_i - \mu, \epsilon') \|_2^k]
\]

\[
\leq O(\sqrt{k})^k E_{X, \epsilon} [\| \sum_{i=1}^{n} (X_i - \mu, \epsilon') \|^2]^{k/2}. \quad \text{(86)}
\]

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 \( \left( \sum_i \langle X_i - \mu, \epsilon' \rangle^2 \right)^{k/2} \) expands out to \( k^{k/2} \) distinct terms, each of which is of the form \( E_{X, \epsilon} [\prod_{i=1}^{n} (X_i - \mu, \epsilon'')] \). Conditioning on \( \epsilon' \) and taking expectation over the \( X_i \), each term is bounded by \( \sigma^{k/2} \| \epsilon' \|^2/2 \), since the \( X_i \) each have bounded \( k \)th moment and are independent. Additionally, we have \( \| \epsilon' \|^2 = d^{k/2} \) since \( \epsilon' \) is a sign vector. Putting this together and plugging back into the inequalities above yields

\[
E_{X, \epsilon} [\| \sum_{i=1}^{n} \epsilon_i (X_i - \mu) \|_2^k] \leq O(\sqrt{k})^k \cdot \sigma^{k/2} \cdot d^{k/2}.
\]

Plugging back into (86) bounds the symmetrized truncated moments by \( O(\sqrt{kd/L})^k \), and plugging back into (81) completes the proof. \( \square \)

**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 \( G_{\text{cov}}(\sigma) = \{ \rho \mid \| E[(X - \mu)(X - \mu)^T] \| \leq \sigma^2 \} \), for \( \sigma^2 = O(1 + d/\log(1/\delta)/n) \). 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:

\[
G_{\text{iso}}(\sigma_1, \sigma_2) \overset{\text{def}}{=} \{ \rho \mid \| E[X - \mu] \|_2 \leq \sigma_1 \text{ and } \| E[(X - \mu)(X - \mu)^T - I] \| \leq (\sigma_2)^2 \}, \quad \text{whenever } r \leq \frac{p}{1 - \epsilon}. \quad (88)
\]

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

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

and

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

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 $p^* \in G'$ for $G' = G_{sao}(O(\epsilon \sqrt{\log(1/\epsilon)}), O(\sqrt{\epsilon \log(1/\epsilon)}))$, at least for large enough $n$. Furthermore, $G' \subseteq M$ for $M = G_{cov}(1 + O(\epsilon \log(1/\epsilon)))$.

Now we bound the generalized modulus of continuity:

Lemma 2.37. Suppose that $p \in G_{sao}(\sigma_1, \sigma_2)$ and $q \in G_{cov}(\sqrt{1 + \sigma_2^2})$, 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 $\|\mu(r) - \mu(p)\|_2$ is already bounded since $p \in G_{sao}$). We have that

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

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

$$\geq (1 - \epsilon)(\text{Cov}_r[X] + (\mu_q - \mu_r)(\mu_q - \mu_r)^\top) + \epsilon (\mu_q - \mu_{q'})(\mu_q - \mu_{q'})^\top.$$ (93)

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

$$\text{Cov}_q[X] \geq (1 - \epsilon)(\text{Cov}_r[X] + (1/\epsilon)(\mu_q - \mu_r)(\mu_q - \mu_r)^\top).$$ (94)

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 $\|\mu_r - \mu_q\|_2^2 \leq \epsilon(2\sigma_2^2 + \epsilon + \sigma_2^2)$, and so $\|\mu_r - \mu_q\|_2 \leq O(\epsilon + \sigma_2 \sqrt{\epsilon})$, which gives the desired result. \qed

In conclusion, projecting onto $G_{cov}(1 + O(\epsilon \log(1/\epsilon)))$ 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 $G_{cov}$ 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 $G$ and $M$ under distances $\text{TV}$ and $\tilde{\text{TV}}$. Here we will consider how to approximately project onto the set $\tilde{G}_{cov}(\sigma)$, 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 $\text{Cov}_{\tilde{p}}[X]$ 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 \( \epsilon \)-delete of \( \hat{p} \).

Before we assumed that \( \text{TV}(p^*, \hat{p}) \leq \epsilon \), but taking \( p' = \frac{\min(p^*, \hat{p})}{1 - \text{TV}(p^*, \hat{p})} \), we have \( p' \leq \frac{\hat{p}}{1-\epsilon} \) and \( \|\text{Cov}_{p'}[X]\| \leq \frac{\sigma^2}{1-\epsilon} \leq 2\sigma^2 \) whenever \( \|\text{Cov}_p[X]\| \leq \sigma^2 \). 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

\[
\|\mu(q) - \mu(p^*)\|^2 \leq O(\max(\|\text{Cov}_q[X]\|, \|\text{Cov}_{p^*}[X]\|) \cdot \text{TV}(p^*, q)) = O(\sigma^2 \epsilon). \tag{95}
\]

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 \( q \) 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 \( c_i / \sum_{i=1}^n c_i \) 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, \( \hat{\mu}_c \overset{\text{def}}{=} \frac{\sum_{i=1}^n c_i x_i}{\sum_{i=1}^n c_i} \).
4. Compute the empirical covariance \( \hat{\Sigma}_c \overset{\text{def}}{=} \frac{\sum_{i=1}^n c_i (x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^\top}{\sum_{i=1}^n c_i} \).
5. Let \( v \) be the maximum eigenvector of \( \hat{\Sigma}_c \), and let \( \hat{\sigma}_c^2 = v^\top \hat{\Sigma}_c v \).
6. If \( \hat{\sigma}_c^2 \leq 20\sigma^2 \), output \( q(c) \).
7. Otherwise, let \( \tau_i = \langle x_i - \hat{\mu}_c, v \rangle^2 \), and update \( c_i \leftarrow c_i \cdot (1 - \tau_i/\tau_{\max}) \), where \( \tau_{\max} = \max_i \tau_i \).
8. Go back to line 3.

The factor \( \tau_{\max} \) in the update \( c_i \leftarrow c_i \cdot (1 - \tau_i/\tau_{\max}) \) 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 \( \hat{\sigma}_c^2 \) 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 \( \tilde{\tau}_i = \langle x_i - \mu, v^* \rangle^2 \), and imagine for now that \( \tau_i \approx \tilde{\tau}_i \). We know that the average of \( \tilde{\tau}_i \) over the good points is at most \( \sigma^2 \), since \( \tilde{\tau}_i \) 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 \( \tilde{\tau}_i \approx \tau_i \) is basically an assumption that \( \mu \approx \hat{\mu}_c \) (which is what we are trying to show in the first place!).
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 $\hat{\mu}$, 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. \hspace{1em} (I_1)

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

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

(\hspace{1em} (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

$$\sum_{i \in T} 1 - c'_i = \sum_{i \in T} (1 - c_i) + \sum_{i \in T} (c_i - c'_i) = \sum_{i \in T} (1 - c_i) + \frac{1}{\tau_{\text{max}}} \sum_{i \in T} c_i \tau_i.$$  \hspace{1em} (96)

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

$$\sum_{i \in S} c_i r_i = \sum_{i \in S} c_i (x_i - \hat{\mu}_c, v^*)^2 \leq \sum_{i \in S} (x_i - \hat{\mu}_c, v^*)^2 = (1 - \epsilon)n E_p^*[(x_i - \hat{\mu}_c, v^*)^2]$$

$$= (1 - \epsilon)n E_p^*[(x_i - \hat{\mu}_c, v^*)^2] = (1 - \epsilon)n \cdot (v^*)^T (\text{Cov}_{p^*}[X] + (\mu - \hat{\mu}_c)(\mu - \hat{\mu}_c)^T)(v^*)$$

