# Lecture Notes for STAT240 (Robust and Nonparametric Statistics)

Jacob Steinhardt

Last updated: April 27, 2021

[Lecture 1]

## 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:

1. 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.

2. 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). There might also be noise e.g. due to a sensor failing.

*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.

Complementary to robustness is the idea of *model mis-specification.* When the true distribution $p^*$ lies within our model family, many robustness issues are less severe: we can rely on the model to extrapolate to new settings, and we can often get well-calibrated uncertainty estimates. This motivates the second focus of the course, *nonparametric modeling*, where we consider broad function classes (e.g. all smooth functions) that are more likely to be correctly specified. Another connection between nonparametrics and robustness is that we often want robust methods to work for any distribution within some large, infinite-dimensional class.

**Overarching framework.** Most robustness questions can be cast in the following way: 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 both of the examples discused at the beginning. However, it will be profitable to think about each case separately due to different emphases:

$$\begin{tikzcd}
\text{test distribution} & \text{train distribution} & \\
p^* \arrow[r, "D(p^*, \tilde{p}) \leq \epsilon"] & \tilde{p} \arrow[dddd] & \\
\rotatebox{90}{$\in$} & & \\
\mathcal{G} & & \\
\text{distributional assumptions} & & \\
 & \text{samples} & \\
 & X_1, \ldots, X_n \arrow[d] & \\
 & \text{estimated parameters} & \text{cost} \\
 & \hat{\theta}(X_1, \ldots, X_n) \arrow[r] & L(p^*, \hat{\theta})
\end{tikzcd}$$

Figure 1: Framework for studying robust learning.

1. For corrupted training data, we think of $\tilde{p}$ as being corrupted and $p^*$ as being nice.
2. 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.

**Distributional shift.** For distributional shift, we often seek invariant features or structure that can transfer information from the train to test distributions. We can also counteract distribution shift by training on more diverse data. Finally, we can use model uncertainty to infer out-of-distribution error, but these inferences can be fraught if the model is mis-specified.

Table 1: Comparison of different robust settings.

| Train robustness | Distributional shift |
|---|---|
| $p^*$ nice | $p^*$ and $\tilde{p}$ both nice |
| undo corruptions | invariant features |
| | diverse training data |

**Nonparametric modeling.** This brings us to nonparametric methods. The simplest way to be nonparametric is through the model–using a rich class such as smooth functions, or neural networks. This mitigates model mis-specification but raises new statistical challenges. Familiar phenomena such as the central limit theorem cease to hold, and error rates are instead governed by the eigenvalue spectrum of an infinite-dimensional kernel. Aside from the model, we can also be nonparametric at inference time; an example is the bootstrap, which constructs confidence intervals that are more robust to model mis-specification than classical parametric tests.

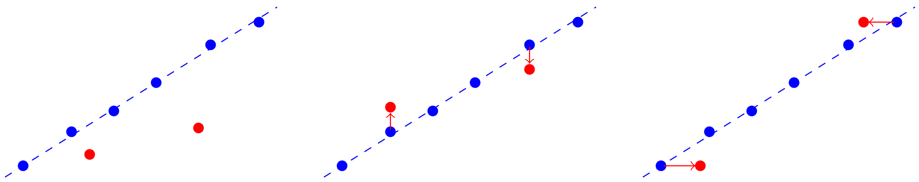


Figure 2: Possible corruptions to be robust to. Left: data contains outliers. Middle: outputs are perturbed (process noise); Right: inputs are perturbed (measurement error).

**Summary.** We will cover each of training time robustness, distribution shift, and nonparametric modeling, drawing connections as we go. The first third will focus on training time, also building up the statistical machinery needed to prove generalization bounds. Then, we will shift focus to model mis-specification, and to nonparametric methods as a remedy. These including kernel regression, the bootstrap, and neural networks, and we will both see how they mitigate model mis-specification and how to analyze their generalization performance. Finally, we will turn to distribution shift, and see that nonparametric models are often robust to distribution shift when trained on the right data. Training robustness will receive a largely theoretical treatment, while for nonparametrics and distribution shift we will see a mix of theoretical and empirical results.

# 2 Training Time Robustness

We will start our investigation with training time robustness. As in Figure 1, we observe samples $X_1, \dots, X_n$ 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.

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 $X_1, \dots, X_n$. 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 $D(p, q) = \mathsf{TV}(p, q)$, where $\mathsf{TV}$ is the *total variation distance*:

$$\mathsf{TV}(p, q) \overset{\text{def}}{=} \sup\{|p(E) - q(E)| \mid E \text{ is a measurable event}\}. \tag{1}$$

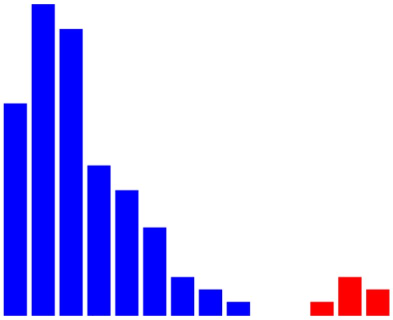
If $p$ and $q$ both have densities then an equivalent characterization is $\mathsf{TV}(p, q) = \frac{1}{2} \int |p(x) - q(x)| dx$.

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 $D = \mathsf{TV}$.

## 2.1 Robustness to Outliers in 1 Dimension

First consider mean estimation in one dimension: we observe $n$ data points $x_1, \dots, x_n \in \mathbb{R}$ 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 $\mathcal{G} = \mathcal{G}_{\mathsf{cov}}(\sigma)$ 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 $\mathsf{TV}$-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, $\mathbb{P}_{X \sim p}[X \geq \mu + \sigma/\sqrt{\delta}] \leq \delta$. Moreover, if $E$ is any event with probability at least $\delta$, then $|\mathbb{E}_{X \sim p}[X \mid E] - \mu| \leq \sigma\sqrt{\frac{2(1-\delta)}{\delta}}$.*

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

*Proof.* If $\mathsf{TV}(\tilde{p}, p^*) \leq \epsilon$, 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 $\frac{\sigma}{\sqrt{\epsilon}}$ of the mean $\mu$ by Chebyshev's inequality.

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

Now let $\epsilon'$ be the fraction of remaining points which are bad, and note that $\epsilon' \leq \frac{\epsilon}{1-4\epsilon}$. The mean of all the remaining points differs from $\mu$ by at most $\epsilon' \cdot \sigma\sqrt{\frac{1}{\epsilon}} + (1 - \epsilon') \cdot \sigma\sqrt{\frac{10\epsilon}{1-5\epsilon}}$, which is at most $(1 + \sqrt{10})\frac{\sqrt{\epsilon}}{1-4\epsilon}\sigma$. This is in turn at most $9\sigma\sqrt{\epsilon}$ assuming that $\epsilon \leq \frac{1}{8}$. □

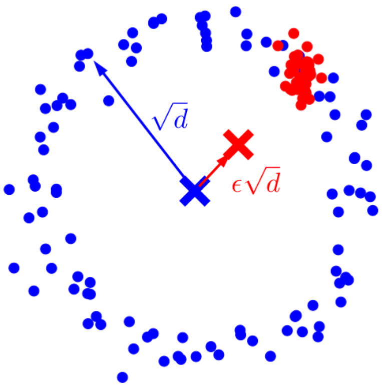


Figure 3: The outliers can lie at distance $\sqrt{d}$ without being detected, skewing the mean by $\epsilon\sqrt{d}$.

**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 $(\sigma_k)^k$ 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.

[Lecture 2]

### 2.3 Minimum Distance Functionals

In the previous section we saw that simple approaches to handling outliers in high-dimensional data, such as the trimmed mean, incur a $\sqrt{d}$ error. We will avoid this error using *minimum distance functionals*, an idea which seems to have first appeared in Donoho and Liu (1988).

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

$$\hat{\theta}(\tilde{p}) = \theta^*(q) = \arg\min_{\theta} L(q, \theta), \text{ where } q = \arg\min_{q \in \mathcal{G}} D(q, \tilde{p}). \tag{2}$$

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 $L(p^*, \hat{\theta}(\tilde{p}))$ of the minimum distance functional is at most the maximum loss between any pair of distributions in $\mathcal{G}$ of distance at most $2\epsilon$:*

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

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

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

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

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

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

$$\mathsf{TV}(\mathcal{N}(-u/2, 1), \mathcal{N}(u/2, 1)) = \frac{1}{2\sqrt{2\pi}} \int_{-\infty}^{\infty} |e^{-(t+u/2)^2/2} - e^{-(t-u/2)^2/2}| dt \tag{5}$$

$$\overset{(i)}{=} \frac{1}{\sqrt{2\pi}} \int_{-u/2}^{u/2} e^{-t^2/2} dt. \tag{6}$$

(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 $\mathcal{G}_{\mathsf{gauss}}$ be the family of isotropic Gaussians, $D = \mathsf{TV}$, and $L$ the difference in means as above. Then $\mathfrak{m}(\mathcal{G}_{\mathsf{gauss}}, \epsilon, D, L) \leq 2\sqrt{2\pi}\epsilon$ 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 $\mathfrak{m}$ 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 $\mathcal{G}_{\mathsf{cov}}(\sigma)$ be the family of distributions whose covariance matrix $\Sigma$ satisfies $\Sigma \preceq \sigma^2 I$. Then $\mathfrak{m}(\mathcal{G}_{\mathsf{cov}}(\sigma), \epsilon) = \mathcal{O}(\sigma\sqrt{\epsilon})$.*

*Proof.* Let $p, q \in \mathcal{G}_{\mathsf{cov}}(\sigma)$ such that $\mathsf{TV}(p, q) \leq \epsilon$. This means that we can get from $p$ to $q$ by first deleting $\epsilon$ 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

$$r = \frac{\min(p, q)}{1 - \mathsf{TV}(p, q)}. \tag{7}$$

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 $\|\mathbb{E}_{X\sim p}[X \mid E] - \mu\|_2 \leq \sigma\sqrt{\frac{2(1-\delta)}{\delta}}$.*

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

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

In Lemma 2.7, the two key properties we needed were:

- The *midpoint property* of $\mathsf{TV}$ 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 $\mathfrak{m}$ is bounded for a general family of distributions containins Gaussians, sub-Gaussians, bounded covariance distributions, and many others. The main observation is that in the proof of Lemma 2.7, all we needed was that the tails of distributions in $\mathcal{G}$ were bounded, in the sense that deleting an $\epsilon$-fraction of the points could not substantially change the mean. This motivates the following definition:

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

$$\|\mathbb{E}_{X\sim p}[X \mid E] - \mathbb{E}_{X\sim p}[X]\| \leq \rho \text{ for all events } E \text{ with } p(E) \geq 1 - \epsilon. \tag{8}$$

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

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

Resilient distributions always have bounded modulus:

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

$$\mathfrak{m}(\mathcal{G}_{\mathsf{TV}}(\rho, \epsilon), 2\epsilon) \leq 2\rho \tag{9}$$

*whenever $\epsilon < 1/2$.*

*Proof.* As in Lemma 2.7, the key idea is that any two distributions $p, q$ that are close in $\mathsf{TV}$ have a *midpoint* distribution $r = \frac{\min(p,q)}{1-\mathsf{TV}(p,q)}$ that is a deletion of both distributions). This midpoint distribution connects the two distributions, and it follows from the triangle inequality that the modulus of $\mathcal{G}_{\mathsf{TV}}$. is bounded. We illustrate this idea in Figure 4 and make it precise below.

$$\begin{tikzcd}
p \arrow[rr, "\mathsf{TV}(p,q) \le \epsilon"] \arrow[dr, "r \le \frac{p}{1-\epsilon}"'] & & q \arrow[dl, "r \le \frac{q}{1-\epsilon}"] \\
& r = \frac{\min(p,q)}{1-\mathsf{TV}(p,q)} &
\end{tikzcd}$$

$$\begin{tikzcd}
p \in \mathcal{G}_{\mathsf{TV}} \arrow[r, Rightarrow] & \|\mu(p) - \mu(r)\| \le \rho \arrow[dr, Rightarrow] & \\
& & \|\mu(p) - \mu(q)\| \le 2\rho \\
q \in \mathcal{G}_{\mathsf{TV}} \arrow[r, Rightarrow] & \|\mu(q) - \mu(r)\| \le \rho \arrow[ur, Rightarrow] &
\end{tikzcd}$$

Figure 4: Midpoint distribution $r$ helps bound the modulus for $\mathcal{G}_{\mathsf{TV}}$.

Recall that

$$\mathfrak{m}(\mathcal{G}_{\mathsf{TV}}(\rho, \epsilon), 2\epsilon) = \sup_{p,q \in \mathcal{G}_{\mathsf{TV}}(\rho,\epsilon) : \mathsf{TV}(p,q) \le 2\epsilon} \|\mu(p) - \mu(q)\|. \tag{10}$$

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

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

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

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

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

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

Definition 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

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

We let $\mathcal{G}_\psi(\sigma)$ 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, $\mathcal{G}_{\mathsf{cov}}(\sigma) = \mathcal{G}_\psi(\sigma)$ for $\psi(x) = x^2$, so Orlicz norms generalize bounded covariance. It is also possible to generalize Definition 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 $\mathcal{G}_\psi(\sigma)$ is contained in $\mathcal{G}_{\mathsf{TV}}(2\sigma\epsilon\psi^{-1}(1/\epsilon), \epsilon)$ for all $0 < \epsilon < 1/2$.*

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

$$\|\mathbb{E}_{X\sim p}[X \mid E]\|_2 \overset{(i)}{=} \frac{\epsilon'}{1-\epsilon'}\|\mathbb{E}_{X\sim p}[X \mid E^c]\|_2 \tag{13}$$

$$= \frac{\epsilon'}{1-\epsilon'} \sup_{\|v\|_2\leq 1} \mathbb{E}_{X\sim p}[\langle X, v\rangle \mid E^c] \tag{14}$$

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

$$\overset{(iii)}{\leq} \frac{\epsilon'}{1-\epsilon'} \sup_{\|v\|_2\leq 1} \sigma\psi^{-1}(\mathbb{E}_{X\sim p}[\psi(|\langle X, v\rangle|/\sigma)]/\epsilon') \tag{16}$$

$$\overset{(iv)}{\leq} \frac{\epsilon'}{1-\epsilon'}\sigma\psi^{-1}(1/\epsilon') \leq 2\epsilon\sigma\psi^{-1}(1/\epsilon), \tag{17}$$

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

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

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

**Lemma 2.14.** *For a fixed vector $v$, let $\tau_\epsilon(v)$ denote the $\epsilon$-quantile of $\langle x - \mu, v\rangle$: $\mathbb{P}_{x\sim p}[\langle x - \mu, v\rangle \geq \tau_\epsilon(v)] = \epsilon$. Then, $p$ is $(\rho, \epsilon)$-resilient in a norm $\|\cdot\|$ if and only if the $\epsilon$-tail of $p$ has bounded mean when projected onto any dual unit vector $v$:*

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

*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:

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

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

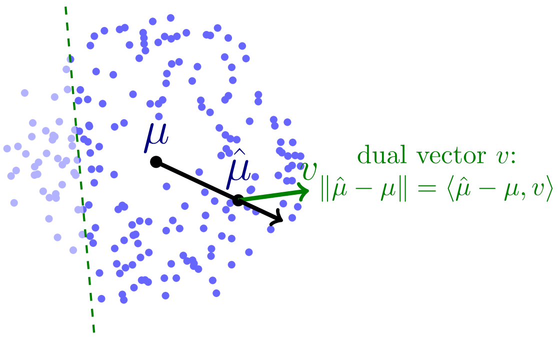


Figure 5: The optimal set $T$ discards the smallest $\epsilon|S|$ elements projected onto a dual unit vector $v$.

**Recap.** We saw that the error of the trimmed mean grew as $\sqrt{d}$ in $d$ dimensions, and introduced an alternative estimator–the minimum distance functional–that achieves better error. Specifically, it achieves error $2\rho$ 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 $\mathcal{G}_{\mathsf{TV}}(\rho, \epsilon)$ and be confident that this will typically yield an optimal error bound. The only complication is that projection still depends on the parameters $\rho$ and $\epsilon$; however, we can do without knowledge of either one of the parameters as long as we know the other.

[Lecture 3]

## 2.5 Concentration Inequalities

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

Concentration inequalities usually involve two steps:

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

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

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

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

> A slot machine has an expected pay-out of \$5 (and its payout is always non-negative). What can we say about the probability that it pays out at least \$100?

We observe that the probability must be at most 0.05, since a 0.05 chance of a \$100 payout would by itself already contribute \$5 to the expected value. Moreover, this bound is achievable by taking a slot machine that pays \$0 with probability 0.95 and \$100 with probability 0.05. Markov's inequality is the generalization of this observation:

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

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, $\mathbb{P}[|X - \mu| \geq t \cdot \sigma] \leq \frac{1}{t^2}$.*

*Proof.* Since $Z = (X - \mu)^2$ is non-negative, we have that $\mathbb{P}[Z \geq t^2 \cdot \sigma^2] \leq \frac{1}{t^2}$ by Markov's inequality. Taking the square-root gives $\mathbb{P}[|X - \mu| \geq t \cdot \sigma] \leq \frac{1}{t^2}$, 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 $\mathsf{Var}[X]$ denote the variance of $X$. Then,*

$$\mathsf{Var}[X_1 + \cdots + X_n] = \mathsf{Var}[X_1] + \cdots + \mathsf{Var}[X_n]. \tag{19}$$

*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 $\mathsf{Var}[X + Y] = \mathbb{E}[(X + Y)^2] = \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\mathbb{E}[XY] = \mathsf{Var}[X] + \mathsf{Var}[Y] + 2\mathbb{E}[X]\mathbb{E}[Y] = \mathsf{Var}[X] + \mathsf{Var}[Y]$, 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, $\mathbb{P}[|S - \mu|/\sigma \geq t/\sqrt{n}] \leq 1/t^2$.*

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

**Higher moments.** Chebyshev's inequality gives bounds in terms of the second moment of $X - \mu$. Can we do better by considering higher moments such as the 4th moment? Supposing that $\mathbb{E}[(X - \mu)^4] \leq \tau^4$, we do get the analogous bound $\mathbb{P}[|X - \mu| \geq t \cdot \tau] \leq 1/t^4$. 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

$$\mathbb{E}[(X + Y)^4] = \mathbb{E}[X^4] + \mathbb{E}[Y^4] + 6\mathbb{E}[X^2]\mathbb{E}[Y^2], \tag{20}$$

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

$$\mathbb{E}[|\sum_i X_i|^p] \leq \mathcal{O}(p)^p \sum_i \mathbb{E}[|X_i|^p] + \mathcal{O}(\sqrt{p})^p (\sum_i \mathbb{E}[X_i^2])^{p/2} \tag{21}$$

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

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

$$m_X(\lambda) \overset{\text{def}}{=} \mathbb{E}[\exp(\lambda(X - \mu))]. \tag{22}$$

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*

$$\mathbb{P}[X - \mu \geq t] \leq \inf_{\lambda \geq 0} m_X(\lambda) e^{-\lambda t}. \tag{23}$$

*Proof.* By Markov's inequality, $\mathbb{P}[X - \mu \geq t] = \mathbb{P}[\exp(\lambda(X - \mu)) \geq \exp(\lambda t)] \leq \mathbb{E}[\exp(\lambda(X - \mu))]/\exp(\lambda t)$, which is equal to $m_X(\lambda)e^{-\lambda t}$ by the definition of $m_X$. Taking inf over $\lambda$ yields the claimed bound. □

**Sub-exponential and sub-Gaussian distributions.** An important special case is sub-exponential random variables; recall these are random variables satisfying $\mathbb{E}[\exp(|X - \mu|/\sigma)] \leq 2$. For these, applying the Chernoff bound with $\lambda = 1/\sigma$ yields $\mathbb{P}[X - \mu \geq t] \leq 2e^{-t/\sigma}$.

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

$$m_X(\lambda) = \mathbb{E}[\exp(\lambda(X - \mu))] \leq \mathbb{E}[\exp(\lambda^2\sigma^2/4 + (X - \mu)^2/\sigma^2)] \leq 2\exp(\lambda^2\sigma^2/4). \tag{24}$$

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

Sub-Gaussians are particularly convenient because the bound $m_X(\lambda) \leq \exp(3\lambda^2\sigma^2/2)$ composes well. Let $X_1, \ldots, X_n$ be independent sub-Gaussians with constants $\sigma_1, \ldots, \sigma_n$. Then we have $m_{X_1+\cdots+X_n}(\lambda) \leq \exp(3\lambda^2(\sigma_1^2 + \cdots + \sigma_n^2)/2)$. We will use this to bound the behavior of sums of bounded random variables using *Hoeffding's inequality*:[1]

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

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

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

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

The cumulant function is defined as

$$K_X(\lambda) \stackrel{\text{def}}{=} \log \mathbb{E}[\exp(\lambda X)]. \tag{25}$$

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:

$$K_X(\lambda) = 1 + \sum_{n=1}^{\infty} \frac{\kappa_n(X)}{n!}\lambda^n. \tag{26}$$

[1] Most of the constants presented here are suboptimal; we have focused on giving simpler proofs at the expense of sharp constants.