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

Here the second-to-last step uses the fact that for any $\theta$, $E[(X - \theta)(X - \theta)^T] = \text{Cov}[X] + (\theta - \mu)(\theta - \mu)^T$. Next note that $\|\text{Cov}_{p^*}\| \leq \sigma^2$ while $\|\mu - \hat{\mu}_c\|_2^2 \leq \frac{8\epsilon}{2\epsilon - 1} \sigma^2$ by the modulus of continuity bound combined with the fact that $p^*, q(c) \in \mathcal{G}_{\text{cov}}(\theta)$ and $\text{TV}(p^*, q(c)) \leq \frac{\epsilon}{1 - \epsilon}$. Therefore, we have

$$\sum_{i \in S} c_i r_i \leq (1 - \epsilon)\sigma^2 n + \frac{8\epsilon(1 - \epsilon)}{1 - 2\epsilon} \hat{\sigma}_c^2 n. \quad (102)$$

On the other hand, we have

$$\sum_{i=1}^n c_i r_i = (\sum_{i=1}^n c_i)\|\text{Cov}_{q(c)}[X]\| = (\sum_{i=1}^n c_i)\hat{\sigma}_c^2 \geq (1 - 2\epsilon)\hat{\sigma}_c^2 n, \quad (103)$$

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

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

which upon re-arrangement yields

$$\hat{\sigma}_c^2 \geq \frac{2(1 - \epsilon)(1 - 2\epsilon)}{1 - 12\epsilon + 12\epsilon^2} \sigma^2. \quad (105)$$

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 $\frac{55}{3} < 20$). 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 $\text{TV}(p^*, q(c)) \leq \frac{\epsilon}{1 - \epsilon}$. 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

$$\text{TV}(p^*, q(c)) = \sum_{i \in S} \max \left( \frac{c_i}{(1 - \beta)n} - \frac{1}{(1 - \epsilon)n}, 0 \right) + \sum_{i \notin S} \frac{c_i}{(1 - \beta)n}. \quad (106)$$

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

$$\text{TV}(p^*, q(c)) = \sum_{i \in S} \max \left( \frac{1}{(1 - \epsilon)n} - \frac{c_i}{(1 - \beta)n}, 0 \right) \quad (107)$$

$$= \frac{1}{(1 - \epsilon)(1 - \beta)n} \sum_{i \in S} \max((1 - \beta)(1 - c_i) + (\epsilon - \beta)c_i, 0). \quad (108)$$

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

[ Lecture 9]
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

$$\|u\|_* = \sup_{\|v\|_2 \leq 1} \langle u,v \rangle.$$  \hspace{1cm} (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 $G_{\text{cov}}(\sigma, \| \cdot \|)$ denote the family of distributions satisfying $\max_{\|v\|_\sigma \leq 1} v^T \text{Cov}_p[X] v \leq \sigma^2$. Then $G_{\text{cov}}$ is resilient exactly analogously to the $\ell_2$-case:

**Proposition 2.42.** If $p \in G_{\text{cov}}(\sigma, \| \cdot \|)$ 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 $G_{\text{cov}}(\sigma, \| \cdot \|)$ are $(\sigma, O(\sigma \sqrt{\epsilon}))$-resilient.

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

When $p^* \in G_{\text{cov}}(\sigma, \| \cdot \|)$, 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 $\max_{\|v\|_\sigma \leq 1} v^T \Sigma c v$ 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

$$\langle M, \Sigma \rangle \geq \sup_{\|v\|_\sigma \leq 1} v^T \Sigma v, \text{ and } \langle M, \Sigma' \rangle \leq \kappa \sup_{\|v\|_\sigma \leq 1} v^T \Sigma' v \text{ for all } \Sigma' \succeq 0.$$  \hspace{1cm} (110)

Thus a $\kappa$-approximate oracle over-approximates $\langle vv^T, \Sigma \rangle$ for the maximizing vector $v$ on $\Sigma$, and it underapproximates $\langle vv^T, \Sigma' \rangle$ 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 $\hat{\Sigma}_c \overset{\text{def}}{=} \sum_{i=1}^n c_i (x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^T / \sum_{i=1}^n c_i$.
4: Let $M = A(\hat{\Sigma}_c)$ be the output of a $\kappa$-approximate oracle.
5: If $\langle M, \hat{\Sigma}_c \rangle \leq 20\kappa \sigma^2$, output $\hat{c}(c)$.
6: Otherwise, let $\tau_i = (x_i - \hat{\mu}_c)^T M (x_i - \hat{\mu}_c)$, and update $c_i \leftarrow c_i \cdot (1 - \tau_i / \tau_{\max})$, where $\tau_{\max} = \max_i \tau_i$.
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 $A$.  

27
Semidefinite programming. As a concrete example, suppose that we wish to estimate $\mu$ in the $\ell_1$-norm $\|v\| = \sum_{j=1}^d |v_j|$. The dual norm is the $\ell_\infty$-norm, and hence our goal is to approximately solve the optimization problem

$$\text{maximize } v^T \Sigma v \text{ subject to } \|v\|_\infty \leq 1.$$  \hspace{1cm} (111)

The issue with (111) is that it is not concave in $v$ because of the quadratic function $v^T \Sigma v$. However, note that $v^T \Sigma v = \langle \Sigma, vv^T \rangle$. Therefore, if we replace $v$ with the variable $M = vv^T$, then we can re-express the optimization problem as

$$\text{maximize } \langle \Sigma, M \rangle \text{ subject to } M_{ij} \leq 1 \text{ for all } j, \ M \succeq 0, \ \text{rank}(M) = 1.$$  \hspace{1cm} (112)

Here the first constraint is a translation of $\|v\|_\infty \leq 1$, while the latter two constrain $M$ to be of the form $vv^T$. This is almost convex in $M$, except for the constraint $\text{rank}(M) = 1$. If we omit this constraint, we obtain the optimization

$$\text{maximize } \langle \Sigma, M \rangle \\text{ subject to } M_{ij} = 1 \text{ for all } j, \ M \succeq 0.$$  \hspace{1cm} (113)

Note that here we replace the constraint $M_{ij} \leq 1$ with $M_{ij} = 1$; this can be done because the maximizer of (113) will always have $M_{ij} = 1$ for all $j$. For brevity we often write this constraint as $\text{diag}(M) = 1$.

The problem (113) is a special instance of a *semidefinite program* and can be solved in polynomial time (in general, a semidefinite program allows arbitrary linear inequality or positive semidefinite constraints between linear functions of the decision variables; we discuss this more below).

The optimizer $M^*$ of (113) will always satisfy $\langle \Sigma, M^* \rangle \geq \sup_{\|v\|_\infty \leq 1} v^T \Sigma v$ because and $v$ with $\|v\|_\infty \leq 1$ yields a feasible $M$. The key is to show that it is not too much larger than this. This turns out to be a fundamental fact in the theory of optimization called Grothendieck’s inequality:

**Theorem 2.44.** If $\Sigma \succeq 0$, then the value of (113) is at most $\frac{\pi}{2} \sup_{\|v\|_\infty \leq 1} v^T \Sigma v$.