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

$$\kappa_1(X) = 0, \tag{27}$$

$$\kappa_2(X) = \mathbb{E}[X^2], \tag{28}$$

$$\kappa_3(X) = \mathbb{E}[X^3], \tag{29}$$

$$\kappa_4(X) = \mathbb{E}[X^4] - 3\mathbb{E}[X^2]^2, \tag{30}$$

$$\kappa_5(X) = \mathbb{E}[X^5] - 10\mathbb{E}[X^3]\mathbb{E}[X^2], \tag{31}$$

$$\kappa_6(X) = \mathbb{E}[X^6] - 16\mathbb{E}[X^4]\mathbb{E}[X^2] - 10\mathbb{E}[X^3]^2 + 30\mathbb{E}[X^2]^3. \tag{32}$$

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

[Lecture 4]

### 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 $\mathcal{O}(\sigma\sqrt{\log(n/\delta)})$ with probability $1-\delta$.*

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

Lemma 2.22 illustrates a typical proof strategy: We first decompose the event we care about as a union of simpler events, then show that each individual event holds with high probability by exploiting independence. As long as the "failure probability" of a single event is much small 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 $\mathcal{O}(\sigma^2 \cdot (1 + d/n + \log(1/\delta)/n))$ with probability $1-\delta$.*

*Proof.* The maximum eigenvalue can be expressed as

$$\|M\| = \sup_{\|v\|_2 \leq 1} v^\top M v = \sup_{\|v\|_2 \leq 1} \frac{1}{n}\sum_{i=1}^n |\langle X_i, v\rangle|^2. \tag{33}$$

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

$$\mathbb{E}[\exp(\frac{n}{\sigma^2}v^\top Mv)] = \mathbb{E}[\exp(\sum_{i=1}^n |\langle X_i, v\rangle|^2/\sigma^2)] \tag{34}$$

$$= \prod_{i=1}^n \mathbb{E}[\exp(|\langle X_i, v\rangle|^2/\sigma^2)] \leq 2^n, \tag{35}$$

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

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

$$\sup_{v \in \mathcal{N}_{1/4}} v^\top Mv \geq \frac{1}{2} \sup_{\|v\|_2 \leq 1} v^\top Mv. \tag{36}$$

Our construction follows Section 5.2.2 of Vershynin (2010). Let $\mathcal{N}_{1/4}$ be a maximal set of points in the unit ball such that $\|x - y\|_2 \geq 1/4$ for all distinct $x, y \in \mathcal{N}_{1/4}$. We observe that $|\mathcal{N}_{1/4}| \leq 9^d$; this is because the balls of radius 1/8 around each point in $\mathcal{N}_{1/4}$ are disjoint and contained in a ball of radius 9/8.

To establish (36), let $v$ maximize $v^\top Mv$ over $\|v\|_2 \leq 1$ and let $u$ maximize $v^\top Mv$ over $\mathcal{N}_{1/4}$. Then

$$|v^\top Mv - u^\top Mu| = |v^\top M(v - u) + u^\top M(v - u)| \tag{37}$$

$$\leq (\|v\|_2 + \|u\|_2)\|M\|\|v - u\|_2 \tag{38}$$

$$\leq 2 \cdot \|M\| \cdot (1/4) = \|M\|/2. \tag{39}$$

Since $v^\top Mv = \|M\|$, we obtain $|\|M\| - u^\top Mu| \leq \|M\|/2$, whence $u^\top Mu \geq \|M\|/2$, which establishes (36). We are now ready to apply the union bound: Recall that from the Chernoff bound on $v^\top Mv$, we had $\mathbb{P}[v^\top Mv \geq t] \leq 2^n \exp(-nt/\sigma^2)$, so

$$\mathbb{P}[\sup_{v \in \mathcal{N}_{1/4}} v^\top Mv \geq t] \leq 9^d 2^n \exp(-nt/\sigma^2). \tag{40}$$

Solving for this quantity to equal $\delta$, we obtain

$$t = \frac{\sigma^2}{n} \cdot (n \log(2) + d \log(9) + \log(1/\delta)) = \mathcal{O}(\sigma^2 \cdot (1 + d/n + \log(1/\delta)/n)), \tag{41}$$

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

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 $\mathsf{vc}(H) = d$, 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 $\nu_n(f) = \frac{1}{n}|\{i \mid f(X_i) = 1\}|$ and let $\nu(f) = p(f(X) = 1)$. Then*

$$\sup_{f \in \mathcal{H}} |\nu_n(f) - \nu(f)| \leq \mathcal{O}\Big(\sqrt{\frac{d + \log(1/\delta)}{n}}\Big) \tag{42}$$

*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 $d = \mathsf{vc}(\mathcal{H})$. 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 $(f(X_1), \dots, f(X_n))$; however, this doesn't work because revealing $X_1, \dots, X_n$ 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 $X_1', \dots, X_n'$ be independent copies of $X_1, \dots, X_n$ and let $\nu_n'(f)$ denote the version of $\nu_n(f)$ computed with the $X_i'$. Then we have

$$\mathbb{E}_X[\sup_{f \in \mathcal{H}} |\nu_n(f) - \nu(f)|] \leq \mathbb{E}_{X,X'}[\sup_{f \in \mathcal{H}} |\nu_n(f) - \nu_n'(f)|] \tag{43}$$

$$= \frac{1}{n}\mathbb{E}_{X,X'}[\sup_{f \in \mathcal{H}} |\sum_{i=1}^{n} f(X_i) - f(X_i')|. \tag{44}$$

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 ±1 with equal probability. Introducing these variables and continuing the inequality, we thus have

$$\frac{1}{n}\mathbb{E}_{X,X'}[\sup_{f \in \mathcal{H}} |\sum_{i=1}^{n} f(X_i) - f(X_i')|] = \frac{1}{n}\mathbb{E}_{X,X',s}[\sup_{f \in \mathcal{H}} |\sum_{i=1}^{n} s_i(f(X_i) - f(X_i'))|]. \tag{45}$$

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 $V_{\mathcal{H}}(n)^2 = \mathcal{O}(n^{2d})$ values. But for fixed $X, X'$, the quantities $s_i(f(X_i) - f(X_i'))$ are independent, zero-mean, bounded random variables and hence for fixed $f$ we have $\mathbb{P}[\sum_i s_i(f(X_i) - f(X_i')) \geq t] \leq \exp(-t^2/9n)$ by Hoeffding's inequality. Union bounding over the $\mathcal{O}(n^{2d})$ effectively distinct $f$, we obtain

$$\mathbb{P}_s[\sup_{f \in \mathcal{H}} |\sum_i s_i(f(X_i) - f(X_i'))| \geq t \mid X, X'] \leq \mathcal{O}(n^{2d}) \exp(-t^2/9n). \tag{46}$$

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

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

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

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

[Lecture 5]

## 2.6 Finite-Sample Analysis

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

$$\hat{\theta}(\tilde{p}) = \theta^*(q), \text{ where } q = \operatorname*{arg\,min}_{q \in \mathcal{G}} \mathsf{TV}(q, \tilde{p}). \tag{47}$$

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

The problem with the minimum distance functional defined above is that projection under $\mathsf{TV}$ usually doesn't make sense for finite samples! For instance, suppose that $p$ is a Gaussian distribution and let $p_n$ and $p_n'$ be the empirical distributions of two different sets of $n$ samples. Then $\mathsf{TV}(p_n, p) = \mathsf{TV}(p_n, p_n') = 1$ almost surely. This is because samples from a continuous probability distribution will almost surely be distinct, and $\mathsf{TV}$ distance doesn't give credit for being "close"—the $\mathsf{TV}$ 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 $\mathsf{TV}$ 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 $\mathsf{TV}$ with a more forgiving distance $\widetilde{\mathsf{TV}}$ and use the minimum distance functional corresponding to this relaxed distance. To show that projection under $\widetilde{\mathsf{TV}}$ still works, we will need to check that the modulus $\mathfrak{m}(\mathcal{G}, \epsilon)$ is still small after we replace $\mathsf{TV}$ with $\widetilde{\mathsf{TV}}$, and we will also need to check that the distance $\widetilde{\mathsf{TV}}(p, p_n)$ between $p$ and its empirical distribution is small with high probability. We do this below in Section 2.6.1.

An alternative to relaxing the distance from $\mathsf{TV}$ to $\widetilde{\mathsf{TV}}$ is to expand the destination set from $\mathcal{G}$ to some $\mathcal{M} \supset \mathcal{G}$, such that even though $p$ is not close to the empirical distribution $p_n$, *some* element of $\mathcal{M}$ is close to $p_n$. Another advantage to expanding the destination set is that projecting onto $\mathcal{G}$ may not be computationally tractable, while projecting onto some larger set $\mathcal{M}$ can sometimes be done efficiently. We will see how to statistically analyze this modified projection algorithm in Section 2.6.2, and study the computational feasibility of projecting onto a set $\mathcal{M}$ starting in Section 2.7.

### 2.6.1 Relaxing the Distance

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

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

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

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

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

$$\widetilde{\mathsf{TV}}_{\mathcal{H}}(p, q) \stackrel{\text{def}}{=} \sup_{f \in \mathcal{H}, \tau \in \mathbb{R}} |\mathbb{P}_{X \sim p}[f(X) \geq \tau] - \mathbb{P}_{X \sim q}[f(X) \geq \tau]|. \tag{48}$$

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

First note that $\widetilde{\mathsf{TV}}_{\mathcal{H}}$ is indeed upper-bounded by $\mathsf{TV}$, since $\mathsf{TV}(p, q) = \sup_E |p(E) - q(E)|$ 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*:

**Lemma 2.28.** *Suppose that $p$ and $q$ are two distributions such that $\widetilde{\mathsf{TV}}_{\mathcal{H}}(p, q) \leq \epsilon$. 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 $\mathbb{E}_{X \sim r_p}[f(X)] \geq \mathbb{E}_{Y \sim r_q}[f(Y)]$.*

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

*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 smallest points of $f(X)$ in $p$ to get $r_p$, and delete $\epsilon$ probability mass corresponding to the largest points $f(Y)$ in $q$ to get $r_q$. Since $\widetilde{\mathsf{TV}}_{\mathcal{H}}(p, q) \le \epsilon$ we have

$$\sup_{\tau \in \mathbb{R}} |\mathbb{P}_{X\sim p}(f(X) \ge \tau) - \mathbb{P}_{Y \sim q}(f(Y) \ge \tau)| \le \epsilon, \tag{49}$$

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

Mean crossing lemmas such as Lemma 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 $\mathcal{G}_{\mathsf{TV}}(\rho, \epsilon)$ of $(\rho, \epsilon)$-resilient distributions and $L(p, \theta) = \|\theta - \mu(p)\|_2$, we have*

$$\mathfrak{m}(\mathcal{G}_{\mathsf{TV}}(\rho, \epsilon), \epsilon, \widetilde{\mathsf{TV}}_{\mathcal{H}_{\mathsf{lin}}}) \le 2\rho. \tag{50}$$

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

*Proof of Corollary 2.29.* Let $p, q \in \mathcal{G}_{\mathsf{TV}}$ such that $\widetilde{\mathsf{TV}}(p, q) \le \epsilon$. Take $v = \arg\max_{\|v\|_2=1} v^\top(\mathbb{E}_p[X] - \mathbb{E}_q[X])$, hence $\mathbb{E}_p[v^\top X] - \mathbb{E}_q[v^\top X] = \|\mathbb{E}_p[X] - \mathbb{E}_q[X]\|_2$. It follows from Lemma 2.28 that there exist $r_p \le \frac{p}{1-\epsilon}, r_q \le \frac{q}{1-\epsilon}$ such that

$$\mathbb{E}_{r_p}[v^\top X] \le \mathbb{E}_{r_q}[v^\top X]. \tag{51}$$

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

$$\mathbb{E}_p[v^\top X] - \mathbb{E}_{r_p}[v^\top X] \le \rho, \tag{52}$$

$$\mathbb{E}_{r_q}[v^\top X] - \mathbb{E}_q[v^\top X] \le \rho. \tag{53}$$

Then,

$$\begin{aligned}
\|\mathbb{E}_p[X] - \mathbb{E}_q[X]\|_2 &= \mathbb{E}_p[v^\top X] - \mathbb{E}_q[v^\top X] && (54)\\
&= \underbrace{\mathbb{E}_p[v^\top X] - \mathbb{E}_{r_p}[v^\top X]}_{\le \rho} + \underbrace{\mathbb{E}_{r_p}[v^\top X] - \mathbb{E}_{r_q}[v^\top X]}_{\le 0} + \underbrace{\mathbb{E}_{r_q}[v^\top X] - \mathbb{E}_q[v^\top X]}_{\le \rho} && (55)\\
&\le 2\rho, && (56)
\end{aligned}$$

which shows the modulus is small as claimed. □

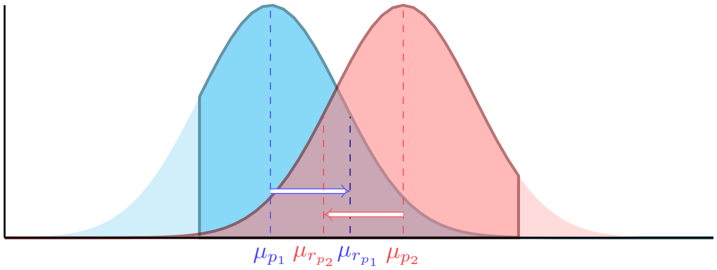


Figure 6: Illustration of mean cross lemma. For any distributions $p_1, p_2$ that are close under $\widetilde{\mathsf{TV}}$, we can truncate the $\epsilon$-tails of each distribution to make their means cross.

**Bounding the distance to the empirical distribution.** Now that we have bounded the modulus, it remains to bound the distance $\widetilde{\mathsf{TV}}(\tilde{p}, \tilde{p}_n)$. Note that $\widetilde{\mathsf{TV}}(\tilde{p}, \tilde{p}_n)$ is exactly the quantity bounded in equation (42) of Proposition 2.24; we thus have that $\widetilde{\mathsf{TV}}_{\mathcal{H}}(\tilde{p}, \tilde{p}_n) \leq \mathcal{O}\big(\sqrt{\frac{\mathsf{vc}(\mathcal{H})+\log(1/\delta)}{n}}\big)$ with probability $1-\delta$. Here $\mathsf{vc}(\mathcal{H})$ is the VC dimension of the family of threshold functions $\{x \mapsto \mathbb{I}[f(x) \geq \tau] \mid f \in \mathcal{H}, \tau \in \mathbb{R}\}$. So, for $\mathcal{H} = \mathcal{H}_{\mathsf{lin}}$ all we need to do is bound the VC dimension of the family of halfspace functions on $\mathbb{R}^d$.

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

$$\sum_{i=1}^{d+2} a_i x_i = 0, \sum_{i=1}^{d+2} a_i = 0. \tag{57}$$

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 $\langle x_i, v\rangle \geq \tau$ iff $i \in S_+$. (This is because both a halfspace and its complement are convex, so if we let $H_{v,\tau}$ denote the half-space, it is impossible to have $S_+ \subset H_{v,\tau}$, $S_- \subset H_{v,\tau}^c$, and $\mathrm{conv}(S_+) \cap \mathrm{conv}(S_-) \neq \emptyset$.)

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

$$\frac{1}{A} \sum_{i \in S_+} a_i x_i = \frac{1}{A} \sum_{i \in S_-} (-a_i) x_i, \tag{58}$$

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

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

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

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

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

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

$$\min_{\mu} \max_{v \in \mathbb{R}^d} |\mathbb{P}_{X \sim \tilde{p}_n}[\langle X, v\rangle \geq \langle \mu, v\rangle] - \tfrac{1}{2}| \tag{59}$$

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

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

$$\min_{q \in \mathcal{G}} \max_{v \in \mathbb{R}^d, \tau \in \mathbb{R}} |\mathbb{P}_{X \sim \tilde{p}_n}[\langle X, v\rangle \geq \tau] - \mathbb{P}_{X \sim q}[\langle X, v\rangle \geq \tau]|. \tag{60}$$

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 $\mathsf{TV}$ projection by relaxing to a weaker distance $\widetilde{\mathsf{TV}}$. We will now study an alternate approach, based on expanding the destination set $\mathcal{G}$ to a larger set $\mathcal{M}$. For this approach we will need to reference the "true empirical distribution" $p_n^*$. What we mean by this is the following: Whenever $\mathsf{TV}(p^*, \tilde{p}) \leq \epsilon$, 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 $\tilde{p}_{|E}$.
3. Otherwise leave $X$ as-is.

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

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

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

In fact, it suffices for $\mathcal{M}$ to satisfy a weaker property; we only need the "generalized modulus" to be small relative to some $\mathcal{G}' \subset \mathcal{M}$:

**Proposition 2.31.** *For a set $\mathcal{G}' \subset \mathcal{M}$, define the generalized modulus of continuity as*

$$\mathfrak{m}(\mathcal{G}', \mathcal{M}, 2\epsilon) \stackrel{\mathrm{def}}{=} \min_{p \in \mathcal{G}', q \in \mathcal{M}: \mathsf{TV}(p,q) \leq 2\epsilon} L(p, \theta^*(q)). \tag{61}$$

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

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

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

**Application: bounded $k$th moments.** First suppose that the distribution $p^*$ has bounded $k$th moments, i.e. $\mathcal{G}_{\mathsf{mom},k}(\sigma) = \{p \mid \|p\|_\psi \leq \sigma\}$, where $\psi(x) = x^k$. When $k > 2$, the empirical distribution $p_n^*$ will not have boune $k$th moments until $n \geq \Omega(d^{k/2})$. This is because if we take a single sample $x_1 \sim p$ and let $v$ be a unit vector in the direction of $x_1 - \mu$, then $\mathbb{E}_{x \sim p_n^*}[\langle x - \mu, v\rangle^k] \geq \frac{1}{n}\|x_1 - \mu\|_2^k \gtrsim d^{k/2}/n$, since the norm of $\|x_1 - \mu\|_2$ is typically $\sqrt{d}$.

Consequently, it is necessary to expand the set and we will choose $\mathcal{G}' = \mathcal{M} = \mathcal{G}_{\mathsf{TV}}(\rho, \epsilon)$ for $\rho = \mathcal{O}(\sigma\epsilon^{1-1/k})$ 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 $\mathcal{O}(\sigma\epsilon^{1-1/k})$, 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 *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

$$\tilde{\psi}(x) = \left\{ \begin{array}{ccc} x^k & : & x \leq x_0 \\ kx_0^{k-1}(x - x_0) + x_0^k & : & x > x_0 \end{array} \right. \tag{62}$$

In other words, $\tilde{\psi}$ is equal to $\psi$ for $x \leq x_0$, and is the best linear lower bound to $\psi$ for $x > x_0$. Note that $\tilde{\psi}$ is $L$-Lipschitz for $L = kx_0^{k-1}$. We will eventually take $x_0 = (k^{k-1}\epsilon)^{-1/k}$ and hence $L = (1/\epsilon)^{(k-1)/k}$. Using a symmetrization argument, we will bound the truncated $\sup_{\|v\|_2 \leq 1} \mathbb{E}_{p_n^*}[\tilde{\psi}(|\langle x - \mu, v\rangle|/\sigma)]$.

**Proposition 2.32.** *Let $X_1, \ldots, X_n \sim p^*$, where $p^* \in \mathcal{G}_{\mathsf{mom},k}(\sigma)$. Then,*

$$\mathbb{E}_{X_1,\ldots,X_n \sim p^*}\Bigg[\Bigg|\sup_{\|v\|_2 \leq 1} \frac{1}{n}\sum_{i=1}^n \tilde{\psi}\Bigg(\frac{|\langle X_i - \mu, v\rangle|}{\sigma}\Bigg) - U(v)\Bigg|^k\Bigg] \leq O\Bigg(2L\sqrt{\frac{dk}{n}}\Bigg)^k, \tag{63}$$

*where $L = kx_0^{k-1}$ and $U(v)$ is a function satisfying $U(v) \leq 1$ for all $v$.*

Before proving Proposition 2.32, let us interpret its significance. Take $x_0 = (k^{k-1}\epsilon)^{-1/k}$ and hence $L = \epsilon^{1-1/k}$. Take $n$ large enough so that the right-hand-side of (63) is at most 1, which requires $n \geq \Omega(kd/\epsilon^{2-2/k})$. We then obtain a high-probability bound on the $\tilde{\psi}$-norm of $p_n^*$, i.e. the $\tilde{\psi}$-norm is at most $\mathcal{O}(\delta^{-1/k})$ with probability $1 - \delta$. This implies that $p_n^*$ is resilient with parameter $\rho = \sigma\epsilon\tilde{\psi}^{-1}(\mathcal{O}(\delta^{-1/k})/\epsilon) = 2\sigma\epsilon^{1-1/k}$. A useful bound on $\tilde{\psi}^{-1}$ is $\tilde{\psi}(-1)(z) \leq x_0 + z/L$, and since $x_0 \leq (1/\epsilon)^{-1/k}$ and $L = (1/\epsilon)^{(k-1)/k}$ in our case, we have

$$\rho \leq \mathcal{O}(\sigma\epsilon^{1-1/k}\delta^{-1/k}) \text{ with probability } 1 - \delta.$$

This matches the population-bound of $\mathcal{O}(\sigma\epsilon^{1-1/k})$, and only requires $kd/\epsilon^{2-2/k}$ samples, in contrast to the $d/\epsilon^2$ samples required before. Indeed, this sample complexity dependence is optimal (other than the factor of $k$); the only drawback is that we do not get exponential tails (we instead obtain tails of $\delta^{-1/k}$, which is worse than the $\sqrt{\log(1/\delta)}$ from before).