See Alon and Naor (2004) for a very well-written exposition on Grothendieck’s inequality and its relation to optimization algorithms. In that text we also see that a version of Theorem 2.44 holds even when $\Sigma$ is not positive semidefinite or indeed even square. Here we produce a proof based on [todo: cite] for the semidefinite case.

**Proof of Theorem 2.44.** The proof involves two key relations. To describe the first, given a matrix $X$ let $\arcsin[X]$ denote the matrix whose $ij$ entry is $\arcsin(X_{ij})$ (i.e. we apply $\arcsin$ element-wise). Then we have (we will show this later)

$$\max_{\|v\|_\infty \leq 1} v^T \Sigma v = \max_{X \geq 0, \text{diag}(X) = 1} \frac{2}{\pi} \langle \Sigma, \arcsin[X] \rangle.$$  \hspace{1cm} (114)

The next relation is that

$$\arcsin[X] \succeq X.$$  \hspace{1cm} (115)

Together, these imply the approximation ratio, because we then have

$$\max_{M \succeq 0, \text{diag}(M) = 1} \langle \Sigma, M \rangle \leq \max_{M \succeq 0, \text{diag}(M) = 1} \langle \Sigma, \arcsin[M] \rangle = \frac{\pi}{2} \sup_{\|v\|_\infty \leq 1} v^T \Sigma v.$$  \hspace{1cm} (116)

We will therefore focus on establishing (115) and (116).

To establish (115), we will show that any $X$ with $X \succeq 0$, $\text{diag}(X) = 1$ can be used to produce a probability distribution over vectors $v$ such that $\mathbb{E}[v^T \Sigma v] = \frac{2}{\pi} \langle \Sigma, \arcsin[X] \rangle$.

First, by Graham decomposition we know that there exist vectors $u_i$ such that $M_{ij} = \langle u_i, u_j \rangle$ for all $i, j$. In particular, $M_{ii} = 1$ implies that the $u_i$ have unit norm. We will then construct the vector $v$ by taking $v_i = \text{sign}(\langle u_i, g \rangle)$ for a Gaussian random variable $g \sim \mathcal{N}(0, 1)$.

We want to show that $\mathbb{E}_g[v_i v_j] = \frac{2}{\pi} \arcsin(\langle u_i, u_j \rangle)$. For this it helps to reason in the two-dimensional space spanned by $u_i$ and $v_j$. Then $v_i v_j = -1$ if the hyperplane induced by $g$ cuts between $u_i$ and $u_j$, and $+1$ if it does not. Letting $\theta$ be the angle between $u_i$ and $u_j$, we then have $\mathbb{P}[v_i v_j = -1] = \frac{\theta}{\pi}$ and hence

$$\mathbb{E}_g[v_i v_j] = (1 - \frac{\theta}{\pi}) - \frac{\theta}{\pi} = \frac{2}{\pi} (\frac{\pi}{2} - \theta) = \frac{2}{\pi} \arcsin(\langle u_i, u_j \rangle),$$  \hspace{1cm} (117)
as desired. Therefore, we can always construct a distribution over $v$ for which $E[v^\top \Sigma v] = \frac{2}{3} \langle \Sigma, \arcsin[M] \rangle$, hence the right-hand-side of (115) is at most the left-hand-side. For the other direction, note that the maximizing $v$ on the left-hand-side is always a $\{-1, +1\}$ vector by convexity of $v^\top \Sigma v$, and for any such vector we have $\frac{2}{3} \arcsin[vv^\top] = vv^\top$. Thus the left-hand-side is at most the right-hand-side, and so the equality (115) indeed holds.

We now turn our attention to establishing (116). For this, let $X^{\otimes k}$ denote the matrix whose $i,j$ entry is $X_{ij}^k$ (we take element-wise power). We require the following lemma:

**Lemma 2.45.** For all $k \in \{1,2,\ldots\}$, if $X \succeq 0$ then $X^{\otimes k} \succeq 0$.

**Proof.** The matrix $X^{\otimes k}$ is a submatrix of $X^{\otimes k}$, where $(X^{\otimes k})_{i_1\cdots i_k,j_1\cdots j_k} = X_{i_1,j_1} \cdots X_{i_k,j_k}$. We can verify that $X^{\otimes k} \succeq 0$ (its eigenvalues are $\lambda_i \cdots \lambda_k$ where $\lambda_i$ are the eigenvalues of $X$), hence so is $X^{\otimes k}$ since submatrices of PSD matrices are PSD.

With this in hand, we also make use of the Taylor series for $\arcsin(z)$: $\arcsin(z) = \sum_{n=0}^{\infty} \frac{(2n)!}{(2^n n!)^2} \frac{z^{2n+1}}{2n+1} = z + \frac{z^3}{3} + \cdots$. Then we have

$$\arcsin[X] = X + \sum_{n=1}^{\infty} \frac{(2n)!}{(2^n n!)^2} \frac{1}{2n+1} X^{\otimes (2n+1)} \succeq X,$$  \hspace{1cm} (118)

as was to be shown. This completes the proof. \qed

**Alternate proof (by Mihaela Curmei):** We can also show that $X^{\otimes k} \succeq 0$ more directly. Specifically, we will show that if $A, B \succeq 0$ then $A \circ B \succeq 0$, from which the result follows by induction. To show this let $A = \sum_i \lambda_i u_i u_i^\top$ and $B = \sum_j \nu_j v_j v_j^\top$ and observe that

$$A \circ B = (\sum_i \lambda_i u_i u_i^\top) \circ (\sum_j \nu_j v_j v_j^\top) \hspace{1cm} (119)$$

$$= \sum_{i,j} \lambda_i \nu_j (u_i u_i^\top) \circ (v_j v_j^\top) \hspace{1cm} (120)$$

$$= \sum_{i,j} \lambda_i \nu_j (u_i \circ v_j)(u_i \circ v_j)^\top \hspace{1cm} (121)$$

from which the claim follows. Here the key step is that for rank-one matrices the $\circ$ operation behaves nicely: $(u_i u_i^\top) \circ (v_j v_j^\top) = (u_i \circ v_j)(u_i \circ v_j)^\top$.

[Lecture 10]

### 2.8 Semidefinite Programming and Sum-of-Squares

In the previous subsection, we saw how to approximately solve $\max_{\|v\|_\infty \leq 1} v^\top \Sigma v$ via the semidefinite program defined by $\max_{M \succeq 0,\, \text{diag}(M) = 1} \langle M, \Sigma \rangle$. In this section we will cover semidefinite programming in more detail, and build up to *sum-of-squares programming*, which will be used to achieve error $O(\epsilon^{1-1/k})$ when $p^*$ has “certifiably bounded” $k$th moments (recall that we earlier achieved error $O(\epsilon^{1-1/k})$ for bounded $k$th moments but did not have an efficient algorithm).

A semidefinite program is an optimization problem of the form

$$\text{maximize } \langle A, X \rangle$$

subject to $X \succeq 0$,

$$\langle X, B_1 \rangle \leq c_1,$$

$$\vdots$$

$$\langle X, B_m \rangle \leq c_m.$$
Here $\langle X, Y \rangle = \text{tr}(X^T Y) = \sum_{ij} X_{ij} Y_{ij}$ is the inner product between matrices, which is the same as the elementwise dot product when considered as $n^2$-dimensional vectors.

Here the matrix $A$ specifies the objective of the program, while $(B_j, c_j)$ specify linear inequality constraints. We additionally have the positive semidefinite cone constraint that $X \succeq 0$, meaning that $X$ must be symmetric with only non-negative eigenvalues. Each of $A$ and $B_1, \ldots, B_m$ are $n \times n$ matrices while the $c_j$ are scalars. We can equally well minimize as maximize by replacing $A$ with $-A$.

While (122) is the canonical form for a semidefinite program, problems that are seemingly more complex can be reduced to this form. For one, we can add linear equality constraints as two-sided inequality constraints. In addition, we can replace $X \succeq 0$ with $L(X) \succeq 0$ for any linear function $L$, by using linear equality constraints to enforce the linear relations implied by $L$. Finally, we can actually include any number of constraints $\mathcal{L}_1(X) \succeq 0, \mathcal{L}_k(X) \succeq 0$, since this is e.g. equivalent to the single constraint $\begin{bmatrix} \mathcal{L}_1(X) & 0 \\ 0 & \mathcal{L}_2(X) \end{bmatrix}$ when $k = 2$. As an example of these observations, the following (arbitrarily-chosen) optimization problem is also a semidefinite program:

$$
\begin{align*}
\text{minimize} & \quad a^T x + \langle A_1, M \rangle + \langle A_2, Y \rangle \\
\text{subject to} & \quad M + Y \succeq \Sigma \\
& \quad \text{diag}(M) = 1 \\
& \quad \text{tr}(Y) \leq 1 \\
& \quad Y \succeq 0 \\
& \quad \begin{bmatrix} 1 & x^T \\ x & M \end{bmatrix} \succeq 0
\end{align*}
$$

(As a brief aside, the constraint $[1 \ x^T; x \ M] \succeq 0$ is equivalent to $xx^T \leq M$ which is in turn equivalent to $x^T M^{-1} x \leq 1$ and $M \succeq 0$.)

**Semidefinite constraints as quadratic polynomials.** An alternative way of viewing the constraint $M \succeq 0$ is that the polynomial $p_M(v) = v^T M v$ is non-negative for all $v \in \mathbb{R}^d$. More generally, if we have a non-homogeneous polynomial $p_{M,y,c}(v) = v^T M v + y^T v + c$, we have $p_{M,y,c}(v) \geq 0$ for all $v$ if and only if $M' \succeq 0$ for $M' = \begin{bmatrix} c & y^T/2 \\ y^T/2 & M \end{bmatrix} \succeq 0$.

This polynomial perspective is helpful for solving eigenvalue-type problems. For instance, $\|M\| \leq \lambda$ if and only if $v^T M v \leq \lambda \|v\|^2_2$ for all $v$, which is equivalent to asking that $v^T (\lambda M - M) v \geq 0$ for all $v$. Thus $\|M\|$ can be expressed as the solution to

$$
\begin{align*}
\text{minimize} & \quad \lambda \\
\text{subject to} & \quad \lambda I - M \succeq 0 \text{ (equivalently, } v^T (\lambda I - M) v \geq 0 \text{ for all } v) 
\end{align*}
$$

We thus begin to see a relationship between moments and polynomial non-negativity constraints.

**Higher-degree polynomials.** It is tempting to generalize the polynomial approach to higher moments. For instance, $M_4(p)$ denote the 4th moment tensor of $p$, i.e. the unique symmetric tensor such that 

$$
\langle M_4, v^\otimes 4 \rangle = \mathbb{E}_{x \sim p}[(x - \mu, v)^4].
$$

Note we can equivalently express $\langle M_4, v^\otimes 4 \rangle = \sum_{ijkl} (M_4)_{ijkl} v_i v_j v_k v_l$, and hence $(M_4)_{ijkl} = \mathbb{E}[(x_i - \mu)(x_j - \mu)(x_k - \mu)(x_l - \mu)]$.

A distribution $p$ has bounded 4th moment if and only if $\langle M, v^\otimes 4 \rangle \leq \lambda \|v\|^4_2$ for all $v$. Letting $p_M(v) \overset{\text{def}}{=} \langle M, v^\otimes 4 \rangle$, we thus can express the 4th moment of $p$ as the polynomial program

$$
\begin{align*}
\text{minimize} & \quad \lambda \\
\text{subject to} & \quad \lambda(v_1^2 + \cdots + v_d^2)^2 - p_M(v) \geq 0 \text{ for all } v \in \mathbb{R}^d
\end{align*}
$$

30
Unfortunately, in contrast to (123), (126) is NP-hard to solve in general. We will next see a way to approximate (126) via a technique called sum-of-squares programming, which is a way of approximately reducing polynomial programs such as (126) to a large but polynomial-size semidefinite program.

**Warm-up: certifying non-negativity over** \( \mathbb{R} \). Consider the one-dimensional polynomial

\[
q(x) = 2x^4 + 2x^3 - x^2 + 5
\]

(127)

Is it the case that \( q(x) \geq 0 \) for all \( x \)? If so, how would we check this?

What if I told you that we had

\[
q(x) = \frac{1}{2}(2x^2 + x - 3)^2 + \frac{1}{2}(3x + 1)^2
\]

(128)

Then, it is immediate that \( q(x) \geq 0 \) for all \( x \), since it is a (weighted) sum of squares.

How can we construct such decompositions of \( q \)? First observe that we can re-write \( q \) as the matrix function

\[
q(x) = \begin{bmatrix} 1 & x & x^2 \end{bmatrix}^\top \begin{bmatrix} 5 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}.
\]

(129)

On the other hand, the sum-of-squares decomposition for \( q \) implies that we can also write

\[
q(x) = \begin{bmatrix} 1 & x & x^2 \end{bmatrix}^\top \begin{bmatrix} -3 & 1 & 2 \\ 1 & -3 & 1 \\ 2 & 1 & 0 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 1 & 3 \\ 3 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ x^2 \end{bmatrix},
\]

(130)

i.e. we can decompose the matrix \( M \) defining \( q(x) = [1; x; x^2]^\top M [1; x; x^2] \) into a non-negative combination of rank-one outer products, which is true if and only if \( M \succeq 0 \).

There is one problem with this, which is that despite our successful decomposition of \( q \), \( M \) is self-evidently not positive semidefinite! (For instance, \( M_{22} = -1 \).) The issue is that the matrix \( M \) defining \( q(x) \) is not unique. Indeed, any \( M(a) = \begin{bmatrix} 5 & 0 & -a \\ 0 & 2a - 1 & 1 \\ -a & 1 & 2 \end{bmatrix} \) would give rise to the same \( q(x) \), and a sum-of-squares decomposition merely implies that \( M(a) \succeq 0 \) for some \( a \). Thus, we obtain the following characterization:

\[
q(x) \text{ is a sum of squares } \sum_{j=1}^k q_j(x)^2 \iff M(a) \succeq 0 \text{ for some } a \in \mathbb{R}.
\]

(131)

For the particular decomposition above we took \( a = 3 \).