Now we discuss some ideas that are needed in the proof. We would like to somehow exploit the fact that $\tilde{\psi}$ is $L$-Lipschitz to prove concentration. We can do so with the following keystone result in probability theory:

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

$$\mathbb{E}_{\epsilon_{1:n}}[g(\sup_{t \in T}\sum_{i=1}^n \epsilon_i \phi(t_i))] \leq \mathbb{E}_{\epsilon_{1:n}}[g(L\sup_{t \in T}\sum_{i=1}^n \epsilon_i t_i)]. \tag{64}$$

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 $\mathbb{E}[g(Y)] \leq \mathbb{E}[g(Z)]$ for all convex increasing functions $g$. When this holds we say that $Y$ *stochastically dominates $Z$ in second order*; intuitively, it is equivalent to saying that $Z$ has larger mean than $Y$ and greater variation around its mean. For distributions supported on just two points, we can formalize this as follows:

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

$$\frac{z_1 + z_2}{2} \geq \frac{y_1 + y_2}{2} \text{ and } \max(z_1, z_2) \geq \max(y_1, y_2). \tag{65}$$

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

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

$$\frac{y_2 - z_1}{z_2 - z_1} g(z_2) + \frac{z_2 - y_2}{z_2 - z_1} g(z_1) \geq g\Big(\frac{z_2(y_2 - z_1) + z_1(z_2 - y_2)}{z_2 - z_1}\Big) = g(y_2), \tag{66}$$

$$\frac{z_2 - y_2}{z_2 - z_1} g(z_2) + \frac{y_2 - z_1}{z_2 - z_1} g(z_1) \geq g\Big(\frac{z_2(z_2 - y_2) + z_1(y_2 - z_1)}{z_2 - z_1}\Big) = g(z_1 + z_2 - y_2) \geq g(y_1). \tag{67}$$

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 $\mathbb{E}[g(Z)] \geq \mathbb{E}[g(Y)]$, as desired. We need only check that the weights $\frac{y_2 - z_1}{z_2 - z_1}$ and $\frac{z_2 - y_2}{z_2 - z_1}$ are positive. The second weight is positive by the assumption $z_2 \geq y_2$. The first weight could be negative if $y_2 < z_1$, meaning that *both* $y_1$ and $y_2$ are smaller than *both* $z_1$ and $z_2$. But in this case, the inequality $\mathbb{E}[g(Y)] \leq \mathbb{E}[g(Z)]$ trivially holds by monotonicity of $g$. This completes the proof. □

We are now ready to prove Theorem 2.33.

*Proof of Theorem 2.33.* Without loss of generality we may take $L = 1$. Our strategy will be to iteratively apply an inequality for a single $\epsilon_i$ to replace all the $\phi(t_i)$ with $t_i$ one-by-one. The inequality for a single $\epsilon_i$ is the following:

**Lemma 2.35.** *For any 1-Lipschitz function $\phi$ with $\phi(0) = 0$, any collection $T$ of ordered pairs $(a, b)$, and any convex increasing function $g$, we have*

$$\mathbb{E}_{\epsilon \sim \{-1,+1\}}[g(\sup_{(a,b) \in T} a + \epsilon\phi(b))] \leq \mathbb{E}_{\epsilon \sim \{-1,+1\}}[g(\sup_{(a,b) \in T} a + \epsilon b)]. \tag{68}$$

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

$$y_1 = a_- - \phi(b_-), \tag{69}$$

$$y_2 = a_+ + \phi(b_+), \tag{70}$$

$$z_1 = \max(a_- - b_-, a_+ - b_+), \tag{71}$$

$$z_2 = \max(a_- + b_-, a_+ + b_+). \tag{72}$$

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

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

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

$$\mathbb{E}_{\epsilon_{1:n}}[g(\sup_{t\in T}\sum_{i=1}^{n}\epsilon_i\phi(t_i))] = \mathbb{E}_{\epsilon_{1:n-1}}[\mathbb{E}_{\epsilon_n}[g(\underbrace{\sup_{t\in T}\sum_{i=1}^{n-1}\epsilon_i\phi(t_i)}_{a} + \epsilon_n\underbrace{\phi(t_n)}_{\phi(b)}) \mid \epsilon_{1:n-1}]] \tag{73}$$

$$\leq \mathbb{E}_{\epsilon_{1:n-1}}[\mathbb{E}_{\epsilon_n}[g(\sup_{t\in T}\sum_{i=1}^{n-1}\epsilon_i\phi(t_i) + \epsilon_n t_n) \mid \epsilon_{1:n-1}]] \tag{74}$$

$$= \mathbb{E}_{\epsilon_{1:n}}[[g(\sup_{t\in T}\sum_{i=1}^{n-1}\epsilon_i\phi(t_i) + \epsilon_n t_n)] \tag{75}$$

$$\vdots \tag{76}$$

$$\leq \mathbb{E}_{\epsilon_{1:n}}[g(\sup_{t\in T}\epsilon_1\phi(t_1) + \sum_{i=2}^{n}\epsilon_i t_i)] \tag{77}$$

$$\leq \mathbb{E}_{\epsilon_{1:n}}[g(\sup_{t\in T}\sum_{i=1}^{n}\epsilon_i t_i)], \tag{78}$$

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

$$\mathbb{E}_{X_1,\ldots,X_n\sim p^*}\Big[\Big|\sup_{\|v\|_2\leq 1}\frac{1}{n}\sum_{i=1}^{n}\tilde{\psi}\Big(\frac{|\langle X_i-\mu, v\rangle|}{\sigma}\Big) - \mu_{\tilde{\psi}}\Big|^k\Big] \tag{79}$$

$$\leq \mathbb{E}_{X,X'\sim p,\epsilon}\Big[\Big|\sup_{\|v\|_2\leq 1}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\Big(\tilde{\psi}\Big(\frac{|\langle X_i-\mu, v\rangle|}{\sigma}\Big) - \tilde{\psi}\Big(\frac{|\langle X'_i-\mu, v\rangle|}{\sigma}\Big)\Big)\Big|^k\Big] \tag{80}$$

$$\leq 2^k\mathbb{E}_{X\sim p,\epsilon}\Big[\Big|\sup_{\|v\|_2\leq 1}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\tilde{\psi}\Big(\frac{|\langle X_i-\mu, v\rangle|}{\sigma}\Big)\Big|^k\Big]. \tag{81}$$

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 $X$ and $X'$ are equal to each other in that case).

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

$$\mathbb{E}_{X\sim p,\epsilon}\Big[\Big|\sup_{\|v\|_2\leq 1}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\tilde{\psi}\Big(\frac{|\langle X_i-\mu, v\rangle|}{\sigma}\Big)\Big|^k\Big] \leq (L/\sigma)^k\mathbb{E}_{X\sim p,\epsilon}\Big[\Big|\sup_{\|v\|_2\leq 1}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\langle X_i-\mu, v\rangle\Big|^k\Big] \tag{82}$$

$$= (L/\sigma)^k\mathbb{E}_{X\sim p,\epsilon}\Big[\Big\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(X_i-\mu)\Big\|_2^k\Big]. \tag{83}$$

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

$$A_k\|z\|_2 \leq \mathbb{E}_\epsilon[|\sum_i \epsilon_i z_i|^k]^{1/k} \leq B_k\|z\|_2, \tag{84}$$

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

$$\mathbb{E}_{X,\epsilon}[\|\sum_{i=1}^{n}\epsilon_i(X_i-\mu)\|_2^k] \leq O(1)^k\mathbb{E}_{X,\epsilon,\epsilon'}[|\sum_{i=1}^{n}\epsilon_i\langle X_i-\mu, \epsilon'\rangle|^k]. \tag{85}$$

Next apply Rosenthal's inequality (Eq. 21), which yields that

$$\mathbb{E}_{X,\epsilon}[\sum_{i=1}^{n} \epsilon_i \langle X_i - \mu, \epsilon' \rangle|^k \mid \epsilon'] \le \mathcal{O}(k)^k \sum_{i=1}^{n} \mathbb{E}_{X,\epsilon}[|\langle X_i - \mu, \epsilon' \rangle|^k \mid \epsilon'] + \mathcal{O}(\sqrt{k})^k (\sum_{i=1}^{n} \mathbb{E}[\langle X_i - \mu, \epsilon' \rangle|^2])^{k/2} \tag{86}$$

$$\le \mathcal{O}(k)^k \cdot n\sigma^k \|\epsilon'\|_2^k + \mathcal{O}(\sqrt{kn})^k \sigma^k \|\epsilon'\|_2^k \tag{87}$$

$$= \mathcal{O}(\sigma k \sqrt{d})^k n + \mathcal{O}(\sigma \sqrt{kd})^k n^{k/2}, \tag{88}$$

where the last step uses that $\|\epsilon'\|_2 = \sqrt{d}$ and the second-to-last step uses the bounded moments of $X$. As long as $n \gg k^{k/(k-2)}$ the latter term dominates and hence plugging back into we conclude that

$$\mathbb{E}_{X,\epsilon}[\| \sum_{i=1}^{n} \epsilon_i (X_i - \mu) \|_2^k]^{1/k} = \mathcal{O}(\sigma \sqrt{kdn}). \tag{89}$$

Thus bounds the symmetrized truncated moments in (82-83) by $O(L\sqrt{kd/n})^k$, and plugging back into (81) completes the proof. □

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

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

We will let $\mathcal{G}_{\mathsf{iso}}$ 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:

$$\mathcal{G}_{\mathsf{iso}}(\sigma_1, \sigma_2) \overset{\text{def}}{=} \{p \mid \|\mathbb{E}_r[X - \mu]\|_2 \le \sigma_1 \text{ and } \|\mathbb{E}_r[(X - \mu)(X - \mu)^\top - I\| \le (\sigma_2)^2, \text{ whenever } r \le \frac{p}{1 - \epsilon}\}. \tag{90}$$

The following improvement on Lemma 2.23 implies that $p_n^* \in \mathcal{G}_{\mathsf{iso}}(\sigma_1, \sigma_2)$ for $\sigma_1 = \mathcal{O}(\epsilon\sqrt{\log(1/\epsilon)})$ and $\sigma_2 = \mathcal{O}(\sqrt{\epsilon \log(1/\epsilon)})$. [Note: the lemma below is wrong as stated. To be fixed.]

**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*

$$\Big\| \frac{1}{|S|} \sum_{i \in S}^{n} X_i X_i^\top - I \Big\| \le \mathcal{O}\Big(\sigma^2 \cdot \Big(\epsilon \log(1/\epsilon) + \frac{d + \log(1/\delta)}{n}\Big)\Big), \text{ and} \tag{91}$$

$$\Big\| \frac{1}{|S|} \sum_{i \in S}^{n} X_i \Big\|_2 \le \mathcal{O}\Big(\sigma \cdot \Big(\epsilon\sqrt{\log(1/\epsilon)} + \sqrt{\frac{d + \log(1/\delta)}{n}}\Big)\Big) \tag{92}$$

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

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

Now we bound the generalized modulus of continuity:

**Lemma 2.37.** *Suppose that $p \in \mathcal{G}_{\mathsf{iso}}(\sigma_1, \sigma_2)$ and $q \in \mathcal{G}_{\mathsf{cov}}(\sqrt{1 + \sigma_2^2})$, and furthermore $\mathsf{TV}(p, q) \le \epsilon$. Then $\|\mu(p) - \mu(q)\|_2 \le \mathcal{O}(\sigma_1 + \sigma_2\sqrt{\epsilon} + \epsilon)$.*

*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 $p \in \mathcal{G}_{\mathsf{iso}}$). We have that

$$\begin{aligned} \mathsf{Cov}_q[X] &= (1-\epsilon)\mathbb{E}_r[(X-\mu_q)(X-\mu_q)^\top] + \epsilon\mathbb{E}_{q'}[(X-\mu_q)(X-\mu_q)^\top] && (93)\\ &= (1-\epsilon)(\mathsf{Cov}_r[X] + (\mu_q-\mu_r)(\mu_q-\mu_r)^\top) + \epsilon\mathbb{E}_{q'}[(X-\mu_q)(X-\mu)q)^\top] && (94)\\ &\succeq (1-\epsilon)(\mathsf{Cov}_r[X] + (\mu_q-\mu_r)(\mu_q-\mu_r)^\top) + \epsilon(\mu_q-\mu_{q'})(\mu_q-\mu_{q'})^\top. && (95) \end{aligned}$$

A computation yields $\mu_q - \mu_{q'} = \frac{(1-\epsilon)^2}{\epsilon}(\mu_q - \mu_r)$. Plugging this into (95) and simplifying, we obtain that

$$\mathsf{Cov}_q[X] \succeq (1-\epsilon)(\mathsf{Cov}_r[X] + (1/\epsilon)(\mu_q-\mu_r)(\mu_q-\mu_r)^\top). \qquad (96)$$

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

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

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

**Proof of Lemma 2.36.** TBD

[Lecture 8]

## 2.7 Efficient Algorithms

We now turn our attention to efficient algorithms. Recall that previously we considered minimum distance functionals projecting onto sets $\mathcal{G}$ and $\mathcal{M}$ under distances $\mathsf{TV}$ and $\widetilde{\mathsf{TV}}$. Here we will show how to approximately project onto the set $\mathcal{G}_{\mathsf{cov}}(\sigma)$, the family of bounded covariance distributions, under $\mathsf{TV}$ distance. The basic idea is to write down a (non-convex) optimization problem that tries to find the projection, and then show that the cost landscape of this optimization is nice enough that all local minima are within a constant factor of the global minimum.

To study efficient computation we need a way of representing the distributions $\tilde{p}$ and $p^*$. To do this we will suppose that $\tilde{p}$ is the empirical distribution over $n$ points $x_1, \ldots, x_n$, while $p^*$ is the empirical distribution over some subset $S$ of these points with $|S| \geq (1-\epsilon)n$. Thus in particular $p^*$ is an $\epsilon$-deletion of $\tilde{p}$.

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

We will construct an efficient algorithm that, given $\tilde{p}$, outputs a distribution $q$ such that $\mathsf{TV}(q, p^*) \leq \mathcal{O}(\epsilon)$ and $\|\mathsf{Cov}_q[X]\|_2 \leq \mathcal{O}(\sigma^2)$. 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 $\tilde{p}$ onto $\mathcal{G}_{\mathsf{cov}}$, and also the covariance of $q$ is bounded by $\mathcal{O}(\sigma^2)$ 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 (by tracking the constants in our previous midpoint argument):

**Lemma 2.38.** *If $\mathsf{TV}(p,q) \leq \epsilon$, then $\|\mu(p) - \mu(q)\|_2 \leq \sqrt{\frac{\|\Sigma_q\|\epsilon}{1-\epsilon}} + \sqrt{\frac{\|\Sigma_p\|\epsilon}{1-\epsilon}}$.*

We will prove Lemma 2.38 at the end of the section.

The main result of this section is the following:

**Proposition 2.39.** *Suppose $\tilde{p}$ and $p^*$ are empirical distributions as above with $p^* \leq \tilde{p}/(1-\epsilon)$, and further suppose that $\|\mathsf{Cov}_{p^*}[X]\| \leq \sigma^2$ and $\epsilon < 1/3$. Then given $\tilde{p}$ (but not $p^*$), there is an algorithm with runtime $\mathrm{poly}(n,d)$ that outputs a $q$ with $\mathsf{TV}(p^*, q) \leq \frac{\epsilon}{1-\epsilon}$ and $\|\mathsf{Cov}_q[X]\| \leq \big(\frac{1-\epsilon}{1-3\epsilon}\big)^2\sigma^2$. In addition, $\|\mu(q) - \mu(p^*)\|_2 \leq \frac{\sqrt{4\epsilon(1-2\epsilon)}}{1-3\epsilon}\sigma$.*

Note that the bound on $\|\mu(p^*) - \mu(q)\|_2$ follows directly from the modulus bound on $\mathcal{G}_{\mathsf{cov}}(\sigma)$ together with the property $\mathsf{TV}(p^*, q) \leq \frac{\epsilon}{1-\epsilon}$.

The algorithm, `MinCovL2`, underlying Proposition 2.39 is given below; it maintains a weighted distribution $q$, which places weight $q_i$ on point $x_i$. It then computes the weighted mean and covariance, picking the weights that minimize the norm of the covariance.

**Algorithm 2** `MinCovL2`

```
1: Input: $x_1, \ldots, x_n \in \mathbb{R}^d$.
2: Find any stationary point $q$ of the optimization problem:
```

$$\min_q \sup_{\|v\|_2 \leq 1} \sum_{i=1}^n q_i \langle v, x_i - \mu_q \rangle^2, \tag{97}$$

$$s.t.\ \mu_q = \sum_i q_i x_i,$$

$$q \geq 0, \sum_i q_i = 1, q_i \leq \frac{1}{(1-\epsilon)n}$$

```
3: Output $\hat{\mu}_q$, the empirical mean for the stationary point $q$.
```

The intuition behind Algorithm 2 is as follows: the constraint $q_i \leq \frac{1}{(1-\epsilon)n}$ ensures that $q$ is an $\epsilon$-deletion of the uniform distribution over $x_1, \ldots, x_n$. Then, subject to that constraint, Algorithm 2 seeks to minimize the weighted covariance matrix: note the objective is exactly $\|\Sigma_q\|^2$.

Algorithm 2 is non-trivial to analyze, because although the constraint set is convex, the objective is non-convex: both $q_i$ and $\mu_q$ are linear in $q$, and so the overall objective (even for a fixed $v$) is thus a non-convex cubic in $q$. On the other hand, for any "reasonable" choice of $q$, $\mu_q$ should be close to $\mu_{p^*}$. If we apply this approximation–substituting $\mu_{p^*}$ for $\mu_q$–then the objective becomes convex again. So the main idea behind the proof is to show that this substitution can be (approximately) done.

Before getting into that, we need to understand what stationary points of (117) look like. In general, a stationary point is one where the gradient is either zero, or where the point is at the boundary of the constraint set and the gradient points outward into the infeasible region for the constraints.

However, the supremum over $v$ can lead to a non-differentiable function (e.g. $\max(x_1, x_2)$ is non-differentiable when $x_1 = x_2$), so these conditions need to be refined slightly to handle that. We can use something called a "Clarke subdifferential", to show that the preceding conditions at least hold for some $v$ that maximizes the supremum:

**Lemma 2.40.** *Suppose that $q$ is a stationary point of (117). Then, for any feasible $p$, there exists a $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$ such that*

$$\mathbb{E}_q[\langle v, X - \mu_q \rangle^2] \leq \mathbb{E}_p[\langle v, X - \mu_q \rangle^2]. \tag{98}$$

*Moreover, $v$ is a maximizer of the left-hand-side, i.e. $v^\top \Sigma_q v = \|\Sigma_q\|$.*

*Proof.* Let $F_v(q) = \mathbb{E}_q[\langle v, X - \mu_q \rangle^2]$. First, compute $\nabla F_v(q)$. We have

$$\frac{\partial}{\partial q_i} F(q) = \frac{\partial}{\partial q_i} \sum_{j=1}^n q_j \langle v, x_j - \mu_q \rangle^2 \tag{99}$$

$$= \langle v, x_i - \mu_q \rangle^2 + 2\Big(\sum_{j=1}^n q_j \langle v, x_j - \mu_q \rangle\Big) \frac{\partial \mu_q}{\partial q_i} \tag{100}$$

$$= \langle v, x_i - \mu_q \rangle^2, \tag{101}$$

where the last equality is because $\sum_j q_j (x_j - \mu_q) = 0$. Consequently, $\nabla F_v(q)_i = \langle v, x_i - \mu_q \rangle^2$.

Now, let $F(q) = \max_{\|v\|_2 = 1} F_v(q) = \|\Sigma_q\|$. If the maximizing $v$ is unique and equal to $v^*$, then $\nabla F(q) = \nabla F_{v^*}(q)$, and $q$ is a stationary point if and only if $\sum_i (q_i - p_i) \nabla F_{v^*}(q)_i \leq 0$ for all feasible $p$, or equivalently $\mathbb{E}_q[\langle v^*, X - \mu_q \rangle^2] - \mathbb{E}_p[\langle v^*, X - \mu_q \rangle^2] \leq 0$, which is exactly the condition (98).