**Sum-of-squares in two dimensions.** We can generalize the insights to higher-dimensional problems. Suppose for instance that we wish to check whether \( q(x, y) = a_{40}x^4 + a_{31}x^3y + a_{22}xy^2 + a_{13}y^3 + a_{04}y^4 + \ldots \) is non-negative for all \( x, y \). Again, this is hard-to-check, but we can hope to check the sufficient condition that \( q \) is a sum-of-squares, which we will express as \( q \succeq_{\text{sos}} 0 \). As before this is equivalent to checking that a certain matrix is positive semidefinite. Observe that

\[
q(x, y) = \begin{bmatrix} x^2 \\ xy \\ y^2 \\ x \\ y \\ 1 \end{bmatrix}^\top \begin{bmatrix} a_{40} & a_{31}/2 & -b & a_{30}/2 & -c & -b' \\ a_{31}/2 & a_{22} + 2b & a_{13}/2 & a_{21}/2 + c & -c' & -c'' \\ -b & a_{13}/2 & a_{04} & a_{21}/2 + c' & a_{03}/2 & -b'' \\ a_{30}/2 & a_{21}/2 + c & a_{21}/2 + c' & a_{20} + 2b' & a_{11}/2 + c'' & a_{10}/2 \\ -c & -c' & a_{03}/2 & a_{11}/2 + c'' & a_{02} + 2b'' & a_{01}/2 \\ -b' & -c'' & a_{03}/2 & a_{10}/2 & a_{01}/2 & a_{00} \end{bmatrix} \begin{bmatrix} x^2 \\ xy \\ y^2 \\ x \\ y \\ 1 \end{bmatrix}\]

(132)

for any \( b, b', b'', c, c', c'' \). Call the above expression \( M(b, b', b'', c, c', c'') \), which is linear in each of its variables. Then we have \( q \succeq_{\text{sos}} 0 \) if and only if \( M(b, b', b'', c, c', c'') \succeq 0 \) for some setting of the \( b \)s and \( c \)s.
Sum-of-squares in arbitrary dimensions. In general, if we have a polynomial \( q(x_1, \ldots, x_d) \) in \( d \) variables, which has degree \( 2t \), then we can embed it as some matrix \( M(b) \) (for decision variables \( b \) that capture the symmetries in \( M \) as above), and the dimensionality of \( M \) will be the number of monomials of degree at most \( t \) which turns out to be \( (d+t)^t = O((d+t)^t) \).

The upshot is that any constraint of the form \( q \sos 0 \), where \( q \) is linear in the decision variables, is a semidefinite constraint in disguise. Thus, we can solve any program of the form

\[
\begin{align*}
\text{maximize} & \quad c^T y \\
\text{subject to} & \quad q_1 \sos 0, \ldots, q_m \sos 0,
\end{align*}
\]

(133)

where the \( q_j \) are linear in the decision variables \( y \). (And we are free to throw in any additional linear inequality or semidefinite constraints as well.) We refer to such optimization problems as sum-of-squares programs, in analogy to semidefinite programs.

Sum-of-squares for \( k \)th moment. Return again to the \( k \)th moment problem. As a polynomial program we sought to minimize \( \lambda \) such that \( \lambda (v_1^2 + \cdots + v_d^2)^{k/2} - \langle M_{2k}, v^{\otimes 2k} \rangle \) was a non-negative polynomial. It is then natural to replace the non-negativity constraint with the constraint that \( \lambda \|v\|_2^2 - \langle M_{2k}, v^{\otimes 2k} \rangle \sos 0 \).

However, we actually have a bit more flexibility and it turns out that the best program to use is

\[
\begin{align*}
\text{minimize} & \quad \lambda \\
\text{subject to} & \quad \lambda \|v\|_2^2 - \langle M_{2k}, v^{\otimes 2k} \rangle + (\|v\|_2^2 - 1)q(v) \sos 0 \text{ for some } q \text{ of degree at most } 2k - 2
\end{align*}
\]

(134)

Note that the family of all such \( q \) can be linearly parameterized and so the above is indeed a sum-of-squares program. It is always at least as good as the previous program because we can take \( q(v) = \lambda (1 + \|v\|_2^2 + \cdots + \|v\|_2^{2k-2}) \).

When the solution \( \lambda^* \) to (134) is at most \( \sigma^{2k} \) for \( M_{2k}(p) \), we say that \( p \) has \( 2k \)th moment certifiably bounded by \( \sigma^{2k} \). In this case a variation on the filtering algorithm achieves error \( O(\sqrt{\epsilon^{-1/2k}}) \). We will not discuss this in detail, but the main issue we need to resolve to obtain a filtering algorithm is to find some appropriate tensor \( T \) such that \( \langle T, M_{2k} \rangle = \lambda^* \) and \( T \) "looks like" the expectation of \( v^{\otimes 2k} \) for some probability distribution over the unit sphere. Then we can filter using \( \tau_T = \langle T, (x_i - \mu)^{\otimes 2k} \rangle \).

To obtain \( T \) requires computing the dual of (134), which requires more optimization theory than we have assumed from the reader, but it can be done in polynomial time. We refer to the corresponding \( T \) as a pseudomoment matrix. Speaking very roughly, \( T \) has all properties of a moment matrix that can be "proved using only sum-of-squares inequalities", which includes all properties that we needed for the filtering algorithm to work. We will henceforth ignore the issue of \( T \) and focus on assumptions on \( p \) that ensure that \( M_{2k}(p) \) is certifiably bounded. The main such assumption is the Poincaré inequality, which we cover in the next section.

[ Lecture 11 ]

2.9 Sum-of-Squares Certifiably from the Poincaré inequality

We now turn our attention to bounding the value of (134). Ignoring finite-sample issues, our goal is to identify assumptions on \( p \) such that \( M_{2k}(p) \defeq \mathbb{E}_{X \sim p}[(X - \mu)^{\otimes 2k}] \) yields a small value for (134).

Before doing so, we will introduce some machinery for establishing bounds on (134). The main idea is that of a sum-of-squares proof:

Definition 2.46. A polynomial inequality \( p(v) \leq q(v) \) has a sum-of-squares proof if \( q(v) - p(v) \sos 0 \). We will also denote this as \( q(v) \sos p(v) \) or \( p(v) \sos q(v) \).

The usefulness of this perspective is that the relation \( \sos \) satisfies many of the same properties as \( \leq \):

- If \( p_1 \sos p_2 \) and \( p_2 \sos p_3 \), then \( p_1 \sos p_3 \).
• If $p_1 \preceq_{\text{sos}} q_1$ and $p_2 \preceq_{\text{sos}} q_2$, then $p_1 + p_2 \preceq_{\text{sos}} q_1 + q_2$.
• If $p_1 \preceq_{\text{sos}} 0$ and $p_2 \preceq_{\text{sos}} 0$, then $p_1 p_2 \preceq_{\text{sos}} 0$.
• If $p_1 \preceq_{\text{sos}} p_2$, $q_1 \preceq_{\text{sos}} q_2$, and $p_2, q_1 \preceq_{\text{sos}} 0$, then $p_1 q_1 \preceq_{\text{sos}} p_2 q_2$.

Moreover, many “standard” inequalities such as Cauchy-Schwarz and Hölder have sum-of-squares proofs. Using these, we can often turn a normal proof that $p \leq q$ into a sum-of-squares proof that $p \leq q$ as long as we give sum-of-squares proofs for a small number of key steps.

For concreteness, we will prove the last two claims properties above. We first prove that $p_1, p_2 \preceq_{\text{sos}} 0 \implies p_1 p_2 \preceq_{\text{sos}} 0$. Indeed we have

$$p_1(v)p_2(v) = (\sum_i p_{1i}(v)^2)(\sum_j p_{2j}(v)^2) = \sum_{ij} (p_{1i}(v)p_{2j}(v))^2 \preceq_{\text{sos}} 0$$

(135)

Next we prove that $p_1 \preceq_{\text{sos}} p_2, q_1 \preceq_{\text{sos}} q_2$, and $p_2, q_1 \preceq_{\text{sos}} 0$ implies $p_1 q_2 \preceq_{\text{sos}} p_2 q_2$. This is because

$$p_2 q_2 - p_1 q_1 = p_2 (q_2 - q_1) + (p_2 - p_1) q_1 \preceq_{\text{sos}} 0,$$

(136)

where the second relation uses $p_2, q_2 - q_1 \preceq_{\text{sos}} 0$ and $p_2 - p_1, q_1 \preceq_{\text{sos}} 0$ together with the previous result.