Suppose (the harder case) that the maximizing $v$ is not unique. Then $F$ is not differentiable at $q$, but the Clark subdifferential is the convex hull of $\nabla F_v(q)$ for all maximizing $v$'s. Stationarity implies that $\sum_i (q_i - p_i) g_i \leq 0$ for some $g$ in this convex hull, and thus by convexity that $\sum_i (q_i - p_i) \nabla F_{v^*}(q)_i \leq 0$ for some maximizing $v^*$. This gives us the same desired condition as before and thus completes the proof. $\square$

Given Lemma 2.40, we are in a better position to analyze Algorithm 2. In particular, for any $p$ (we will eventually take the global minimizer $\overline{p}$ of (117)), Lemma 2.40 yields

$$\begin{aligned}
\|\mathsf{Cov}_q\| &= \mathbb{E}_q[\langle v, X - \mu_q \rangle^2] && (102)\\
&\leq \mathbb{E}_p[\langle v, X - \mu_q \rangle^2] && (103)\\
&= \mathbb{E}_p[\langle v, X - \mu_p \rangle^2] + \langle v, \mu_p - \mu_q \rangle^2 && (104)\\
&\leq \|\mathsf{Cov}_p\| + \|\mu_p - \mu_q\|_2^2. && (105)
\end{aligned}$$

The $\|\mu_p - \mu_q\|_2^2$ quantifies the "error due to non-convexity"–recall that if we replace $\mu_q$ with a fixed $\mu_p$ in (117), the problem becomes convex, and hence any stationary point would be a global minimizer. The distance $\|\mu_p - \mu_q\|_2^2$ is how much we pay for this discrepancy.

Fortunately, $\mu_p - \mu_q$ is small, precisely due to the modulus of continuity! We can show that any feasible $p, q$ for (117) satisfies $\mathsf{TV}(p, q) \leq \frac{\epsilon}{1-\epsilon}$ (see Lemma 2.41), hence Lemma 2.38 gives $\|\mu_p - \mu_q\|_2 \leq \sqrt{\frac{\|\Sigma_q\|\epsilon}{1-2\epsilon}} + \sqrt{\frac{\|\Sigma_p\|\epsilon}{1-2\epsilon}}$. Plugging back in to (105), we obtain

$$\|\mathsf{Cov}_q\| \leq \|\mathsf{Cov}_p\| + \frac{\epsilon}{1-2\epsilon}\Big(\sqrt{\|\Sigma_p\|} + \sqrt{\|\Sigma_q\|}\Big)^2. \tag{106}$$

For fixed $\|\mathsf{Cov}_p\|$, we can view this as a quadratic inequality in $\sqrt{\|\Sigma_p\|}$. Solving the quadratic then yields

$$\|\mathsf{Cov}_q\| \leq \Big(\frac{1-\epsilon}{1-3\epsilon}\Big)^2 \|\mathsf{Cov}_p\|. \tag{107}$$

In particular, taking $p$ to be the global minimum $\overline{p}$ of (117), we have $\|\mathsf{Cov}_{\overline{p}}\| \leq \|\mathsf{Cov}_{p^*}\| \leq \sigma^2$, so $\|\mathsf{Cov}_q\| \leq \Big(\frac{1-\epsilon}{1-3\epsilon}\Big)^2 \sigma^2$. Plugging back into Lemma 2.38 again, we then have

$$\|\mu_q - \mu_{p^*}\|_2 \leq \sqrt{\frac{\epsilon}{1-2\epsilon}}\Big(\sigma + \frac{1-\epsilon}{1-3\epsilon}\sigma\Big) = 2\frac{\sqrt{1-2\epsilon}}{1-3\epsilon}\sqrt{\epsilon}\sigma, \tag{108}$$

which proves Proposition 2.39.

### 2.7.1 Lower Bound (Breakdown at $\epsilon = 1/3$)

The $1 - 3\epsilon$ in the denominator of our bound means that Proposition 2.39 becomes vacuous once $\epsilon \geq 1/3$. Is this necessary? We will show that it indeed is.

Specifically, when $\epsilon = 1/3$, it is possible to have:

$$\begin{aligned}
p^* &= \frac{1}{2}\delta_{-a} + \frac{1}{2}\delta_0, && (109)\\
\tilde{p} &= \frac{1}{3}\delta_{-a} + \frac{1}{3}\delta_0 + \frac{1}{3}\delta_b, && (110)\\
q &= \frac{1}{2}\delta_0 + \frac{1}{2}\delta_b, && (111)
\end{aligned}$$

where $q$ is a stationary point no matter how large $b$ is. In particular, $\mu_q = \frac{b}{2}$ can be arbitrarily far away from the true mean of $-\frac{a}{2}$.

To see this more intuitively, note that an equivalent minimization problem to (117) would be to minimize $\sum_{i=1}^n q_i (x_i - \mu)^2$ with respect to both $q_i$ and $\mu$ (since the minimizer for fixed $q$ is always at $\mu = \mu_q$). Therefore, a stationary point is one such that:

- The $q_i$ are concentrated on the $(1-\epsilon)n$ smallest values of $(x_i - \mu)^2$

- $\mu$ is equal to $\mu_q$

The distribution $q$ clearly satisfies this: we have $\mu_q = b/2$, and both 0 and $b$ are closer to $\mu_q$ than $-a$ is.

This also shows why the breakdown point is 1/3 and not smaller. If $\epsilon$ were slightly smaller than 1/3, then some of the mass of $q$ would have to remain on $\delta_{-a}$. Then as $b$ increases, the mean $\mu_q$ would increase more slowly, and eventually $-a$ would be closer to $\mu_q$ than $b$.

### 2.7.2 Auxiliary Lemmas

*Proof of Lemma 2.38.* Note that the proof of Lemma 2.1 implies that if $E$ is an event with $q(E) \geq 1-\epsilon$, then $\|\mathbb{E}_q[X] - \mathbb{E}_q[X \mid E]\| \leq \sqrt{\|\Sigma_q\|\frac{\epsilon}{1-\epsilon}}$. Now if $q, p$ satisfy $\mathsf{TV}(p, q) \leq \epsilon$, there is a midpoint $r$ that is an $\epsilon$-deletion of both $p$ and $q$. Applying the preceding result, we thus have $\|\mathbb{E}_q[X] - \mathbb{E}_r[X]\|_2 \leq \sqrt{\|\Sigma_q\|\frac{\epsilon}{1-\epsilon}}$. Similarly $\|\mathbb{E}_p[X] - \mathbb{E}_r[X]\|_2 \leq \sqrt{\|\Sigma_p\|\frac{\epsilon}{1-\epsilon}}$. The result then follows by the triangle inequality. □

**Lemma 2.41.** *Suppose that $q, q'$ are both $\epsilon$-deletions of a distribution $p$. Then $\mathsf{TV}(q, q') \leq \frac{\epsilon}{1-\epsilon}$.*

*Proof.* Conceptually, the reason Lemma 2.41 is true is that $q'$ can be obtained from $q$ by first adding an $\epsilon$-fraction of points (to get to $p$), then deleting an $\epsilon$-fraction. Since $\mathsf{TV} \leq \epsilon$ exactly allows an $\epsilon$-fraction of of additions and deletions, this should yield the result. The reason we get $\frac{\epsilon}{1-\epsilon}$ is because $q$ and $q'$ are only a $(1-\epsilon)$-“fraction” of $p$, so the $\epsilon$-deletions are more like $\frac{\epsilon}{1-\epsilon}$-deletions relative to $q$ and $q'$.

To be more formal, for any set $A$, note that we have

$$q(A) \leq \frac{p(A)}{1-\epsilon} \text{ and } q'(A) \leq \frac{r(A)}{1-\epsilon}. \tag{112}$$

Also, using $A^c$ instead of $A$, we also get

$$q(A) \geq \frac{p(A)-\epsilon}{1-\epsilon} \text{ and } q'(A) \geq \frac{p(A)-\epsilon}{1-\epsilon}. \tag{113}$$

Combining these inequalities yields

$$q(A) \leq \frac{\epsilon + (1-\epsilon)q'(A)}{1-\epsilon} \leq \frac{\epsilon}{1-\epsilon} + q'(A), \tag{114}$$

and similarly $q'(A) \leq \frac{\epsilon}{1-\epsilon} + q(A)$, which together implies $|q(A) - q'(A)| \leq \frac{\epsilon}{1-\epsilon}$. Since this holds for all $A$, we obtain our $\mathsf{TV}$ distance bound. □

[Lecture 9]

## 2.8 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 $\mathcal{O}(\sigma\sqrt{\epsilon})$ error, as long as we can approximately solve the analogous eigenvector problem. To formalize this, we will make use of the *dual norm*:

**Definition 2.42.** Given a norm $\|\cdot\|$, the *dual norm* $\|\cdot\|_*$ is defined as

$$\|u\|_* = \sup_{\|v\|_2 \leq 1} \langle u, v \rangle. \tag{115}$$

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 2.43.** *If $\|\cdot\|$ is a norm on a finite-dimensional vector space, then $\|\cdot\|_{**} = \|\cdot\|$.*

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

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

**Proposition 2.44.** *If $p \in \mathcal{G}_{\mathsf{cov}}(\sigma, \|\cdot\|)$ and $r \leq \frac{p}{1-\epsilon}$, then $\|\mu(r) - \mu(p)\| \leq \sqrt{\frac{2\epsilon}{1-\epsilon}}\sigma$. In other words, all distributions in $\mathcal{G}_{\mathsf{cov}}(\sigma, \|\cdot\|)$ are $(\epsilon, \mathcal{O}(\sigma\sqrt{\epsilon}))$-resilient.*

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

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

**Definition 2.45.** A set $\mathcal{C}$ of positive semidefinie matrices is a $\kappa$-approximate relaxation if $\mathcal{C}$ contains $vv^\top$ for all $v$ satisfying $\|v\|_* \leq 1$, and furthermore

$$\sup_{M \in \mathcal{C}} \langle M, \Sigma \rangle \leq \kappa \sup_{\|v\|_* \leq 1} v^\top \Sigma v \text{ for all } \Sigma \succeq 0. \tag{116}$$

Thus a $\kappa$-approximate relaxation over-approximates $\langle vv^\top, \Sigma\rangle$, but also it underapproximates it within a factor of $\kappa$. Given such a relaxation, we have the following analog to Algorithm 2:

**Algorithm 3** `MinCovNorm`

```
1: Input: $x_1, \ldots, x_n \in \mathbb{R}^d$ and $\kappa$-approximate relaxation $\mathcal{C}$.
2: Find any stationary point $q$ of the optimization problem:
       $$\min_q \sup_{M \in \mathcal{C}} \sum_{i=1}^n q_i (x_i - \mu_q)^\top M (x_i - \mu_q), \tag{117}$$
       $$s.t.\ \mu_q = \sum_i q_i x_i,$$
       $$q \geq 0, \sum_i q_i = 1, q_i \leq \frac{1}{(1-\epsilon)n}$$
3: Output $\hat{\mu}_q$, the empirical mean for the stationary point $q$.
```

Algorithm 3 outputs an estimate of the mean with error $\mathcal{O}(\sigma\sqrt{\kappa\epsilon})$. The proof is almost exactly the same as Algorithm 2; the main difference is that we need to ensure that $\langle \Sigma_p, M\rangle$, 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 relaxation $\mathcal{C}$.

**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

$$\text{maximize } v^\top \Sigma v \text{ subject to } \|v\|_\infty \leq 1. \tag{118}$$

The issue with (118) is that it is not concave in $v$ because of the quadratic function $v^\top \Sigma v$. However, note that $v^\top \Sigma v = \langle \Sigma, vv^\top \rangle$. Therefore, if we replace $v$ with the variable $M = vv^\top$, then we can re-express the optimization problem as

$$\text{maximize } \langle \Sigma, M \rangle \text{ subject to } M_{jj} \leq 1 \text{ for all j}, M \succeq 0, \operatorname{rank}(M) = 1. \tag{119}$$

Here the first constraint is a translation of $\|v\|_\infty \leq 1$, while the latter two constrain $M$ to be of the form $vv^\top$.

This is almost convex in $M$, except for the constraint $\mathrm{rank}(M) = 1$. If we omit this constraint, we obtain the optimization

$$\begin{aligned} \text{maximize } & \langle \Sigma, M \rangle \\ \text{subject to } & M_{jj} = 1 \text{ for all } j, \\ & M \succeq 0. \end{aligned} \tag{120}$$

Note that here we replace the constraint $M_{jj} \leq 1$ with $M_{jj} = 1$; this can be done because the maximizer of (120) will always have $M_{jj} = 1$ for all $j$. For brevity we often write this constraint as $\mathrm{diag}(M) = 1$.

The problem (120) 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 (120) will always satisfy $\langle \Sigma, M^* \rangle \geq \sup_{\|v\|_\infty \leq 1} v^\top \Sigma v$ 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 2.46.** *If $\Sigma \succeq 0$, then the value of* (120) *is at most $\frac{\pi}{2} \sup_{\|v\|_\infty \leq 1} v^\top \Sigma v$.*

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 2.46 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 2.46.* The proof involves two key relations. To describe the first, given a matrix $X$ let $\arcsin[X]$ denote the matrix whose $i, j$ entry is $\arcsin(X_{ij})$ (i.e. we apply arcsin element-wise). Then we have (we will show this later)

$$\max_{\|v\|_\infty \leq 1} v^\top \Sigma v = \max_{X \succeq 0, \mathrm{diag}(X)=1} \frac{2}{\pi} \langle \Sigma, \arcsin[X] \rangle. \tag{121}$$

The next relation is that

$$\arcsin[X] \succeq X. \tag{122}$$

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

$$\max_{M \succeq 0, \mathrm{diag}(M)=1} \langle \Sigma, M \rangle \leq \max_{M \succeq 0, \mathrm{diag}(M)=1} \langle \Sigma, \arcsin[M] \rangle = \frac{\pi}{2} \max_{\|v\|_\infty \leq 1} v^\top \Sigma v. \tag{123}$$

We will therefore focus on establishing (121) and (122).

To establish (121), we will show that any $X$ with $X \succeq 0$, $\mathrm{diag}(X) = 1$ can be used to produce a probability distribution over vectors $v$ such that $\mathbb{E}[v^\top \Sigma v] = \frac{2}{\pi} \langle \Sigma, \arcsin[X] \rangle$.

First, by Graham/Cholesky 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 $v_i = \mathrm{sign}(\langle u_i, g \rangle)$ for a Gaussian random variable $g \sim \mathcal{N}(0, I)$.

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 $v_i$ 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 $\mathbb{P}[v_j v_j = -1] = \frac{\theta}{\pi}$ and hence

$$\mathbb{E}_g[v_i v_j] = (1 - \frac{\theta}{\pi}) - \frac{\theta}{\pi} = \frac{2}{\pi}(\frac{\pi}{2} - \theta) = \frac{2}{\pi} \arcsin(\langle u_i, u_j \rangle), \tag{124}$$

as desired. Therefore, we can always construct a distribution over $v$ for which $\mathbb{E}[v^\top \Sigma v] = \frac{2}{\pi} \langle \Sigma, \arcsin[M] \rangle$, hence the right-hand-side of (121) 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 $\frac{2}{\pi} \arcsin[vv^\top] = vv^\top$. Thus the left-hand-side is at most the right-hand-side, and so the equality (121) indeed holds.

We now turn our attention to establishing (122). For this, let $X^{\odot k}$ denote the matrix whose $i, j$ entry is $X_{ij}^k$ (we take element-wise power). We require the following lemma:

**Lemma 2.47.** *For all $k \in \{1, 2, \dots\}$, if $X \succeq 0$ then $X^{\odot k} \succeq 0$.*

*Proof.* The matrix $X^{\odot k}$ 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 $\lambda_{i_1} \cdots \lambda_{i_k}$ where $\lambda_i$ are the eigenvalues of $X$), hence so is $X^{\odot k}$ since submatrices of PSD matrices are PSD. □

With this in hand, we also make use of the Taylor series for $\arcsin(z)$: $\arcsin(z) = \sum_{n=0}^{\infty} \frac{(2n)!}{(2^n n!)^2} \frac{z^{2n+1}}{2n+1} = z + \frac{z^3}{6} + \cdots$. Then we have

$$\arcsin[X] = X + \sum_{n=1}^{\infty} \frac{(2n)!}{(2^n n!)^2} \frac{1}{2n+1} X^{\odot(2n+1)} \succeq X, \tag{125}$$

as was to be shown. This completes the proof. □

**Alternate proof (by Mihaela Curmei):** We can also show that $X^{\odot k} \succeq 0$ more directly. Specifically, we will show that if $A, B \succeq 0$ then $A \odot B \succeq 0$, 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

$$\begin{aligned} A \odot B &= (\sum_i \lambda_i u_i u_i^\top) \odot (\sum_j \nu_j v_j v_j^\top) && (126) \\ &= \sum_{i,j} \lambda_i \nu_j (u_i u_i^\top) \odot (v_j v_j^\top) && (127) \\ &= \sum_{i,j} \underbrace{\lambda_i \nu_j}_{\geq 0} \underbrace{(u_i \odot v_j)(u_i \odot v_j)^\top}_{\succeq 0}, && (128) \end{aligned}$$

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

[Bonus Material (Optional)]

## 2.9 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 $\max_{M \succeq 0, \mathrm{diag}(M)=1} \langle M, \Sigma \rangle$. 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 $\mathcal{O}(\epsilon^{1-1/k})$ when $p^*$ has "certifiably bounded" $k$th moments (recall that we earlier achieved error $\mathcal{O}(\epsilon^{1-1/k})$ for bounded $k$th moments but did not have an efficient algorithm).

A **semidefinite program** is an optimization problem of the form

$$\begin{aligned} \text{maximize } & \langle A, X \rangle && (129) \\ \text{subject to } & X \succeq 0, \\ & \langle X, B_1 \rangle \leq c_1, \\ & \quad \vdots \\ & \langle X, B_m \rangle \leq c_m. \end{aligned}$$

Here $\langle X, Y \rangle = \mathrm{tr}(X^T Y) = \sum_{ij} X_{ij} Y_{ij}$ 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 $B_1, \dots, B_m$ are $n \times n$ matrices while the $c_j$ are scalars. We can equally well minimize as maximize by replacing $A$ with $-A$.

While (129) 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 $\mathcal{L}(X) \succeq 0$ for any linear function $\mathcal{L}$, by using linear equality constraints to enforce the linear relations implied by $\mathcal{L}$. 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 $\left[\begin{array}{cc} \mathcal{L}_1(X) & 0 \\ 0 & \mathcal{L}_2(X) \end{array}\right]$ when $k = 2$. As an example of these observations, the following (arbitrarily-chosen) optimization problem is also a semidefinite program:

$$\begin{aligned} \operatorname*{minimize}_{x,M,Y} \ & a^\top x + \langle A_1, M \rangle + \langle A_2, Y \rangle \\ \text{subject to } & M + Y \succeq \Sigma \\ & \operatorname{diag}(M) = 1 \\ & \operatorname{tr}(Y) \leq 1 \\ & Y \succeq 0 \\ & \left[\begin{array}{cc} 1 & x^\top \\ x & M \end{array}\right] \succeq 0 \end{aligned} \tag{130}$$

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

**Semidefinite constraints as quadratic polynomials.** An alternative way of viewing the constraint $M \succeq 0$ is that the polynomial $p_M(v) = v^\top M v$ is non-negative for all $v \in \mathbb{R}^d$. More generally, if we have a non-hogoneous polynomial $p_{M,y,c}(v) = v^\top M v + y^\top v + c$, we have $p_{M,y,c}(v) \geq 0$ for all $v$ if and only if $M' \succeq 0$ for $M' = \left[\begin{array}{cc} c & y^\top/2 \\ y/2 & M \end{array}\right] \succeq 0$.

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

$$\begin{aligned} & \text{minimize } \lambda \\ & \text{subject to } \lambda I - M \succeq 0 \text{ (equivalently, } v^\top(\lambda I - M)v \geq 0 \text{ for all } v) \end{aligned} \tag{131}$$

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

$$\langle M_4, v^{\otimes 4} \rangle = \mathbb{E}_{x \sim p}[\langle x - \mu, v \rangle^4]. \tag{132}$$

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

$$\begin{aligned} & \text{minimize } \lambda \\ & \text{subject to } \lambda(v_1^2 + \cdots + v_d^2)^2 - p_M(v) \geq 0 \text{ for all } v \in \mathbb{R}^d \end{aligned} \tag{133}$$

Unfortunately, in constrast to (130), (133) is NP-hard to solve in general. We will next see a way to approximate (133) via a technique called *sum-of-squares programming*, which is a way of approximately reducing polynomial programs such as (133) to a large but polynomial-size semidefinite program.

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

$$q(x) = 2x^4 + 2x^3 - x^2 + 5 \tag{134}$$

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

$$q(x) = \frac{1}{2}(2x^2 + x - 3)^2 + \frac{1}{2}(3x + 1)^2 \tag{135}$$

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

$$q(x) = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}^\top \underbrace{\begin{bmatrix} 5 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & 2 \end{bmatrix}}_{M} \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}. \tag{136}$$

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

$$q(x) = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}^\top \left( \frac{1}{2} \begin{bmatrix} -3 \\ 1 \\ 2 \end{bmatrix} \begin{bmatrix} -3 \\ 1 \\ 2 \end{bmatrix}^\top + \frac{1}{2} \begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix}^\top \right) \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}, \tag{137}$$

i.e. we can decompose the matrix $M$ defining $q(x) = [1; x; x^2]^\top M [1; x;^2\,]$ 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:

$$q(x) \text{ is a sum of squares } \sum_{j=1}^{k} q_j(x)^2 \iff M(a) \succeq 0 \text{ for some } a \in \mathbb{R}. \tag{138}$$

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 $q(x, y) = a_{40}x^4 + a_{31}x^3y + a_{22}x^2y^2 + a_{13}xy^3 + a_{04}y^4 + a_{30}x^3 + a_{21}x^2y + a_{12}xy^2 + a_{03}y^3 + a_{20}x^2 + a_{11}xy + a_{02}y^2 + a_{10}x + a_{01}y + a_{00}$ 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

$$q(x, y) = \begin{bmatrix} x^2 \\ xy \\ y^2 \\ x \\ y \\ 1 \end{bmatrix}^\top \begin{bmatrix} a_{40} & a_{31}/2 & -b & a_{30}/2 & -c & -b' \\ a_{31}/2 & a_{22} + 2b & a_{13}/2 & a_{21}/2 + c & -c' & -c'' \\ -b & a_{13}/2 & a_{04} & a_{21}/2 + c' & a_{03}/2 & -b'' \\ a_{30}/2 & a_{21}/2 + c & a_{21}/2 + c' & a_{20} + 2b' & a_{11}/2 + c'' & a_{10}/2 \\ -c & -c' & a_{03}/2 & a_{11}/2 + c'' & a_{02} + 2b'' & a_{01}/2 \\ -b' & -c'' & -b'' & a_{10}/2 & a_{01}/2 & a_{00} \end{bmatrix} \begin{bmatrix} x^2 \\ xy \\ y^2 \\ x \\ y \\ 1 \end{bmatrix} \tag{139}$$

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 $q(x_1, \dots, x_d)$ 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 $\binom{d+t}{t} = \mathcal{O}((d+t)^t)$.

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

$$\begin{aligned} &\underset{y}{\text{maximize}}\ c^\top y && (140)\\ &\text{subject to } q_1 \succeq_{\mathrm{sos}} 0, \dots, q_m \succeq_{\mathrm{sos}} 0, \end{aligned}$$

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 $\lambda \|v\|_2^k - \langle M_{2k}, v^{\otimes 2k} \rangle \succeq_{\mathrm{sos}} 0$. However, we actually have a bit more flexibility and it turns out that the best program to use is

$$\begin{aligned} &\text{minimize } \lambda && (141)\\ &\text{subject to } \lambda - \langle M_{2k}, v^{\otimes 2k} \rangle + (\|v\|_2^2 - 1)q(v) \succeq_{\mathrm{sos}} 0 \text{ for some } q \text{ of degree at most } 2k-2 \end{aligned}$$

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 (141) 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 $\mathcal{O}(\sigma\epsilon^{1-1/2k})$. 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 $\tau_i = \langle T, (x_i - \hat{\mu})^{\otimes 2k} \rangle$.

To obtain $T$ requires computing the dual of (141), 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.

## 2.10 Sum-of-Squares Certifiably from the Poincaré inequality

We now turn our attention to bounding the value of (141). Ignoring finite-sample issues, our goal is to identify assumptions on $p$ such that $M_{2k}(p) \overset{\text{def}}{=} \mathbb{E}_{X \sim p}[(X - \mu)^{\otimes 2k}]$ yields a small value for (141).

Before doing so, we will introduce some machinery for establishing bounds on (141). The main idea is that of a sum-of-squares proof:

**Definition 2.48.** A polynomial inequality $p(v) \le q(v)$ has a *sum-of-squares proof* if $q(v) - p(v) \succeq_{\mathrm{sos}} 0$. We will also denote this as $q(v) \succeq_{\mathrm{sos}} p(v)$ or $p(v) \preceq_{\mathrm{sos}} q(v)$.

The usefulness of this perspective is that the relation $\preceq_{\mathrm{sos}}$ satisfies many of the same properties as $\le$:

- If $p_1 \preceq_{\mathrm{sos}} p_2$ and $p_2 \preceq_{\mathrm{sos}} p_3$, then $p_1 \preceq_{\mathrm{sos}} p_3$.
- If $p_1 \preceq_{\mathrm{sos}} q_1$ and $p_2 \preceq_{\mathrm{sos}} q_2$, then $p_1 + p_2 \preceq_{\mathrm{sos}} q_1 + q_2$.
- If $p_1 \succeq_{\mathrm{sos}} 0$ and $p_2 \succeq_{\mathrm{sos}} 0$, then $p_1 p_2 \succeq_{\mathrm{sos}} 0$.

- If $p_1 \preceq_{\mathrm{sos}} p_2$, $q_1 \preceq_{\mathrm{sos}} q_2$, and $p_2, q_1 \succeq_{\mathrm{sos}} 0$, then $p_1 q_1 \preceq_{\mathrm{sos}} p_2 q_2$.
- 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 $p \preceq q$ 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 $p_1, p_2 \succeq_{\mathrm{sos}} 0 \implies p_1 p_2 \succeq_{\mathrm{sos}} 0$. Indeed we have

$$p_1(v)p_2(v) = (\sum_i p_{1i}(v)^2)(\sum_j p_{2j}(v)^2) = \sum_{ij} (p_1 i(v) p_{2j}(v))^2 \succeq_{\mathrm{sos}} 0 \tag{142}$$

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

$$p_2 q_2 - p_1 q_1 = p_2(q_2 - q_1) + (p_2 - p_1) q_1 \succeq_{\mathrm{sos}} 0, \tag{143}$$

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

In view of this, we can reframe bounding (141) as the following goal:

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

**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 $p = \mathcal{N}(\mu, \Sigma)$, then

$$\begin{aligned}
\langle M_{2k}(\mathcal{N}(\mu, \Sigma)), v^{\otimes 2k} \rangle &= \langle M_{2k}(\mathcal{N}(0, I)), (\Sigma^{1/2} v)^{\otimes 2k} \rangle & (144)\\
&= \big(\prod_{i=1}^k (2i-1)\big) \langle \mathcal{I}, (\Sigma^{1/2} v)^{\otimes 2k} \rangle & (145)\\
&= \big(\prod_{i=1}^k (2i-1)\big) \|\Sigma^{1/2} v\|_2^{2k} & (146)\\
&\leq (2k)^k \|\Sigma\|^k \|v\|_2^{2k}, & (147)
\end{aligned}$$

so we may take $\lambda = (2k\|\Sigma\|)^k$. (Here $\mathcal{I}$ 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

$$\mathsf{Var}_{x \sim p}[f(x)] \leq \sigma^2 \mathbb{E}_{x \sim p}[\|\nabla f(x)\|_2^2] \tag{148}$$

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" (regines 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 $\mathcal{N}(\mu, \sigma^2)$ is Poincaré with constant $\sigma$.
- If $p$, $q$ are $\sigma$-Poincaré then their product $p \times q$ is $\sigma$-Poincaré. In particular a multivariate Gausssian $\mathcal{N}(\mu, \sigma^2 I)$ 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 $\mathcal{N}(\mu, \Sigma)$ 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 $Z = \mathcal{N}(0, \tau^2 I)$ for $\tau \geq 2R$. Then $X + Z$ is $(\tau\sqrt{e})$-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 $X \sim \mathcal{N}(0, I)$ 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 $\mathsf{Var}[f(X, Y)] \approx d^2$. 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 $\mathcal{O}(1)$, 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: $\mathbb{P}[|f(x) - \mathbb{E}[f(x)]| \geq t] \leq 6 \exp(-t/(\sigma L))$ 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 $p(A)p(A_\epsilon^c) \leq 36 \exp(-\epsilon/\sigma)$.

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

To prove the second property, let $f(x) = \min(\inf_{y \in A} \|x - y\|_2 . \epsilon)$. Then $f$ is Lipschitz, is 0 on $A$, and is $\epsilon$ on $A_\epsilon^c$. 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 $p(A_\epsilon^c) = p(f(X) = \epsilon) \leq 6 \exp(-(\epsilon - \mu))/\sigma)$. Multiplying these together yields the claimed result.

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

**Theorem 2.49.** *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 $\mathsf{Var}[f(X)] \leq C_k \sigma^{2k} \mathbb{E}[\|\nabla^k f(X)\|_F^2]$.*

Note that $k = 1$ is the original Poincaré property, so we can think of Theorem 2.49 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 2.49, 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 Latała (2006). We will use Theorem 2.49 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 $M_{2k}(v) \overset{\text{def}}{=} \mathbb{E}_p[\langle x - \mu, v \rangle^{2k}] \preceq_{\text{sos}} C_k' \sigma^{2k} \|v\|_2^{2k}$ 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

$$M_k = \mathbb{E}[(x-\mu)^{\otimes k}], \tag{149}$$

$$M_k(v) = \langle M_k, v^{\otimes k}\rangle = \mathbb{E}[\langle x-\mu, v\rangle^k]. \tag{150}$$

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

$$\mathbb{E}_p[\langle x-\mu, v\rangle^2] = \mathsf{Var}[f_v(x)] \le \sigma^2\mathbb{E}[\|\nabla f_v(x)\|_2^2] = \sigma^2\mathbb{E}[\|v\|_2^2] = \sigma^2\|v\|_2^2. \tag{151}$$

Thus $M_2(v) \le \sigma^2\|v\|_2^2$ (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.

*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 $\nabla f_v(x) = 2\langle x-\mu, v\rangle v$ and hence $\mathbb{E}[\nabla f_v(x)] = 0$. We also have $\nabla^2 f_v(x) = 2v \otimes v$. Thus applying Theorem 2.49 we obtain

$$\mathsf{Var}[f_v(x)] \le C_2\sigma^4\mathbb{E}[\|2v\otimes v\|_F^2] = 4C_2\sigma^4\|v\|_2^4. \tag{152}$$

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

$$M_4(v) = (M_4(v) - M_2(v)^2) + M_2(v)^2 \tag{153}$$

$$\le 4C_2\sigma^4\|v\|_2^4 + \sigma^4\|v\|_2^4 = (4C_2+1)\sigma^4\|v\|_2^4. \tag{154}$$

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

For this we will apply Theorem 2.49 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 2.49 we have

$$\mathsf{Var}[f_A(x)] \le C_2\sigma^4\mathbb{E}[\|2A\|_F^2] = 4C_2\sigma^4\|A\|_F^2. \tag{155}$$

On the other hand, we have $\mathsf{Var}[f_A(x)] = \langle M_4, A\otimes A\rangle - \langle M_2, A\rangle^2 = \langle M_4 - M_2\otimes M_2, A\otimes A\rangle$. Thus (155) implies that

$$\langle M_4 - M_2\otimes M_2, A\otimes A\rangle \le 4C_2\sigma^4\|A\|_F^2. \tag{156}$$

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 $4C_2\sigma^4\|v\|_2^4 - \langle M_4 - M_2\otimes M_2, v^{\otimes 4}\rangle \preceq_{\mathrm{sos}} 0$, giving us our desired sum-of-squares proof. To recap, we have:

$$M_4(v) = (M_4(v) - M_2(v)^2) + M_2(v)^2 \tag{157}$$

$$\preceq_{\mathrm{sos}} 4C_2\sigma^4\|v\|_2^4 + \sigma^4\|v\|_2^4 = (4C_2+1)\sigma^4\|v\|_2^4, \tag{158}$$

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

*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 $\nabla f_v(x) = 3\langle x-\mu, v\rangle^2 v$ and so $\mathbb{E}[\nabla f_v(x)] = 3M_2(v)v \neq 0$. We instead should take

$$f_v(x) = \langle x-\mu, v\rangle^3 - 3M_2(v)\langle x-\mu, v\rangle, \text{ which yields} \tag{159}$$

$$\mathbb{E}[\nabla f_v(x)] = \mathbb{E}[3\langle x-\mu, v\rangle^2 v - 3M_2(v)v] = 0, \tag{160}$$

$$\mathbb{E}[\nabla^2 f_v(x)] = \mathbb{E}[6\langle x-\mu, v\rangle(v\otimes v)] = 0, \tag{161}$$

$$\nabla^3 f_v(x) = 6(v\otimes v\otimes v). \tag{162}$$

Applying Theorem 2.49 to $f_v(x)$ yields

$$\mathsf{Var}[f_v(x)] \le C_3\sigma^6\|6(v\otimes v\otimes v)\|_F^2 = 36C_3\sigma^6\|v\|_2^6. \tag{163}$$

We can additionally compute

$$\mathsf{Var}[f_v(x)] = \mathbb{E}[(\langle x-\mu, v\rangle^3 - 3M_2(v)\langle x-\mu, v\rangle)^2] - \mathbb{E}[\langle x-\mu, v\rangle^3 - 3M_2(v)\langle x-\mu, v\rangle]^2 \tag{164}$$

$$= M_6(v) - 6M_2(v)M_4(v) + 9M_2(v)^3 - M_3(v)^2. \tag{165}$$

Since our goal is to bound $M_6(v)$, we re-arrange to obtain

$$M_6(v) = \mathsf{Var}[f_v(x)] + 6M_2(v)M_4(v) + M_3(v)^2 - 9M_2(v)^2 \tag{166}$$

$$\leq 36C_3\sigma^6\|v\|_2^6 + 6(\sigma^2\|v\|_2^2)(C_2'\sigma^4\|v\|_2^4) + M_3(v)^2 + 0 \tag{167}$$

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 $M_6(v) \leq (36C_3 + 12C_2')\sigma^6\|v\|_2^6$.

We now turn this into a sum-of-squares proof. We need to show the following four relations:

$$(i)\ \mathsf{Var}[f_v(x)] \preceq_{\text{sos}} 36C_3\sigma^6\|v\|_2^6, \quad (ii)\ M_2(v)M_4(v) \preceq_{\text{sos}} (\sigma^2\|v\|_2^2)(C_2'\sigma^4\|v\|_2^4), \tag{168}$$

$$(iii)\ M_3(v) \preceq_{\text{sos}} M_2(v)M_4(v), \quad (iv)\ -9M_2(v)^2 \preceq_{\text{sos}} 0. \tag{169}$$

The relation (ii) again follows by the product property of $\preceq_{\text{sos}}$, while $-9M_2(v)^2 \preceq_{\text{sos}} 0$ 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 $f_A(x) = \langle A, (x-\mu)^{\otimes 3} - 3M_2 \otimes (x-\mu)\rangle$. Then just as before we have $\mathbb{E}[\nabla f_A(x)] = 0$, $\mathbb{E}[\nabla^2 f_A(x)] = 0$, and so $\mathsf{Var}[f_A(x)] \leq 36C_3\sigma^6\|A\|_F^2$, which implies that[3]

$$M_6 - 6M_2 \otimes M_4 + 9M_2 \otimes M_2 \otimes M_2 - M_3 \otimes M_3 \preceq 36C_3\sigma^6 I, \tag{170}$$

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

In summary, we have $M_6(v) \preceq_{\text{sos}} (36C_3 + 12C_2')\sigma^6\|v\|_2^6$, 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 $k-1$ derivates all have zero mean. This always exists and is unique up to scaling by constants. For instance, when $k = 4$ we can take $f_v(x) = \langle x-\mu, v\rangle^4 - 6M_2(v)\langle x-\mu, v\rangle^2 - 4M_3(v)\langle x-\mu, v\rangle - M_4(v) + 6M_2(v)^2$. This appears somewhat clunky but is a special case of a combinatorial sum. For the general case, let $\mathcal{T}_k$ 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

$$f_{v,k}(x) = \sum_{(i_0,\ldots,i_r)\in\mathcal{T}_k} (-1)^r \binom{k}{i_0 \ \cdots \ i_r} \langle x-\mu, v\rangle^{i_0} M_{i_1}(v)M_{i_2}(v)\cdots M_{i_r}(v). \tag{171}$$

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 2.49 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.

[Lectures 10-11]

# 3 Resilience Beyond Mean Estimation

We have so far focused primarily on mean estimation, first considering information theoretic and then algorithmic issues. We now turn back to information theoretic issues with a focus on generalizing our results from mean estimation to other statistical problems.

Let us recall our general setup: for true (test) distribution $p^*$ and corrupted (train) distribution $\tilde{p}$, we observe samples $X_1, \ldots, X_n$ from $\tilde{p}$ (oblivious contamination, although we can also consider adaptive

[3] Actually this is not quite true because we only bound $\mathsf{Var}[f_A(x)]$ for symmetric tensors $A$. What is true is that this holds if we symmetrize the left-hand-side of (170), which involves averaging over all ways of splitting $M_2$ and $M_4$ over the 3 copies of $\mathbb{R}^d$ in $\mathbb{R}^{d\times d\times d}$.

contamination as in Section 2.6.2). We wish to estimate a parameter $\theta$ and do so via en estimator $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$. Our goal is to construct an estimator such that $L(p^*, \hat{\theta})$ is small according to a given loss function $L$. This was summarized in Figure 1 from Section 1.

As before, we will start by allowing our estimator $\hat{\theta}$ to directly access the population distribution $\tilde{p}$ rather than samples. Thus we wish to control the error $L(p^*, \hat{\theta}(\tilde{p}))$. Since this is hopeless without further assumptions, we assume that $D(p^*, \tilde{p}) \le \epsilon$ for some distance $D$, and that $p^*$ lies in some family $\mathcal{G}$.

For now we continue to take $D = \mathsf{TV}$ and focus on more general losses $L$, corresponding to tasks beyond mean estimation. Two key examples will be:

- **Second moment estimation** in spectral norm, which corresponds to the loss $L(p, S) = \|\mathbb{E}_p[XX^\top] - S\|$.
- **Linear regression**, which corresponds to the loss $L(p, \theta) = \mathbb{E}_{x,y\sim p}[(y - \theta^\top x)^2 - (y - \theta^*(p)^\top x)^2]$. Note that here $L$ measures the *excess predictive loss* so that $L(p, \theta^*(p)) = 0$.

As in the mean estimation case, we will define the modulus of continuity and the family of resilience distributions, and derive sufficient conditions for resilience.

**Modulus of continuity.** The modulus of continuity generalizes straightforwardly from the mean estimation case. We define

$$\mathfrak{m}(\mathcal{G}, 2\epsilon, L) = \sup_{p,q\in\mathcal{G}:\mathsf{TV}(p,q)\le 2\epsilon} L(p, \theta^*(q)). \tag{172}$$

As before, the modulus $\mathfrak{m}$ upper-bounds the minimax loss. Specifically, consider the projection estimator that outputs $\hat{\theta}(\tilde{p}) = \theta^*(q)$ for any $q \in \mathcal{G}$ with $\mathsf{TV}(\tilde{p}, q) \le \epsilon$. Then the error of $\hat{\theta}$ is at most $\mathfrak{m}$ because $\mathsf{TV}(q, p^*) \le 2\epsilon$ and $p^*, q \in \mathcal{G}$.

**Resilience.** Generalizing resilience requires more care. Recall that for mean estimation the set of $(\rho, \epsilon)$-resilient distributions was

$$\mathcal{G}^{\mathsf{TV}}_{\mathsf{mean}}(\rho, \epsilon) \overset{\text{def}}{=} \{p \mid \|\mathbb{E}_r[X] - \mathbb{E}_p[X]\| \le \rho \text{ for all } r \le \frac{p}{1-\epsilon}\}. \tag{173}$$

We saw in Section 2.4 that robust mean estimation is possible for the family $\mathcal{G}_{\mathsf{mean}}$ of resilient distributions; the two key ingredients were the existence of a midpoint distribution and the triangle inequality for $L(p, \theta^*(q)) = \|\mu_p - \mu_q\|$. We now extend the definition of resilience to arbitrary cost functions $L(p, \theta)$ that may not satisfy the triangle inequality. The general definition below imposes two conditions: (1) the parameter $\theta^*(p)$ should do well on all distributions $r \le \frac{p}{1-\epsilon}$, and (2) any parameter that does well on some $r \le \frac{p}{1-\epsilon}$ also does well on $p$. We measure performance on $r$ with a *bridge function* $B(r, \theta)$, which is often the same as the loss $L$ but need not be.

**Definition 3.1** ($\mathcal{G}^{\mathsf{TV}}(\rho_1, \rho_2, \epsilon)$)**.** Given an arbitrary loss function $L(p, \theta)$, we define $\mathcal{G}^{\mathsf{TV}}(\rho_1, \rho_2, \epsilon) = \mathcal{G}^{\mathsf{TV}}_{\downarrow}(\rho_1, \epsilon) \cap \mathcal{G}^{\mathsf{TV}}_{\uparrow}(\rho_1, \rho_2, \epsilon)$, where:

$$\mathcal{G}^{\mathsf{TV}}_{\downarrow}(\rho_1, \epsilon) \triangleq \{p \mid \sup_{r\le\frac{p}{1-\epsilon}} B(r, \theta^*(p)) \le \rho_1\}, \tag{174}$$

$$\mathcal{G}^{\mathsf{TV}}_{\uparrow}(\rho_1, \rho_2, \epsilon) \triangleq \{p \mid \text{ for all } \theta, r \le \frac{p}{1-\epsilon}, (B(r, \theta) \le \rho_1 \Rightarrow L(p, \theta) \le \rho_2)\}, \tag{175}$$

The function $B(p, \theta)$ is an arbitrary cost function that serves the purpose of bridging.