In view of this, we can reframe bounding (134) as the following goal:

**Goal:** Find a sum-of-squares proof that $\langle M_{2k}(p), v^{\otimes 2k} \rangle \preceq_{\text{sos}} \lambda \|v\|_2^{2k}$.

**Certifiability for Gaussians.** We now return to the assumptions needed on $p$ that will enable us to provide the desired sum-of-squares proof. Let us start by observing that a sum-of-squares proof exists for any Gaussian distribution: If $p = N(\mu, \Sigma)$, then

$$\langle M_{2k}(N(\mu, \Sigma)), v^{\otimes 2k} \rangle = \langle M_{2k}(N(0, I)), (\Sigma^{1/2} v)^{\otimes 2k} \rangle$$

(137)

$$= \left( \prod_{i=1}^k (2i - 1) \right) \|\Sigma^{1/2} v\|_2^{2k}$$

(138)

$$\leq (2k)^k \|\Sigma\|^{k/2} \|v\|_2^{2k},$$

(139)

so we may take $\lambda = (2k\|\Sigma\|)^k$. (Here $I$ denotes the identity tensor that is 1 along the diagonal and zero elsewhere.) Therefore normal distributions have certifiably bounded moments, but the proof above heavily exploited the rotational symmetry of a normal distribution. We can provide similar proofs for other highly symmetric distributions (such as the uniform distribution on the hypercube), but these are unsatisfying as they only apply under very specific distributional assumptions. We would like more general properties that yield certifiably bounded moments.

**Poincaré inequality.** The property we will use is the *Poincaré inequality*. A distribution $p$ on $\mathbb{R}^d$ is said to satisfy the Poincaré inequality with parameter $\sigma$ if

$$\text{Var}_{x \sim p}[f(x)] \leq \sigma^2 \mathbb{E}_x[p(\|\nabla f(x)\|_2^2)]$$

(141)

for all differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$. This is a “global to local property”–it says that for any function that for any function $f$ that varies under $p$, that variation can be picked up in terms of local variation (the gradient). In particular, it says that $p$ doesn’t have any “holes” (regions with low probability density that lie between two regions of high probability density). Indeed, suppose that $A$ and $B$ were two disjoint convex regions with $p(A) = p(B) = \frac{1}{2}$. Then $p$ cannot satisfy the Poincaré inequality with any constant, since there is a function that is 1 on $A$, 0 on $B$, and constant on both $A$ and $B$.

Below are some additional examples and properties of Poincaré distributions:
• A one-dimensional Gaussian $N(\mu, \sigma^2)$ is Poincaré with constant $\sigma$.

• If $p$, $q$ are $\sigma$-Poincaré then their product $p \times q$ is $\sigma$-Poincaré. In particular a multivariate Gaussian $N(\mu, \sigma^2 I)$ is $\sigma$-Poincaré.

• If $X \sim p$ is $\sigma$-Poincaré and $A$ is a linear map, then $AX$ is $(\sigma \| A \|)$-Poincaré. In particular, $aX_1 + aX_2$ is $(\sqrt{a^2 + b^2} \sigma)$-Poincaré when $X_1$ and $X_2$ are both $\sigma$-Poincaré, and $N(\mu, \Sigma)$ is $\| \Sigma \|^{1/2}$-Poincaré.

• More generally, if $X \sim p$ is $\sigma$-Poincaré and $f$ is $L$-Lipschitz, then $f(X)$ is $(\sigma L)$-Poincaré.

Together these imply that Poincaré distributions contain multivariate Gaussians, arbitrary Lipschitz functions of Gaussians, and independent sums of such distributions. The above properties (except the initial Gaussian property) are all straightforward computations. Let us next state two substantially deeper results:

• If $p$ is $\sigma$-strongly log-concave (meaning that the log-probability density $\log p(x)$ satisfies $\nabla^2 \log p(x) \preceq -\frac{1}{\sigma^2} I$), then $p$ is $\sigma$-Poincaré (Bakry and Émery, 1985).

• Suppose that the support of $X \sim p$ has $\ell_2$-radius at most $R$, and let $Z = N(0, \tau^2 I)$ for $\tau \geq 2R$. Then $X + Z$ is $(\tau \sqrt{\epsilon})$-Poincaré (Bardet et al., 2018).

Thus Poincaré encompasses all strongly log-concave densities, and effectively any product of bounded random variables (after adding Gaussian noise, which we can always do ourselves).

It is instructive to compare Poincaré to the sub-Gaussian property that we have so far relied on. Poincaré is neither strictly stronger or weaker than sub-Gaussian, but it is stronger than sub-exponential (we will see this below). In general, we should think of Poincaré as being substantially stronger than sub-exponential: it implies that not only is the distribution itself sub-exponential, but so is any Lipschitz function of the density.

As an example, consider the random variable $(X, Y) \in \mathbb{R}^d$ where $X \sim N(0, I)$ and $Y = \epsilon X$ for a Rademacher random variable $\epsilon$. Then $(X, Y)$ is sub-Gaussian, but not Poincaré with good constant: if we take $f(X, Y) = \sum_i X_i Y_i$, then $f$ is with high probability close to either $+d$ or $-d$, so $\Var[f(X, Y)] \approx d^2$. However, $\nabla f(X, Y) = (Y_1, \ldots, Y_d, X_1, \ldots, X_d)$ and so $\|\nabla f(X, Y)\|_2^2$ is close to $2d$ with high probability. Thus while the sub-Gaussian constant is $O(1)$, the Poincaré constant in this case is $\Omega(\sqrt{d})$.

Consequences of Poincaré. So far we have seen conditions that imply Poincaré, but we would also like to derive consequences of this property. Below are some of the most useful ones:

• If $X \sim p$ is $\sigma$-Poincaré, then Lipschitz functions concentrate: $\Pr[|f(x) - \mathbb{E}[f(x)]| \geq l] \leq 6 \exp(-l/(\sigma L))$ for any $L$-Lipschitz $f$.

• As a corollary, we have volume expansion: For any set $A$, let $A_\epsilon$ be the set of points within $\ell_2$-distance $\epsilon$ of $A$. Then $p(A)p(A_\epsilon) \leq 36 \exp(-\epsilon/\sigma)$.

This second property implies, for instance, that if $p(A) \geq \delta$, then almost all points will be within distance $O(\sigma \log(1/\delta))$ of $A$.

To prove the second property, let $f(x) = \min_{y \in A} \|x - y\|_2$. Then $f$ is Lipschitz, is 0 on $A$, and is $\epsilon$ on $A_\epsilon$. Let $\mu$ be the mean of $f(X)$. Since $f$ is sub-exponential we have $p(A) = p(f(X) = 0) \leq 6 \exp(-\mu/\sigma)$, and $p(A_\epsilon) = p(f(X) = \epsilon) \leq 6 \exp(-\epsilon/\sigma)/\sigma$. Multiplying these together yields the claimed result.