If we take $B(p, \theta) = L(p, \theta) = \|\mathbb{E}_p[X] - \mathbb{E}_\theta[X]\|, \rho_2 = 2\rho_1$, then this exactly reduces to the resilient set $\mathcal{G}^{\mathsf{TV}}_{\mathsf{mean}}(\rho_1, \epsilon)$ for mean estimation. To see the reduction, note that $\mathcal{G}^{\mathsf{TV}}_{\mathsf{mean}}$ is equivalent to $\mathcal{G}^{\mathsf{TV}}_{\downarrow}$ in Equation (174). Thus we only need to show that $\mathcal{G}^{\mathsf{TV}}_{\uparrow}$ is a subset of $\mathcal{G}^{\mathsf{TV}}_{\downarrow}$. By our choice of $B, L$ and $\rho_2$, the implication condition in $\mathcal{G}^{\mathsf{TV}}_{\uparrow}$ follows from the triangle inequality..

We will show that $\mathcal{G}^{\mathsf{TV}}$ is *not too big* by bounding its modulus of continuity, and that it is *not too small* by exhibiting reasonable sufficient conditions for resilience.

$$\begin{tikzcd}
p_1 \arrow[rr, "\mathsf{TV}(p_1, p_2) \leq \eta"] \arrow[dr, "r \leq \frac{p_1}{1-\eta}"'] & & p_2 \arrow[dl, "r \leq \frac{p_2}{1-\eta}"] \\
& r = \frac{\min(p_1, p_2)}{1-\mathsf{TV}(p_1, p_2)} &
\end{tikzcd}$$

$$p_1 \in \mathcal{G}_{\downarrow}^{\mathsf{TV}} \Longrightarrow B(r, \theta^*(p_1)) \leq \rho_1 \xRightarrow{p_2 \in \mathcal{G}_{\uparrow}^{\mathsf{TV}}} L(p_2, \theta^*(p_1)) \leq \rho_2$$

Figure 7: Midpoint distribution helps bridge the modulus for $\mathcal{G}^{\mathsf{TV}}$.

**Not too big: bounding $\mathfrak{m}$.** We show that the designed $\mathcal{G}^{\mathsf{TV}}(\rho_1, \rho_2, \epsilon)$ has small modulus of continuity (and thus population minimax limit) in the following theorem:

**Theorem 3.2.** *For $\mathcal{G}^{\mathsf{TV}}(\rho_1, \rho_2, \epsilon)$ in Definition 3.1, we have $\mathfrak{m}(\mathcal{G}^{\mathsf{TV}}(\rho_1, \rho_2, \epsilon), \epsilon) \leq \rho_2$.*

*Proof.* As illustrated in Figure 7, we still rely on the midpoint distribution $r$ to bridge the modulus. Consider any $p_1, p_2$ satisfying $\mathsf{TV}(p_1, p_2) \leq \epsilon$. Then there is a midpoint $r$ such that $r \leq \frac{p_1}{1-\epsilon}$ and $r \leq \frac{p_2}{1-\epsilon}$. From the fact that $p_1 \in \mathcal{G}^{\mathsf{TV}}(\rho_1, \rho_2, \epsilon) \subset \mathcal{G}_{\downarrow}^{\mathsf{TV}}(\rho_1, \epsilon)$, we have $B(r, \theta^*(p_1)) \leq \rho_1$. From this and the fact that $p_2 \in \mathcal{G}^{\mathsf{TV}}(\rho_1, \rho_2, \epsilon) \subset \mathcal{G}_{\uparrow}^{\mathsf{TV}}(\rho_1, \rho_2, \epsilon)$, we then have $L(p_2, \theta^*(p_1)) \leq \rho_2$. Since $p_1$ and $p_2$ are arbitrary, this bounds the modulus of continuity by $\rho_2$. □

**Not too small: concrete examples.** We next show that $\mathcal{G}^{\mathsf{TV}}$ yields sensible conditions for second moment estimation and linear regression. We start with second moment estimation:

**Proposition 3.3.** *Let $B(p, S) = L(p, S) = \|\mathbb{E}_p[XX^\top] - S\|$, and let $p$ be a distribution on $\mathbb{R}^d$ such that $p \in \mathcal{G}_{\mathsf{mom},k})(\sigma)$, i.e. $p^*$ has bounded $k$th moments. Then assuming $k > 2$, we have $p \in \mathcal{G}^{\mathsf{TV}}(\rho, 2\rho, \epsilon)$ for $\rho = \mathcal{O}(\sigma^2\epsilon^{1-2/k})$.*

This is essentially the same statement as for mean estimation, except with $\sigma^2\epsilon^{1-2/k}$ instead of $\sigma\epsilon^{1-1/k}$.

*Proof.* First we show that $p \in \mathcal{G}^{\downarrow}(\rho, \epsilon)$, for which we need to show that

$$\|\mathbb{E}_r[XX^\top] - \mathbb{E}_p[XX^\top]\| \leq \rho \text{ for all } r \leq \frac{p}{1-\epsilon}. \tag{176}$$

Letting $Y = XX^\top$, this asks that $Y$ is resilient in operator norm, which in turn asks that $\langle Y, Z\rangle$ is resilient for any $\|Z\|_* \leq 1$, where $\|\cdot\|_*$ is dual to the operator norm. Recalling that the operator norm is the maximum singular value, it turns out that $\|\cdot\|_*$ is the *nuclear norm*, or the sum of the singular values. Thus for $Z = U\Lambda V^\top$ we have $\|Z\|_* = \sum_i \Lambda_{ii}$. (Proving this duality requires some non-trivial but very useful matrix inequalities that we provide at the end of this section.)

Conveniently, the extreme points of the nuclear norm ball are exactly rank-one matrices of the form $\pm vv^\top$ where $\|v\|_2 = 1$. Thus we exactly need that $\langle v, X\rangle^2$ is resilience for all $v$. Fortunately we have that $\mathbb{E}[|\langle v, X\rangle^2 - \mathbb{E}[\langle v, X\rangle^2]|^{k/2}] \leq \mathbb{E}[|\langle v, X\rangle|^k] \leq \sigma^k$, so $p$ is $(\rho_1, \epsilon)$-resilient with $\rho_1 = \sigma^2\epsilon^{1-2/k}$, which gives that $p \in \mathcal{G}^{\downarrow}$.

Next we need to show that $p \in \mathcal{G}^{\uparrow}$. We want

$$\|\mathbb{E}_r[XX^\top] - S\| \leq \rho_1 \implies \|\mathbb{E}_p[XX^\top] - S\| \leq \rho_2 \text{ whenever } r \leq \frac{p}{1-\epsilon}, \tag{177}$$

but this is the same as $\rho_2 - \rho_1 \leq \|\mathbb{E}_r[XX^\top] - \mathbb{E}_p[XX^\top]\|$, and we already know that the right-hand-side is bounded by $\rho_1$, so we can take $\rho_2 = 2\rho_1$, which proves the claimed result. □

We move on to linear regression. In the proof for second moment estimation, we saw that $p \in \mathcal{G}^{\uparrow}$ was essentially implied by $p \in \mathcal{G}^{\downarrow}$. This was due to the symmetry of the second moment loss together with the triangle inequality for $\|\cdot\|$, two properties that we don't have in general. The proof for second moment estimation will require somewhat more different proofs for $\mathcal{G}^{\uparrow}$ and $\mathcal{G}^{\downarrow}$. For simplicity we state the result only for fourth moments:

**Proposition 3.4.** *For a distribution $p$ on $\mathbb{R}^d \times \mathbb{R}$, let $B(p,\theta) = L(p,\theta) = \mathbb{E}_p[(y - \langle \theta, x\rangle)^2 - (y - \langle \theta^*(p), x\rangle)^2]$. Let $Z = Y - \langle \theta^*(p), X\rangle$ and suppose that the following two conditions holds:*

$$\mathbb{E}_p[XZ^2X^\top] \preceq \sigma^2 \mathbb{E}[XX^\top], \tag{178}$$

$$\mathbb{E}_p[\langle X, v\rangle^4] \leq \kappa \mathbb{E}_p[\langle X, v\rangle^2]^2 \text{ for all } v. \tag{179}$$

*Then $p \in \mathcal{G}^{\mathsf{TV}}(\rho, 5\rho, \epsilon)$ for $\rho = 2\sigma^2\epsilon$ as long as $\epsilon(\kappa - 1) \leq \frac{1}{6}$ and $\epsilon \leq \frac{1}{8}$.*

Let us interpret the two conditions. First, as long as $X$ and $Z$ are independent (covariates are independent of noise), we have $\mathbb{E}_p[XZ^2X^\top] = \mathbb{E}[Z^2]\mathbb{E}[XX^\top]$, so in that case $\sigma^2$ is exactly a bound on the noise $Z$. Even when $X$ and $Z$ are not independent, the first condition holds when $Z$ has bounded 4th moment.

The second condition is a *hypercontractivity condition* stating that the fourth moments of $X$ should not be too large compared to the second moments. It is a bit unusual from the perspective of mean estimation, because it does not require $X$ to be well-concentrated, but only well-concentrated relative to its variance. For regression, this condition makes sense because $\kappa$ bounds how close the covariates are to being rank-deficient (the worst-case is roughly an $\epsilon$-mass at some arbitrary distance $t/\sqrt{\epsilon}$, which would have second moment $t^2$ and fourth moment $t^4/\epsilon$, so we roughly want $\kappa < 1/\epsilon$). We will show later that such a hypercontractivity condition is needed, i.e. simply assuming sub-Gaussianity (without making it relative to the variance) allows for distributions that are hard to robustly estimate due to the rank-deficiency issue.

*Proof.* First note that $L(p,\theta) = (\theta - \theta^*(p))^\top S_p(\theta - \theta^*(p))$, where $S_p = \mathbb{E}_p[XX^\top]$, and analogously for $L(r,\theta)$. At a high level our strategy will be to show that $\theta^*(r) \approx \theta^*(p)$ and $S_r \approx S_p$, and then use this to establish membership in $\mathcal{G}^{\downarrow}$ and $\mathcal{G}^{\uparrow}$.

We first use the hypercontractivity condition to show that $S_r \approx S_p$. We have

$$\begin{aligned}
\mathbb{E}_r[\langle v, X\rangle^2] &\geq \mathbb{E}_p[\langle v, X\rangle^2] - \frac{1}{1-\epsilon}\sqrt{\epsilon \mathsf{Var}_p[\langle v, X\rangle^2]} && (180)\\
&= \mathbb{E}_p[\langle v, X\rangle^2] - \frac{1}{1-\epsilon}\sqrt{\epsilon(\mathbb{E}_p[\langle v, X\rangle^4] - \mathbb{E}_p[\langle v, X\rangle^2]^2)} && (181)\\
&\geq \mathbb{E}_p[\langle v, X\rangle^2] - \frac{1}{1-\epsilon}\sqrt{\epsilon(\kappa-1)}\mathbb{E}_p[\langle v, X\rangle^2] && (182)\\
&= (1 - \frac{1}{1-\epsilon}\sqrt{\epsilon(\kappa-1)})\mathbb{E}_p[\langle v, X\rangle^2]. && (183)
\end{aligned}$$

Thus $S_r \succeq (1 - \frac{1}{1-\epsilon}\sqrt{\epsilon(\kappa-1)})S_p$, and similarly $S_r \preceq (1 + \frac{1}{1-\epsilon}\sqrt{\epsilon(\kappa-1)})S_p$. Assuming $\epsilon \leq \frac{1}{8}$ and $\epsilon(\kappa - 1) \leq \frac{1}{6}$, we have $\frac{1}{1-\epsilon}\sqrt{\epsilon(\kappa-1)} \leq \frac{8}{7}\sqrt{1/6} < \frac{1}{2}$, and so $\frac{1}{2}S_p \preceq S_r \preceq \frac{3}{2}S_p$.

We now turn to $\mathcal{G}^{\uparrow}$ and $\mathcal{G}^{\downarrow}$. A useful relation is $\theta^*(p) = S_p^{-1}\mathbb{E}_p[XY]$, and $\theta^*(r) - \theta^*(p) = S_r^{-1}\mathbb{E}_r[XZ]$. To prove that $p \in \mathcal{G}^{\downarrow}$ we need to show that $(\theta^*(r) - \theta^*(p))^\top S_r(\theta^*(r) - \theta^*(p))$ is small. We have

$$\begin{aligned}
(\theta^*(r) - \theta^*(p))^\top S_r(\theta^*(r) - \theta^*(p)) &\leq \frac{3}{2}(\theta^*(r) - \theta^*(p))^\top S_p(\theta^*(r) - \theta^*(p)) && (184)\\
&= \frac{3}{2}\mathbb{E}_r[XZ]^\top S_p^{-1}\mathbb{E}_r[XZ] = \frac{3}{2}\|\mathbb{E}_r[S_p^{-1/2}XZ] - \mathbb{E}_p[S_p^{-1/2}XZ]\|_2^2. && (185)
\end{aligned}$$

This final condition calls for $S_p^{-1/2}XZ$ to be resilient, and bounded variance of this distribution can be seen to exactly correspond to the condition $\mathbb{E}[XZ^2X^\top] \preceq \sigma^2\mathbb{E}[XX^\top]$. Thus we have resilience with $\rho = \frac{3\sigma^2\epsilon}{2(1-\epsilon)^2} \leq 2\sigma^2\epsilon$ (since $\epsilon < \frac{1}{8}$).

Now we turn to $\mathcal{G}^{\uparrow}$. We want that $(\theta - \theta^*(r))^\top S_r(\theta - \theta^*(r)) \leq \rho$ implies $(\theta - \theta^*(p))^\top S_p(\theta - \theta^*(p)) \leq 5\rho$.

By the triangle inequality we have

$$\sqrt{(\theta - \theta^*(p))^\top S_p(\theta - \theta^*(p))} \le \sqrt{(\theta - \theta^*(r))^\top S_p(\theta - \theta^*(r))} + \sqrt{(\theta^*(r) - \theta^*(p))^\top S_p(\theta^*(r) - \theta^*(p))} \quad (186)$$

$$\le \sqrt{2(\theta - \theta^*(r))^\top S_r(\theta - \theta^*(r))} + \sqrt{(4/3)\sigma^2\epsilon} \quad (187)$$

$$\le \sqrt{2\rho} + \sqrt{(4/3)\sigma^2\epsilon} = \sqrt{\rho}(\sqrt{2} + \sqrt{2/3}) \le \sqrt{5\rho}, \quad (188)$$

which completes the proof. □

**Lower bound.** TBD

**Proving that nuclear norm is dual to operator norm.** Here we establish a series of matrix inequalities that are useful more broadly, and use these to analyze the nuclear norm. The first allows us to reduce dot products of arbitrary matrices to symmetric PSD matrices:

**Proposition 3.5.** *For any (rectangular) matrices $A$, $B$ of equal dimensions, we have*

$$\langle A, B\rangle^2 \le \langle (A^\top A)^{1/2}, (B^\top B)^{1/2}\rangle\langle (AA^\top)^{1/2}, (BB^\top)^{1/2}\rangle. \quad (189)$$

In a sense, this is like a "matrix Cauchy-Schwarz".

*Proof.* We first observe that $\begin{bmatrix} (AA^\top)^{1/2} & A \\ A^\top & (A^\top A)^{1/2} \end{bmatrix} \succeq 0$. This is because, if $A = U\Lambda V^\top$ is the singular value decomposition, we can write the above matrix as $\begin{bmatrix} U\Lambda U^\top & U\Lambda V^\top \\ V\Lambda U^\top & V\Lambda V^\top \end{bmatrix}$, which is PSD because it can be factorized as $[U; V]\Lambda[U; V]^\top$. More generally this is true if we multiply $(AA^\top)^{1/2}$ by $\lambda$ and $(A^\top A)^{1/2}$ by $\frac{1}{\lambda}$. We therefore have

$$\left\langle \begin{bmatrix} \lambda(AA^\top)^{1/2} & A \\ A^\top & \frac{1}{\lambda}(A^\top A)^{1/2} \end{bmatrix}, \begin{bmatrix} \lambda(BB^\top)^{1/2} & -B \\ -B^\top & \frac{1}{\lambda}(B^\top B)^{1/2} \end{bmatrix} \right\rangle \ge 0, \quad (190)$$

since both terms in the inner product are PSD. This gives $\lambda^2\langle (AA^\top)^{1/2}, (BB^\top)^{1/2}\rangle + \frac{1}{\lambda^2}\langle (A^\top A)^{1/2}, (B^\top B)^{1/2}\rangle \ge 2\langle A, B\rangle$. Optimizing $\lambda$ yields the claimed result. □

Next we show:

**Theorem 3.6.** *If $A$ and $B$ are matrices of the same dimensions with (sorted) lists of singular values $\sigma_1, \dots, \sigma_n$ and $\tau_1, \dots, \tau_n$, then*

$$\langle A, B\rangle \le \sum_{i=1}^n \sigma_i\tau_i. \quad (191)$$

This says that the dot product between two matrices is bounded by the dot product between their sorted singular values.

*Proof.* By Proposition 3.5, it suffices to show this in the case that $A$ and $B$ are both PSD and $\sigma$, $\tau$ are the eigenvalues. Actually we will only need $A$ and $B$ to be symmetric (which implies that, oddly, the inequality can hold even if some of the $\sigma_i$ and $\tau_i$ are negative).

By taking similarity transforms we can assume without loss of generality that $A = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$ with $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n$. We thus wish to prove that $\sum_{i=1}^n \sigma_i B_{ii} \le \sum_{i=1}^n \sigma_i\tau_i$, where $\tau_i$ are the eigenvalues of $B$. We make use of the following lemma:

**Lemma 3.7.** *For all $1 \le k \le n$, we have $\sum_{i=1}^k B_{ii} \le \sum_{i=1}^k \tau_i$.*

*Proof.* Let $B_k$ be the $k \times k$ top-left submatrix of $B$. Then $\sum_{i=1}^k B_{ii} = \operatorname{tr}(B_k)$ is the sum of the eigenvalues of $B_k$. We will show that the $j$th largest eigenvalue of $B_k$ is smaller than the $j$th largest eigenvalue of $B$ (this is a special case of the *Cauchy interlacing theorem*). We prove this using the min-max formulation of eigenvalues: $\lambda_i(M) = \min_{W:\dim(W)=i-1} \max_{v \in W^\perp, \|v\|_2 \leq 1} v^\top M v$. Let $W^*$ be the $W$ that attains the min for $\lambda_j(B)$, and let $P_k$ denote projection onto the first $k$ coordinates. We have

$$\lambda_j(B_k) = \min_{W:\dim(W)=i-1} \max_{v \in W^\perp : \|v\|_2 \leq 1} v^\top B_k v \tag{192}$$

$$\leq \max_{v \in (W^*)^\perp : \|v\|_2 \leq 1} (P_k v)^\top B_k (P_k v) \tag{193}$$

$$\leq \max_{v \in (W^*)^\perp : \|v\|_2 \leq 1} v^\top B v = \lambda_j(B), \tag{194}$$

which proves the lemma. □

Now with the lemma in hand we observe that, if we let $\sigma_{n+1} = 0$ for convenience, we have

$$\sum_{i=1}^n \sigma_i B_{ii} = \sum_{i=1}^n (\sigma_i - \sigma_{i+1})(B_{11} + \cdots + B_{ii}) \tag{195}$$

$$\leq \sum_{i=1}^n (\sigma_i - \sigma_{i+1})(\tau_1 + \cdots + \tau_i) \tag{196}$$

$$= \sum_{i=1}^n \sigma_i \tau_i, \tag{197}$$

which yields the desired result. In the above algebra we have used *Abel summation*, which is the discrete version of integration by parts. □

Now that we have Theorem 3.6 in hand, we can easily analyze the operator and nuclear norms. Letting $\vec{\sigma}(A)$ denote the vector of non-decreasing singular values of $A$, we have

$$\langle Y, Z \rangle \leq \langle \vec{\sigma}(Y), \vec{\sigma}(Z) \rangle \leq \|\vec{\sigma}(Y)\|_\infty \|\vec{\sigma}(Z)\|_1. \tag{198}$$

This shows that the dual of the operator norm is at most the nuclear norm, since $\|\vec{\sigma}(Z)\|_1$ is the nuclear norm of $Z$. But we can achieve equality when $Y = U \Lambda V^\top$ by taking $Z = u_1 v_1^\top$ (then $\|Z\|_* = 1$ while $\langle Y, Z \rangle = \Lambda_{11} = \|Y\|$). So operator and nuclear norm are indeed dual to each other.

[Lecture 12]

## 3.1 Efficient Algorithm for Robust Regression

We now turn to the question of efficient algorithms, focusing on linear regression (we will address finite-sample issues later). Recall that information-theoretically, we found that two conditions are sufficient to imply resilience:

- *Hypercontractivity:* For all $v$, $\mathbb{E}_{x \sim p}[\langle x, v \rangle^4] \leq \kappa \mathbb{E}_{x \sim p}[\langle x, v \rangle^2]^2$.
- *Bounded noise:* $\mathbb{E}_{x \sim p}[x z^2 x^\top] \preceq \sigma^2 \mathbb{E}_{x \sim p}[x x^\top]$.

As for mean estimation under bounded covariance, our strategy will be to write down a non-convex optimization problem that tries to find small $\kappa$ and $\sigma$, then show that this problem can be approximately solved. Specifically, let

$$F_1(q) = \sup_v \frac{\mathbb{E}_{x \sim q}[\langle x, v \rangle^4]}{\mathbb{E}_{x \sim q}[\langle x, v \rangle^2]^2}, \text{ and} \tag{199}$$

$$F_2(q) = \sup_v \frac{\mathbb{E}_{x \sim q}[\langle x, v \rangle^2 (y - \langle \theta(q), x \rangle)^2]}{\mathbb{E}_{x \sim q}[\langle x, v \rangle^2]}. \tag{200}$$

Then we seek to find a $q$ such that $F_1(q) \le \kappa$, $F_2(q) \le \sigma^2$, and $q \in \Delta_{n,\epsilon}$, where $\Delta_{n,\epsilon}$ is the set of $\epsilon$-deletions of $p$.

However, there are a couple wrinkles. While with mean estimation we could minimize the objective with gradient descent, in this case we will need to use *quasigradient* descent–following a direction that is not the gradient, but that we can show makes progress towards the optimum. The rough reason for this is that, since $p$ appears on both the left- and right-hand sides of the inequalities above, the gradients become quite messy, e.g. $\nabla F_1(q)$ has a mix of positive and negative terms:

$$\nabla F_1(q)_i = \frac{\langle x_i, v\rangle^4}{\mathbb{E}_{x\sim q}[\langle x, v\rangle^2]^2} - 2\frac{\mathbb{E}_q[\langle x, v\rangle^4]\langle x_i, v\rangle^2}{\mathbb{E}_q[\langle x, v\rangle^2]^3}, \tag{201}$$

and it isn't clear that following them will not land us at bad local minima. To address this, we instead construct a simpler *quasigradient* for $F_1$ and $F_2$:

$$g_1(x_i; q) = \langle x_i, v\rangle^4, \qquad \text{where } v = \operatorname*{arg\,max}_{\|v\|_2=1} \frac{\mathbb{E}_q[\langle x, v\rangle^4]}{\mathbb{E}_q[\langle x, v\rangle^2]^2}, \tag{202}$$

$$g_2(x_i; q) = \langle x_i, v\rangle^2(y_i - \langle\theta^*(q), x_i\rangle)^2, \qquad \text{where } v = \operatorname*{arg\,max}_{\|v\|_2=1} \frac{\mathbb{E}_q[\langle x, v\rangle^2(y - \langle\theta^*(q), x\rangle)^2]}{\mathbb{E}_q[\langle x, v\rangle^2]}. \tag{203}$$

We will then follow $g_1$ until $F_1$ is small, and then follow $g_2$ until $F_2$ is small.

The other wrinkle is that computationally, the hypercontractivity condition is difficult to certify, because it involves maximizing $\frac{\mathbb{E}_p[\langle x,v\rangle^4]}{\mathbb{E}_p[\langle x,v\rangle^2]^2}$, which is no longer a straightforward eigenvalue problem as in the mean estimation case. We've already seen this sort of difficulty before–for norms beyond the $\ell_2$-norm, we had to use SDP relaxations and Grothendieck's inequality in order to get a constant factor relaxation. Here, there is also an SDP relaxation called the *sum-of-squares* relaxation, but it doesn't always give a constant factor relaxation. We'll mostly ignore this issue and assume that we can find the maximizing $v$ for hypercontractivity.

We are now ready to define our efficient algorithm for linear regression, Algorithm 4. It is closely analogous to the algorithm for mean estimation (Algorithm 2), but specifies the gradient steps more explicitly.

**Algorithm 4** `QuasigradientDescentLinReg`

```
Input: $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$.
Initialize $q \in \Delta_{n,\epsilon}$ arbitrarily.
while $F_1(q) \ge 2\kappa$ or $F_2(q) \ge 4\sigma^2$ do
    if $F_1(q) \ge 2\kappa$ then
        Find the unit vector $v$ that maximizes $\mathbb{E}_q[\langle x, v\rangle^4]/\mathbb{E}_q[\langle x, v\rangle^2]^2$.
        Take a projected gradient step in the direction $(g_1)_i = \langle x_i, v\rangle^4$.
    else
        Compute the empirical least squares regressor: $\theta^*(q) = (\sum_{i=1}^n q_i x_i x_i^\top)^{-1}(\sum_{i=1}^n q_i x_i y_i)$.
        Find the unit vector $v$ that maximizes $\mathbb{E}_q[\langle x, v\rangle^2(y - \langle\theta^*(q)x\rangle)^2]/\mathbb{E}_q[\langle x, v\rangle^2]$.
        Take a projected gradient step in the direction $(g_2)_i = \langle x_i, v\rangle^2(y_i - \langle\theta^*(q), x_i\rangle)^2$.
    end if
end while
Output $\theta^*(q)$.
```

Analyzing Algorithm 4 enjoys the following loss bound:

**Proposition 3.8.** *Suppose that a set $S$ of $(1-\epsilon)n$ of the $x_i$ satisfy:*

$$\mathbb{E}_{p_S}[\langle x, v\rangle^4] \le \kappa\mathbb{E}_{p_S}[\langle x, v\rangle^2]^2 \; \forall v \in \mathbb{R}^d, \text{ and } \mathbb{E}_{p_S}[(y - \langle\theta^*(p_S), x\rangle)^2 xx^\top] \preceq \sigma^2\mathbb{E}_{p_S}[xx^\top]. \tag{204}$$

*Then assuming $\kappa\epsilon \le \frac{1}{80}$, Algorithm 4 terminates and its output has excess loss $L(p_S, \theta^*(q)) \le 40\sigma^2\epsilon$.*

To prove Proposition 3.8, we need a few ideas. The first is a result from optimization justifying the use of the quasigradients:

**Lemma 3.9** (Informal)**.** *Asymptotically, the iterates of Algorithm 4 (or any other low-regret algorithm) satisfy the conditions*

$$\mathbb{E}_{X\sim q}[g_j(X;q)] \le \mathbb{E}_{X\sim p_S}[g_j(X;q)] \text{ for } j=1,2. \tag{205}$$

We will establish this more formally later, and for now assume that (205) holds. Then, asuming this, we will show that $q$ is both hypercontractive and has bounded noise with constants $\kappa'$, $\sigma'$ that are only a constant factor worse than $\kappa$ and $\sigma$.

First, we will show this for hypercontractivity:

**Lemma 3.10.** *Suppose that $\mathbb{E}_{X\sim q}[g_1(X;q)] \le \mathbb{E}_{X\sim p_S}[g_1(X;q)]$ and that $\kappa\epsilon \le \frac{1}{80}$. Then $q$ is hypercontractive with parameter $\kappa' = 1.5\kappa$.*

*Proof.* Take the maximizing $v$ such that $g_1(x_i;q) = \langle x_i, v\rangle^4$. To establish hypercontractivity, we want to show that $\mathbb{E}_q[\langle x,v\rangle^4]$ is small while $\mathbb{E}_q[\langle x,v\rangle^2]^2$ is large. Note the quasigradient condition gives us that $\mathbb{E}_q[\langle x,v\rangle^4] \le \mathbb{E}_{p_S}[\langle x,v\rangle^4]$; so we will mainly focus on showing that $\mathbb{E}_q[\langle x,v\rangle^2]$ is large, and in fact not much smaller than $\mathbb{E}_{p_S}[\langle x,v\rangle^2]$. Specifically, by the fact that $\mathsf{TV}(q,p_S) \le \frac{\epsilon}{1-\epsilon}$, together with resilience applied to $\langle x,v\rangle^2$, we have

$$|\mathbb{E}_q[\langle x,v\rangle^2] - \mathbb{E}_{p_S}[\langle x,v\rangle^2]| \le \Big(\frac{\epsilon}{(1-2\epsilon)^2}(\mathbb{E}_q[\langle x,v\rangle^4] + \mathbb{E}_{p_S}[\langle x,v\rangle^4])\Big)^{\frac{1}{2}} \tag{206}$$

$$\overset{(i)}{\le} \Big(\frac{2\epsilon}{(1-2\epsilon)^2}\mathbb{E}_{p_S}[\langle x,v\rangle^4]\Big)^{\frac{1}{2}} \tag{207}$$

$$\overset{(ii)}{\le} \Big(\frac{2\kappa\epsilon}{(1-2\epsilon)^2}\Big)^{\frac{1}{2}}\mathbb{E}_{p_S}[\langle x,v\rangle^2], \tag{208}$$

where (i) uses the quasigradient condition and (ii) uses hypercontractivity for $p_S$. Now assuming that $\kappa\epsilon \le \frac{1}{80}$ (and hence also $\epsilon \le \frac{1}{80}$), the coefficient on the right-hand-side is at most $\sqrt{(1/40)/(1-1/40)^2} < \frac{1}{6}$. Consequently $|\mathbb{E}_q[\langle x,v\rangle^2] - \mathbb{E}_{p_S}[\langle x,v\rangle]^2| \le \frac{1}{6}\mathbb{E}_{p_S}[\langle x,v\rangle^2]$, and re-arranging then yields

$$\mathbb{E}_q[\langle x,v\rangle^2] \ge \frac{5}{6}\mathbb{E}_{p_S}[\langle x,v\rangle^2]. \tag{209}$$

But we already have $\mathbb{E}_q[\langle x,v\rangle^4] \le \mathbb{E}_{p_S}[\langle x,v\rangle^2]$, and so the ratio $\mathbb{E}_q[\langle x,v\rangle^2]/\mathbb{E}_q[\langle x,v\rangle^2]^2$ is at most $(6/5)^2$ the same ratio under $p_S$, and in particular at most $(6/5)^2\kappa \le 1.5\kappa$. □

Next, we will show this for bounded noise assuming that hypercontractivity holds:

**Lemma 3.11.** *Suppose that $F_1(q) \le 2\kappa$ and that $\mathbb{E}_{X\sim q}[g_2(X;q)] \le \mathbb{E}_{X\sim p_S}[g_2(X;q)]$, and that $\kappa\epsilon \le \frac{1}{80}$. Then $q$ has bounded noise with parameter $(\sigma')^2 = 4\sigma^2$, and furthermore satisfies $L(p_S, \theta^*(q)) \le 40\sigma^2\epsilon$.*

*Proof.* Again take the maximizing $v$ such that $g_2(x_i;q) = \langle x_i, v\rangle^2(y_i - \langle\theta^*(q), x_i\rangle)^2$. We want to show that $q$ has bounded noise, or in other words that $\mathbb{E}_q[g_2(x;q)]$ is small relative to $\mathbb{E}_q[\langle x,v\rangle^2]$. By the quasigradient assumption, we have

$$\mathbb{E}_q[g_2(x;q)] + \mathbb{E}_q[\langle x,v\rangle^2(y - \langle\theta^*(q),x\rangle)^2] \tag{210}$$

$$\le \mathbb{E}_{p_S}[\langle x,v\rangle^2(y - \langle\theta^*(q),x\rangle)^2]. \tag{211}$$

Intuitively, we want to use the bounded noise condition for $p_S$ to upper-bound the right-hand-side. The problem is that the term inside the expectation contains $\theta^*(q)$, rather than $\theta^*(p_S)$. But we can handle this using the AM-RMS inequality. Specifically, we have

$$\mathbb{E}_{p_S}[\langle x,v\rangle^2(y - \langle\theta^*(q),x\rangle)^2] \le 2\big(\underbrace{\mathbb{E}_{p_S}[\langle x,v\rangle^2(y - \langle\theta^*(p_S),x\rangle)^2]}_{(a)} + \underbrace{\mathbb{E}_{p_S}[\langle x,v\rangle^2\langle\theta^*(q) - \theta^*(p_S),x\rangle^2]}_{(b)}\big). \tag{212}$$

We will bound (a) and (b) in turn. To bound (a) note that by the bounded noise condition we simply have $(a) \le \sigma^2\mathbb{E}_{p_S}[\langle x,v\rangle^2]$.

To bound (b), let $R = \mathbb{E}_{p_S}[\langle \theta^*(q) - \theta^*(p_S), x\rangle^2] = L(p_S, \theta^*(q))$ be the excess loss of $\theta^*(q)$ under $p_S$. We will upper-bound (b) in terms of $R$, and then apply resilience to get a bound for $R$ in terms of itself. Solving the resulting inequality will provide an absolute bound on $R$ and hence also on (b).

More specifically, we have

$$\begin{aligned}
\mathbb{E}_{p_S}[\langle x, v\rangle^2 \langle \theta^*(q) - \theta^*(p_S), x\rangle^2] &\overset{(i)}{\leq} \Big(\mathbb{E}_{p_S}[\langle x, v\rangle^4\Big)^{1/4} \Big(\mathbb{E}_{p_S}[\langle \theta^*(q) - \theta^*(p_S), x\rangle^4]\Big)^{1/2} && (213) \\
&\overset{(ii)}{\leq} \kappa\Big(\mathbb{E}_{p_S}[\langle x, v\rangle^2]\Big)\Big(\mathbb{E}_{p_S}[\langle \theta^*(q) - \theta^*(p_S), x\rangle^2\Big) && (214) \\
&= \kappa R \mathbb{E}_{p_S}[\langle x, v\rangle^2]. && (215)
\end{aligned}$$

Here (i) is Cauchy-Schwarz and (ii) invokes hypercontractivity of $p_S$. Combining the bounds on (a) and (b) and plugging back in to (212), we obtain

$$\mathbb{E}_{p_S}[\langle x, v\rangle^2 (y - \langle \theta^*(q), x\rangle)^2] \leq 2(\sigma^2 + \kappa R)\mathbb{E}_{p_S}[\langle x, v\rangle^2]. \tag{216}$$

Remember that we would like a bound such as the above, but with expectations taken with respect to $q$ instead of $p_S$. For the left-hand-side, we can directly move to $\mathbb{E}_q[\cdot]$ using the quasigradient assumption. For the right-hand-side, since $F_1(q) \leq 2\kappa$, the same argument as in (206)-(209) yields (with modified constants) that $\mathbb{E}_q[\langle x, v\rangle^2] \geq \frac{4}{5}\mathbb{E}_{p_S}[\langle x, v\rangle^2]$. Applying both of these, we have that

$$\mathbb{E}_q[\langle x, v\rangle^2 (y - \langle \theta^*(q), x\rangle)^2] \leq 2.5(\sigma^2 + \kappa R)\mathbb{E}_q[\langle x, v\rangle^2]. \tag{217}$$

This establishes bounded noise with parameter $(\sigma')^2 = 2.5(\sigma^2 + \kappa R)$. By assumption, we also have hypercontractivity with parameter $\kappa' = 2\kappa$. We are not yet done, because we do not know $R$. But recall that $R$ is the excess loss, and so by the resilience conditions for linear regression (Proposition 3.4) we have

$$R \leq 5\rho(\kappa', \sigma') \leq 10(\sigma')^2\epsilon = 25(\sigma^2 + \kappa R)\epsilon, \tag{218}$$

as long as $\epsilon \leq \frac{1}{8}$ and $\kappa'\epsilon = 2\kappa\epsilon \leq \frac{1}{6}$. Re-arranging, we have

$$R \leq \frac{25\sigma^2\epsilon}{1 - 25\kappa\epsilon} \leq 40\sigma^2\epsilon, \tag{219}$$

since $\kappa\epsilon \leq \frac{1}{80}$. Plugging back into $\sigma'$, we also have $(\sigma')^2 \leq 2.5\sigma^2(1 + 40\kappa\epsilon) \leq 4\sigma^2$, as claimed. □

**Quasigradient bounds via low-regret algorithms.** Combining Lemmas 3.9, 3.10, and 3.11 together yields Proposition 3.8. However, we still need to formalize Lemma 3.9, showing that we can drive the quasigradients to be small. We can do so with the idea of low-regret *online optimization algorithms.*

An online optimization algorithm is one that takes a sequence of losses $\ell_1(\cdot), \ell_2(\cdot)$ rather than a fixed loss $\ell$. In traditional optimization, we have a single $\ell$ with parameter $w$, and seek to produce iterates $w_1, w_2, \ldots$ such that $\ell(w_T) - \ell(w^*) \to 0$ as $T \to \infty$. In online optimization algorithms, we instead consider the regret, defined as

$$\mathrm{Regret}_T = \max_w \frac{1}{T}\sum_{t=1}^{T} \ell_t(w_t) - \ell_t(w). \tag{220}$$

This is the average excess loss compared to the best fixed $w$, picked in hindsight. We then seek to produce iterates $w_t$ such that $\mathrm{Regret}_T \to 0$ as $T \to \infty$. Note that when $\ell_t = \ell$ is fixed for all $t$, this is exactly the same as traditional optimization.

Remarkably, for most "nice" loss functions $\ell_t$, it is possible to ensure that $\mathrm{Regret}_T \to 0$; in fact, projected gradient descent with an appropriate step size will achieve this.

How does this relate to quasigradients? For any quasigradient $g(X; q)$, define the loss function $\ell_t(q) = \mathbb{E}_{x \sim q}[g(x; q_t)]$. Even though this loss depends on $q_t$, the regret is well-defined:

$$\mathrm{Regret}_T = \max_{q'} \frac{1}{T}\mathbb{E}_{x \sim q_t}[g(x; q_t)] - \mathbb{E}_{x \sim q'}[g(x; q_t)] \geq \frac{1}{T}\mathbb{E}_{x \sim q_t}[g(x; q_t)] - \mathbb{E}_{x \sim p_S}[g(x; q_t)]. \tag{221}$$

In particular, as long as $\mathrm{Regret}_T \to 0$, we asymptotically have that $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x\sim q_t}[g(x;q_t)] \leq \sum_{t=1}^{T} \mathbb{E}_{x\sim p_S}[g(x;q_t)]$, so eventually the quasigradient bound $\mathbb{E}_{x\sim q}[g(x;q)] \leq \mathbb{E}_{x\sim p_S}[g(x;q)]$ must (almost) hold for one of the $q_t$.

This is enough to show that, for any fixed quasigradient, a statement such as Lemma 3.9 holds[4]. However, we want *both* $g_1$ and $g_2$ to be small simultaneously.

There are two ways to handle this. The first, crude way is to first use $g_1$ until we have hypercontractivity, then take the resulting $q$ as our new $\tilde{p}$ (so that we restrict to $\epsilon$-deletions of $q$) and running gradient descent with $g_2$ until we have bounded noise. This uses the fact that $\epsilon$-deletions of hypercontractive distributions are still hypercontractive, but yields worse constants (since we need $2\epsilon$-deletions instead of $\epsilon$-deletions).

A slicker approach is to alternate between $g_1$ and $g_2$ as in Algorithm 4. Note that we then still (asymptotically) have

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x\sim q_t}[g_{j_t}(x;q_t)] \leq \sum_{t=1}^{T} \mathbb{E}_{x\sim p_S}[g_{j_t}(x;q_t)], \tag{222}$$

where $j_t \in \{1,2\}$ denotes the choice of quasigradient at iteration $t$ of Algorithm 4.

Next note that asymptotically, only a vanishingly small fraction of $j_t$ can equal 1, since we only take quasigradient steps in $g_1$ when $F_1(q) \geq 2\kappa$, and in these cases $\mathbb{E}_{q_t}[g_1(x;q_t)]$ is quite a bit larger than $\mathbb{E}_{p_S}[g_1(x;q_t)]$, since if they were equal we would have $F_1(q) \leq 1.5\kappa$. Therefore, eventually almost all of the quasigradient steps are with respect to $g_2$, and so low regret of the entire sum implies low regret of $g_2$. We therefore both have $F_1(q_t) \leq 2\kappa$ and the quasigradient condition for $g_2$:

**Lemma 3.12** (Formal version of Lemma 3.9)**.** *Suppose that $|g_1(x_i,q)| \leq B$ and $|g_2(x_i,q)| \leq B$ for all $i$, where $B$ is at most polynomially-large in the problem parameters. Then for any $\delta$, within polynomially many steps Algorithm 4 generates an iterate $q_t$ such that $F_1(q_t) \leq 2\kappa$ and $\mathbb{E}_{q_t}[g_2(x,q_t)] \leq \mathbb{E}_{p_S}[g_2(x,q_t)] + \delta$.*

Combining Lemma 3.9 with Lemma 3.11 then yields the desired Proposition 3.8.