The most important property for our purposes, however, will be the following:

**Theorem 2.47.** Suppose that $p$ is $\sigma$-Poincaré and let $f$ be a differentiable function such that $\mathbb{E}[\nabla^j f(X)] = 0$ for $j = 1, \ldots, k - 1$. Then there is a universal constant $C_k$ such that $\Var[f(X)] \leq C_k \sigma^2 k \mathbb{E}[\|\nabla^k f(X)\|_F^2]$.

Note that $k = 1$ is the original Poincaré property, so we can think of Theorem 2.47 as a generalization of Poincaré to higher derivatives. Note also that $\nabla^k f(X)$ is a tensor in $\mathbb{R}^{d^k}$; the notation $\|\nabla^k f(X)\|_F^2$ denotes the squared Frobenius norm of $\nabla^k f(X)$, i.e., the sum of the squares of its entries.

Theorem 2.47, while it may appear to be a simple generalization of the Poincaré property, is a deep result that was established in Adamczak and Wolff (2015), building on work of Latala (2006). We will use Theorem 2.47 in the sequel to construct our sum-of-squares proofs.
Sum-of-squares proofs for Poincaré distributions. Here we will construct sum-of-squares proofs that $M_k(v) \stackrel{\text{def}}{=} \mathbb{E}_p[(x - \mu, v)^{2k}] \preceq_{\text{sos}} C'_k \sigma^{2k} \|v\|_2^{2k}$ whenever $p$ is $\sigma$-Poincaré, for some universal constants $C'_k$. We will exhibit the proof for $k = 1, 2, 3$ (the proof extends to larger $k$ and the key ideas appear already by $k = 3$). We introduce the notation

$$M_k = \mathbb{E}[(x - \mu)^{\otimes k}],$$

$$M_k(v) = \langle M_k, v^{\otimes k} \rangle = \mathbb{E}[(x - \mu, v)^k].$$

\textbf{Proof for $k = 1$.} We wish to show that $\mathbb{E}_p[(x - \mu, v)^2] \preceq_{\text{sos}} \sigma^2 \|v\|_2^2$. To do this take $f_v(x) = \langle x, v \rangle$. Then the Poincaré inequality applied to $f_v$ yields

$$\mathbb{E}_p[(x - \mu, v)^2] = \text{Var}[f_v(x)] \leq \sigma^2 \mathbb{E}[\|\nabla f_v(x)\|_2^2] = \sigma^2 \mathbb{E}[\|v\|_2^2] = \sigma^2 \|v\|_2^2.$$

Thus $M_2(v) \leq \sigma^2 \|v\|_2^2$ (this is just saying that Poincaré distributions have bounded covariance). This property has a sum-of-squares proof because it is equivalent to $\sigma^2 I - M_2 \succeq 0$, and we know that all positive semidefiniteness relations are sum-of-squares certifiable.

\textbf{Proof for $k = 2$.} Extending to $k = 2$, it makes sense to try $f_v(x) = \langle x - \mu, v \rangle^2$. Then we have $\nabla f_v(x) = 2(x - \mu, v)v$ and hence $\mathbb{E}[\nabla f_v(x)] = 0$. We also have $\nabla^2 f_v(x) = 2v \otimes v$. Thus applying Theorem 2.47 we obtain

$$\text{Var}[f_v(x)] \leq C_2 \sigma^4 \mathbb{E}[\|2v \otimes v\|_2^2] = 4C_2 \sigma^4 \|v\|_2^4.$$

We also have $\text{Var}[f_v(x)] = \mathbb{E}[\|x - \mu, v\|_4^4] - \mathbb{E}[(x - \mu, v)^2]^2 = M_4(v) - M_2(v)^2$. Thus

$$M_4(v) = (M_4(v) - M_2(v)^2) + M_2(v)^2 \leq 4C_2 \sigma^4 \|v\|_2^2 + \sigma^4 \|v\|_2^4 = (4C_2 + 1) \sigma^4 \|v\|_2^4.$$

This shows that the fourth moment is bounded, but how can we construct a sum-of-squares proof? We already have that $M_2(v)^2 \preceq_{\text{sos}} \sigma^4 \|v\|_2^4$ (by $0 \preceq_{\text{sos}} M_2(v) \preceq_{\text{sos}} \sigma^2 \|v\|_2^2$ and the product property). Therefore we focus on bounding $M_4(v) - M_2(v)^2 = \text{Var}[f_v(x)]$.

For this we will apply Theorem 2.47 to a modified version of $f_v(x)$. For a matrix $A$, let $f_A(x) = (x - \mu)^\top A(x - \mu) = \langle A, (x - \mu)^{\otimes 2} \rangle$. Then $f_v(x) = f_A(x)$ for $A = vv^\top$. By the same calculations as above we have $\mathbb{E}[\nabla f_A(x)] = 0$ and $\nabla^2 f_A(x) = 2A$. Thus by Theorem 2.47 we have

$$\text{Var}[f_A(x)] \leq C_2 \sigma^4 \mathbb{E}[\|2A\|_F^2] = 4C_2 \sigma^4 \|A\|_F^2.$$

On the other hand, we have $\text{Var}[f_A(x)] = (M_4, A \otimes A) - (M_2, A)^2 = (M_4 - M_2 \otimes M_2, A \otimes A)$. Thus (148) implies that

$$M_4 - M_2 \otimes M_2, A \otimes A \leq 4C_2 \sigma^4 \|A\|_F^2.$$

Another way of putting this is that $M_4 - M_2 \otimes M_2$, when considered as a matrix in $\mathbb{R}^{d^2 \times d^2}$, is smaller than $4C_2 \sigma^4 I$ in the semidefinite ordering. Hence $4C_2 \sigma^4 I - (M_4 - M_2 \otimes M_2) \succeq 0$ and so $4C_2 \sigma^4 \|v\|_2^4 - (M_4 - M_2 \otimes M_2, v^{\otimes 4}) \preceq_{\text{sos}} 0$, giving us our desired sum-of-squares proof. To recap, we have:

$$M_4(v) = (M_4(v) - M_2(v)^2) + M_2(v)^2 \preceq_{\text{sos}} 4C_2 \sigma^4 \|v\|_2^4 + \sigma^4 \|v\|_2^4 = (4C_2 + 1) \sigma^4 \|v\|_2^4,$$

so we can take $C'_2 = 4C_2 + 1$.

\textbf{Proof for $k = 3$.} Inspired by the $k = 1, 2$ cases, we try $f_v(x) = \langle x - \mu, v \rangle^3$. However, this choice runs into problems, because $\nabla f_v(x) = 3(x - \mu, v)^2 v$ and so $\mathbb{E}[\nabla f_v(x)] = 3M_2(v)v \neq 0$. We instead should take

$$f_v(x) = \langle x - \mu, v \rangle^3 - 3M_2(v)(x - \mu, v),$$

which yields

$$\mathbb{E}[\nabla f_v(x)] = \mathbb{E}[3(x - \mu, v)^2 v - 3M_2(v)v] = 0,$$

$$\mathbb{E}[\nabla^2 f_v(x)] = \mathbb{E}[6(x - \mu, v)(v \otimes v)] = 0,$$

$$\nabla^3 f_v(x) = 6(v \otimes v \otimes v).$$
Applying Theorem 2.47 to \( f_v(x) \) yields
\[
\text{Var}[f_v(x)] \leq C_3 \sigma^6 \|6(v \otimes v \otimes v)\|_F^2 = 36C_3 \sigma^6 \|v\|_2^6.
\] (156)

We can additionally compute
\[
\text{Var}[f_v(x)] = E[(x - \mu, v)^3 - 3M_2(v)(x - \mu, v)^2] - E[(x - \mu, v)^3 - 3M_2(v)(x - \mu, v)^2] = M_6(v) - 6M_2(v)M_4(v) + 9M_2(v)^3 - M_3(v)^2.
\] (157)

Since our goal is to bound \( M_6(v) \), we re-arrange to obtain
\[
M_6(v) = \text{Var}[f_v(x)] + 6M_2(v)M_4(v) + M_3(v)^2 - 9M_2(v)^2 \leq 36C_3 \sigma^6 \|v\|_2^6 + 6(\sigma^2 \|v\|_2^2)(C_d' \sigma^4 \|v\|_2^4) + M_3(v)^2 + 0
\] (159)

We can also use Hölder’s inequality to obtain \( M_3(v)^2 \leq M_2(v)M_4(v) \), which yields an overall bound of \( M_6(v) \leq (36C_3 + 12C_d') \sigma^6 \|v\|_2^6 \).

We now turn this into a sum-of-squares proof. We need to show the following four relations:
\[
(i) \ Var[f_v(x)] \lesssim_{\text{sos}} 36C_3 \sigma^6 \|v\|_2^6, \quad (ii) \ M_2(v)M_4(v) \lesssim_{\text{sos}} (\sigma^2 \|v\|_2^2)(C_d' \sigma^4 \|v\|_2^4), \quad (161)
\]
\[
(iii) \ M_3(v) \lesssim_{\text{sos}} M_2(v)M_4(v), \quad (iv) \ -9M_2(v)^2 \lesssim_{\text{sos}} 0. \quad (162)
\]

The relation (ii) again follows by the product property of \( \lesssim_{\text{sos}} \), while \( -9M_2(v)^2 \lesssim_{\text{sos}} 0 \) is direct because \( M_2(v)^2 \) is already a square. We will show in an exercise that the Hölder inequality in (iii) has a sum-of-squares proof, and focus on (i).

The relation (i) holds for reasons analogous to the \( k = 2 \) case. For a symmetric tensor \( A \in \mathbb{R}^{d^3} \), let \( f_A(x) = (A, (x - \mu)^{\otimes 3} - 3M_2 \otimes (x - \mu)) \). Then just as before we have \( E[\nabla f_A(x)] = 0, E[\nabla^2 f_A(x)] = 0 \), and so \( \text{Var}[f_A(x)] \leq 36C_3 \sigma^6 \|A\|_F^2 \), which implies that
\[
M_6 - 6M_2 \otimes M_4 + 9M_2 \otimes M_2 - M_3 \otimes M_3 \leq 36C_3 \sigma^6 I,
\] (163)

and hence \( \text{Var}[f_v(x)] \lesssim_{\text{sos}} 36C_3 \sigma^6 \|v\|_2^6 \) (again because semidefinite relations have sum-of-squares proofs).

In summary, we have \( M_6(v) \lesssim_{\text{sos}} (36C_3 + 12C_d') \sigma^6 \|v\|_2^6 \), as desired.

Generalizing to higher \( k \). For higher \( k \) the proof is essentially the same. What is needed is a function \( f_v(x) \) whose first \( k - 1 \) derivatives all have zero mean. This always exists and is unique up to scaling by constants. For instance, when \( k = 4 \) we can take \( f_v(x) = (x - \mu, v)^4 - 6M_2(v)(x - \mu, v)^2 - 4M_2(v)(x - \mu, v) - M_4(v) + 6M_2(v)^2 \). This appears somewhat chunky but is a special case of a combinatorial sum. For the general case, let \( T_k \) be the set of all integer tuples \( (i_0, i_1, \ldots) \) such that \( i_0 \geq 0, i_s \geq 2 \) for \( s > 0 \), and \( i_0 + i_1 + \cdots = k \). Then the general form is
\[
f_{v,k}(x) = \sum_{(i_0, \ldots, i_r) \in T_k} (-1)^r \binom{k}{i_0 \ldots i_r} (x - \mu, v)^{i_0} M_{i_1}(v) M_{i_2}(v) \cdots M_{i_r}(v).
\] (164)

The motivation for this formula is that it is the solution to \( \nabla f_{v,k}(x) = kf_{v,k-1}(x)v \). Using \( f_{v,k} \), one can construct sum-of-squares proofs by applying Theorem 2.47 to the analogous \( f_{A,k} \) function as before, and then use induction, the product rule, and Hölder’s inequality as in the \( k = 3 \) case.

References


3Actually this is not quite true because we only bound \( \text{Var}[f_A(x)] \) for symmetric tensors \( A \). What is true is that this holds if we symmetrize the left-hand-side of (163), which involves averaging over all ways of splitting \( M_2 \) and \( M_4 \) over the 3 copies of \( \mathbb{R}^d \) in \( \mathbb{R}^{d \times d \times d} \).
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

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

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 (167) 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

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

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 \((1 - \epsilon, \frac{1 - \epsilon}{\epsilon} \rho)\)-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 \(\langle \mathbb{E}[X - \mu \mid E], v \rangle = \|\mathbb{E}[X - \mu \mid E]\|\). Then we have

\[
\|\mathbb{E}[X - \mu \mid E]\| = \langle \mathbb{E}[X - \mu \mid E], v \rangle = \langle \mathbb{E}[X - \mu, v] \mid E \rangle \tag{170}
\]

\[
\leq \mathbb{E}[\langle X - \mu, v \rangle \mid \langle X - \mu, v \rangle \geq \tau_\epsilon(v)] \tag{171}
\]

\[
\leq \frac{1 - \epsilon}{\epsilon} \rho. \tag{172}
\]

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

\[
\mathbb{E}[\langle X - \mu, v \rangle \mid \langle X - \mu, v \rangle \geq \tau_\epsilon(v)] = \mathbb{E}[\langle X - \mu, v \rangle \mid E_v] \tag{174}
\]

\[
= \mathbb{E}[X - \mu \mid E_v, v] \tag{175}
\]

\[
\leq \|\mathbb{E}[X - \mu \mid E_v]\| \tag{176}
\]

\[
\leq \frac{1 - \epsilon}{\epsilon} \rho, \tag{177}
\]

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 \(\langle x_i - \mu, v \rangle\) 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

\[
\mathbb{E}_{x \sim p}[\|\langle x - \mu, v \rangle\|] = 2\mathbb{E}_{x \sim p}[\max(\langle x - \mu, v \rangle, 0)] = 2\mathbb{P}[E_+]\mathbb{E}_{x \sim p}[\langle x - \mu, v \rangle \mid E_+] \tag{178}
\]

\[
\leq 2\mathbb{P}[E_+]\|\mathbb{E}_{x \sim p}[x - \mu \mid E_+]\| \leq 2\rho, \tag{179}
\]

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

\[
\mathbb{E}[\langle X - \mu, v \rangle \mid \langle X - \mu, v \rangle \geq \tau_{1/2}(v)] \leq \mathbb{E}[\|\langle X - \mu, v \rangle\|/\mathbb{P}[\langle X - \mu, v \rangle \geq \tau_{1/2}(v)] = 2\mathbb{E}[\|\langle X - \mu, v \rangle\|] \leq 2\rho, \tag{181}
\]

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