[Lectures 13-14]

# 4 Resilience Beyond TV Distance

We now turn our attention to distances other than the distance $D = \mathsf{TV}$ that we have considered so far. The family of distances we will consider are called *Wasserstein distances.* Given a cost function $c(x,y)$ (which is usually assumed to be a metric), we define the distance $W_c(p,q)$ between two distributions $p$ and $q$ as

$$W_c(p,q) = \inf_{\pi} \mathbb{E}_{x,y\sim\pi}[c(x,y)] \tag{223}$$

$$\text{subject to } \int \pi(x,y)dy = p(x), \int \pi(x,y)dx = q(y). \tag{224}$$

This definition is a bit abstruse so let us unpack it. The decision variable $\pi$ is called a *coupling* between $p$ and $q$, and can be thought of as a way of matching points in $p$ with points in $q$ ($\pi(x,y)$ is the amount of mass in $p(x)$ that is matched to $q(y)$). The Wasserstein distance is then the minimum cost coupling (i.e., minimum cost matching) between $p$ and $q$. Some special cases include:

- $c(x,y) = \mathbb{I}[x \neq y]$. Then $W_c$ is the total variation distance, with the optimal coupling being $\pi(x,x) = \min(p(x), q(x))$ (the off-diagonal $\pi(x,y)$ can be arbitrary as long as the total mass adds up correctly).
- $c(x,y) = \|x-y\|_2$. Then $W_c$ is the *earth-mover distance*—the average amount that we need to move points around to "move" $p$ to $q$.
- $c(x,y) = \|x-y\|_0$. Then $W_c$ is the average number of coordinates we need to change to move $p$ to $q$.
- $c(x,y) = \|x-y\|_2^{\alpha}$, for $\alpha \in [0,1]$. This is still a metric and interpolates between $\mathsf{TV}$ and earthmover distance.

[4] We have to be a bit careful because outliers could make $g(x;q)$ arbitrarily large, which violates the assumptions needed to achieve low regret. This can be addressed with a pre-filtering step that removes data points that are obviously too large to be inliers, but we will not worry about this here.

There are a couple key properties of Wasserstein distance we will want to use. The first is that $W_c$ is a metric if $c$ is:

**Proposition 4.1.** *Suppose that $c$ is a metric. Then $W_c$ is also a metric.*

*Proof.* TBD □

The second, called *Kantorovich-Rubinstein duality*, provides an alternate definition of $W_c$ distance in terms of functions that are Lipschitz under $c$, meaning that $|f(x) - f(y)| \leq c(x, y)$.

**Theorem 4.2** (Kantorovich-Rubinstein)**.** *Call a function $f$* Lipschitz in $c$ *if $|f(x) - f(y)| \leq c(x, y)$ for all $x, y$, and let $\mathcal{L}(c)$ denote the space of such functions. If $c$ is a metric, then we have*

$$W_c(p, q) = \sup_{f \in \mathcal{L}(c)} \mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{x \sim q}[f(x)]. \tag{225}$$

As a special case, take $c(x, y) = \mathbb{I}[x \neq y]$ (corresponding to $\mathsf{TV}$ distance). Then $f \in \mathcal{L}(c)$ if and only if $|f(x) - f(y)| \leq 1$ for all $x \neq y$. By translating $f$, we can equivalently take the supremum over all $f$ mapping to $[0, 1]$. This says that

$$\mathsf{TV}(p, q) = \sup_{f : \mathcal{X} \to [0,1]} \mathbb{E}_p[f(x)] - \mathbb{E}_q[f(x)], \tag{226}$$

which recovers the definition of $\mathsf{TV}$ in terms of the maximum difference in probability of any event $E$.

As another special case, take $c(x, y) = \|x - y\|_2$. Then the supremum is over all 1-Lipschitz functions (in the usual sense).

In the next section, we will see how to generalize the definition of resilience to any Wasserstein distance.

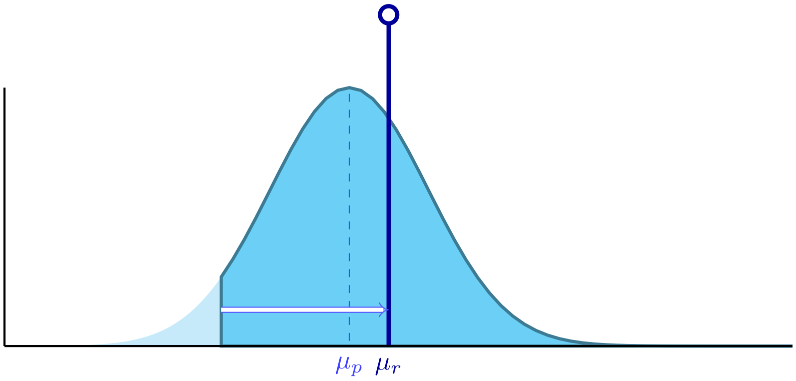
## 4.1 Resilience for Wasserstein distances

We show how to extend the idea of resilience to Wasserstein distances $W_c$. Recall that for $\mathsf{TV}$ distance, we showed that resilient sets have bounded modulus $\mathfrak{m}$; this crucially relied on the midpoint property that any $p_1$, $p_2$ have a midpoint $r$ obtained via *deletions* of $p_1$ or $p_2$. In other words, we used the fact that any $\mathsf{TV}$ perturbation can be decomposed into a "friendly" operation (deletion) and its opposite (addition). We think of deletion as friendlier than addition, as the latter can move the mean arbitrarily far by adding probability mass at infinity.

To extend this to other Wasserstein distances, we need to identify a similar way of decomposing a Wasserstein perturbation into a friendly perturbation and its inverse. Unfortunately, deletion is closely tied to the $\mathsf{TV}$ distance in particular. To get around this, we use the following re-interpretation: *Deletion is equivalent to movement towards the mean under* $\mathsf{TV}$. More precisely:

> $\hat{\mu}$ is a possible mean of an $\epsilon$-deletion of $p$ if and only if some $r$ with mean $\hat{\mu}$ can be obtained from $p$ by moving points *towards* $\hat{\mu}$ with $\mathsf{TV}$ distance at most $\epsilon$.

This is more easily seen in the following diagram:



Here we can equivalently either delete the left tail of $p$ or shift all of its mass to $\mu_r$; both yield a modified distribution with the same mean $\mu_r$. Thus we can more generally say that a perturbation is friendly if it only moves probability mass towards the mean. This motivates the following definition:

**Definition 4.3** (Friendly perturbation)**.** For a distribution $p$ over $\mathcal{X}$, fix a function $f : \mathcal{X} \to \mathbb{R}$. A distribution $r$ is an $\epsilon$-friendly perturbation of $p$ for $f$ under $W_c$ if there is a coupling $\pi$ between $X \sim p$ and $Y \sim r$ such that:

- The cost $(\mathbb{E}_\pi[c(X, Y)])$ is at most $\epsilon$.
- All points move towards the mean of $r$: $f(Y)$ is between $f(X)$ and $\mathbb{E}_r[f(Y)]$ almost surely.

Note that friendliness is defined only in terms of one-dimensional functions $f : \mathcal{X} \to \mathbb{R}$; we will see how to handle higher-dimensional objects later. Intuitively, a friendly perturbation is a distribution $r$ for which there exists a coupling that 'squeezes' $p$ to $\mu_r$.

The key property of deletion in the $\mathsf{TV}$ case was the existence of a *midpoint*: for any two distributions that are within $\epsilon$ in $\mathsf{TV}$, one can find another distribution that is an $\epsilon$-deletion of both distributions. We would like to show the analogous result for $W_c$–i.e. that if $W_c(p, q) \leq \epsilon$ then there exists an $r$ that is an $\epsilon$-friendly perturbation of *both* $p$ and $q$ for the function $f$.

The intuitive reason this is true is that any coupling between two one-dimensional distributions can be separated into two stages: in one stage all the mass only moves towards some point, in the other stage all the mass moves away from that point. This is illustrated in Figure 8.

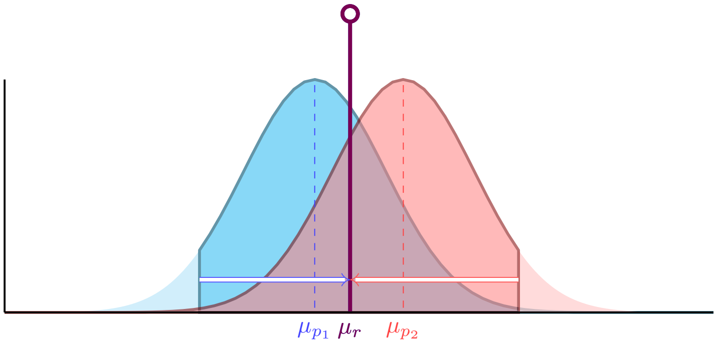


Figure 8: Illustration of midpoint lemma. For any distributions $p_1, p_2$ that are close under $W_c$, the coupling between $p_1$ and $p_2$ can be split into couplings $\pi_{p_1,r}$, $\pi_{p_2,r}$ such that $p_1, p_2$ only move towards $\mu_r$ under the couplings. We do this by "stopping" the movement from $p_1$ to $p_2$ at $\mu_r$.

To formalize this intuitive argument, we need a mild topological property:

**Assumption 4.4** (Intermediate value property)**.** *For any $x$ and $y$ and any $u$ with $f(x) < u < f(y)$, there is some $z$ satisfying $f(z) = u$ and $\max(c(x, z), c(z, y)) \leq c(x, y)$.*

This holds for any $f$ if $c = \mathbb{I}[x \neq y]$ ($\mathsf{TV}$ distance), and for any continuous $f$ if $c$ is a path metric (a metric with "nice" paths between points, which includes the $\ell_2$-distance). Under this assumption we can prove the desired midpoint lemma:

**Lemma 4.5** (Midpoint lemma for $W_c$)**.** *Suppose Assumption 4.4 holds. Then for any $p_1$ and $p_2$ such that $W_c(p_1, p_2) \leq \epsilon$ and any $f$, there exists a distribution $r$ that is an $\epsilon$-friendly perturbation of both $p_1$ and $p_2$ with respect to $f$.*

*Proof.* Given any two points $x$ and $y$, without loss of generality we assume $f(x) \leq f(y)$. Define

$$s_{xy}(u) = \begin{cases} \min(f(x), f(y)), & u \leq \min(f(x), f(y)) \\ u, & u \in [f(x), f(y)] \\ \max(f(x), f(y)), & u \geq \max(f(x), f(y)). \end{cases} \tag{227}$$

If we imagine $u$ increasing from $-\infty$ to $+\infty$, we can think of $s_{xy}$ as a "slider" that tries to be as close to $u$ as possible while remaining between $f(x)$ and $f(y)$.

By Assumption 4.4, there must exist some point $z$ such that $\max(c(x,z), c(z,y)) \leq c(x,y)$ and $f(z) = s_{xy}(u)$. Call this point $z_{xy}(u)$.

Given a coupling $\pi(x,y)$ from $p_1$ to $p_2$, if we map $y$ to $z_{xy}(u)$, we obtain a coupling $\pi_1(x,z)$ to some distribution $r(u)$, which by construction satisfies the squeezing property, except that it squeezes towards $u$ rather than towards the mean $\mu(u) = \mathbb{E}_{X\sim r(u)}[f(X)]$. However, note that $u - \mu(u)$ is a continuous, monotonically non-decreasing function (since $u - s_{xy}(u)$ is non-decreasing) that ranges from $-\infty$ to $+\infty$. It follows that there is a $u^*$ with $\mu(u^*) = u^*$. Then the couplings to $r(u^*)$ squeeze towards its mean $\mu(u^*)$.

Moreover, $\mathbb{E}_{(X,Z)\sim\pi_1}[c(X,Z)] \leq \mathbb{E}_{(X,Y)\sim\pi}[c(X,Y)] = W_c(p_1,p_2)$. The coupling $\pi_1$ therefore also has small enough cost, and so is a friendly perturbation. Similarly, the coupling $\pi_2$ mapping $y$ to $z_{xy}(u^*)$ satisfies the squeezing property and has small enough cost by the same argument. □

**Defining resilience: warm-up.** With Lemma 4.5 in hand, we generalize resilience to Wasserstein distances by saying that a distribution is resilient if $\mathbb{E}_r[f(X)]$ is close to $\mathbb{E}_p[f(X)]$ for every $\eta$-friendly perturbation $r$ and every function $f$ lying within some appropriate family $\mathcal{F}$. For now, we will focus on second moment estimation under $W_{\|\cdot\|_2}$ (we consider second moment estimation because mean estimation is trivial under $W_{\|\cdot\|_2}$). This corresponds to the loss function

$$L(p,S) = \|\mathbb{E}_{x\sim p}[xx^\top] - S\|. \tag{228}$$

For notational convenience we also typically denote $W_{\|\cdot\|_2}$ as $W_1$.

For the loss $L(p,S)$, we will take our family $\mathcal{F}$ to be all functions of the form $f_v(x) = \langle x,v\rangle^2$ with $\|v\|_2 = 1$. Thus we define the $(\rho,\epsilon)$-resilient distributions under $W_1$ as

$$\mathcal{G}^{W_1}_{\text{sec}}(\rho,\epsilon) = \{p \mid |\mathbb{E}_r[\langle x,v\rangle^2] - \mathbb{E}_p[\langle x,v\rangle^2]| \leq \rho \text{ whenever } r \text{ is } \epsilon\text{-friendly under } \langle x,v\rangle^2 \text{ and } \|v\|_2 = 1\}. \tag{229}$$

Note the twist in the definition of $\mathcal{G}^{W_1}_{\text{sec}}$–the allowed $r$ depends on the current choice of $v$, since friendliness is specific to the function $f_v = \langle x,v\rangle^2$, which is different from deletions in the $\mathsf{TV}$ case.

We will first show that $\mathcal{G}^{W_1}_{\text{sec}}$ has small modulus, then derive sufficient moment conditions for $p$ to be $(\rho,\epsilon)$-resilient.

**Proposition 4.6.** *The set of $(\rho,\epsilon)$-resilient distributions for $W_1$ has modulus $\mathfrak{m}(\mathcal{G}^{W_1}_{\text{sec}}(\rho,2\epsilon),\epsilon) \leq 2\rho$.*

*Proof.* For a distribution $q$, let $S_q = \mathbb{E}_q[xx^\top]$. Suppose that $p_1, p_2 \in \mathcal{G}^{W_1}_{\text{sec}}(\rho,\epsilon)$ and $W_1(p_1,p_2) \leq 2\epsilon$. For any $v$, by Lemma 4.5, there exists an $r$ that is a $(2\epsilon)$-friendly perturbation of both $p_1$ and $p_2$ with respect to $\langle x,v\rangle^2$. We conclude that $|\mathbb{E}_{p_i}[\langle x,v\rangle^2] - \mathbb{E}_r[\langle x,v\rangle^2]| \leq \rho$ for $i = 1,2$, and hence $|\mathbb{E}_{p_1}[\langle x,v\rangle^2] - \mathbb{E}_{p_2}[\langle x,v\rangle^2]| \leq 2\rho$, which can be written as $|v^\top(S_{p_1} - S_{p_2})v| \leq 2\rho$. Taking the sup over $\|v\|_2 = 1$ yields $\|S_{p_1} - S_{p_2}\| \leq 2\rho$. Since $L(p_1,\theta^*(p_2)) = \|S_{p_1} - S_{p_2}\|$, this gives the desired modulus bound. □

**Sufficient conditions for $W_1$-resilience.** Recall that for mean estimation under $\mathsf{TV}$ perturbation, any distribution with bounded $\psi$-norm was $(\mathcal{O}(\epsilon\psi^{-1}(1/\epsilon)),\epsilon)$-resilient. In particular, bounded covariance distributions were $(\mathcal{O}(\sqrt{\epsilon}),\epsilon)$-resilient. We have an analogous result for $W_1$-resilience, but with a modified $\psi$ function:

**Proposition 4.7.** *Let $\psi$ be an Orlicz function, and define $\tilde{\psi}(x) = x\psi(2x)$. Suppose that $X$ (not $X-\mu$) has bounded $\tilde{\psi}$-norm: $\mathbb{E}_p[\tilde{\psi}(|v^\top X|/\sigma)] \leq 1$ for all unit vectors $v$. Also assume that the second moment of $p$ is at most $\sigma^2$. Then $p$ is $(\rho,\epsilon)$ resilient for $\rho = \max(\sigma\epsilon\psi^{-1}(\frac{2\sigma}{\epsilon}), 4\epsilon^2 + 2\epsilon\sigma)$.*

Let us interpret Proposition 4.7 before giving the proof. Take for instance $\psi(x) = x^2$. Then Proposition 4.7 asks for the 3rd moment to be bounded by $\sigma^3/4$. In that case we have $\rho = \sigma\epsilon\psi^{-1}(2\sigma/\epsilon) = \sqrt{2}\sigma^{3/2}\epsilon^{1/2}$. If the units seem weird, remember that $\epsilon$ has units of distance (before it was unitless) and hence $\sigma^{3/2}\epsilon^{1/2}$ has quadratic units, which matches the second moment estimation task.

More generally, taking $\psi(x) = x^k$, we ask for a $(k+1)$st moment bound and get error $\mathcal{O}(\sigma^{1+1/k}\epsilon^{1-1/k})$.

We now turn to proving Proposition 4.7. A helpful auxiliary lemma (here and later) proves a way to use Orlicz norm bounds:

**Lemma 4.8.** *Let $p$ and $q$ be two distributions over $\mathcal{X}$, $g : \mathcal{X} \to \mathbb{R}$ be any function, $c$ be a non-negative cost function, and $\psi$ be an Orlicz function. Then for any coupling $\pi_{p,q}$ between $p$ and $q$ and any $\sigma > 0$ we have*

$$|\mathbb{E}_{X\sim p}[g(X)] - \mathbb{E}_{Y\sim q}[g(Y)]| \le \sigma \mathbb{E}_{\pi_{p,q}}[c(X,Y)]\psi^{-1}\left(\frac{\mathbb{E}_{\pi_{p,q}}\big[c(X,Y)\psi\big(\frac{|g(X)-g(Y)|}{\sigma c(X,Y)}\big)\big]}{\mathbb{E}_{\pi_{p,q}}[c(X,Y)]}\right). \tag{230}$$

*Proof.* Note that $|\mathbb{E}_p[g(X)] - \mathbb{E}_q[g(Y)]| = |\mathbb{E}_\pi[g(X) - g(Y)]|$. We weight the coupling $\pi$ by the cost $c$ to obtain a new probability measure $\pi'(x,y) = c(x,y)\pi(x,y)/\mathbb{E}[c(x,y)]$. We apply Jensen's inequality under $\pi'$ as follows:

$$\psi\Big(\Big|\frac{\mathbb{E}_\pi[g(X)-g(Y)]}{\sigma\mathbb{E}_\pi[c(X,Y)]}\Big|\Big) = \psi\Big(\Big|\mathbb{E}_\pi\Big[\frac{c(X,Y)}{\mathbb{E}[c(X,Y)]}\cdot\frac{g(X)-g(Y)}{\sigma c(X,Y)}\Big]\Big|\Big) \tag{231}$$

$$= \psi\Big(\Big|\mathbb{E}_{\pi'}\Big[\frac{g(X)-g(Y)}{\sigma c(X,Y)}\Big]\Big|\Big) \tag{232}$$

$$\le \mathbb{E}_{\pi'}\Big[\psi\Big(\frac{|g(X)-g(Y)|}{\sigma c(X,Y)}\Big)\Big] \tag{233}$$

$$= \mathbb{E}_\pi\Big[c(X,Y)\psi\big(\frac{|g(X)-g(Y)|}{\sigma c(X,Y)}\big)\Big]/\mathbb{E}_\pi[c(X,Y)]. \tag{234}$$

Inverting $\psi$ yields the desired result. □

*Proof of Proposition 4.7.* We apply Lemma 4.8 with $q = r$ an $\epsilon$-friendly perturbation of $p$ under $\langle x, v\rangle^2$, and $g = \langle x, v\rangle^2$; we will also use cost $c'(x,y) = |v^\top(x-y)|$, which satisfies $c'(x,y) \le c(x,y)$. Taking $\pi$ to be the $\epsilon$-friendly coupling (under $c$, not $c'$) between $p$ and $r$ yields

$$|\mathbb{E}_p[\langle x,v\rangle^2] - \mathbb{E}_r[\langle x,v\rangle^2]| \le \sigma\epsilon\psi^{-1}\left(\frac{\mathbb{E}_\pi\big[|\langle x-y,v\rangle|\psi\big(\frac{|\langle x,v\rangle^2 - \langle y,v\rangle^2|}{\sigma|\langle x-y,v\rangle|}\big)\big]}{\epsilon}\right) \tag{235}$$

$$= \sigma\epsilon\psi^{-1}\left(\frac{\mathbb{E}_\pi\big[|\langle x-y,v\rangle|\psi\big(|\langle x,v\rangle + \langle y,v\rangle|/\sigma\big)\big]}{\epsilon}\right). \tag{236}$$

Now we will split into two cases. First, we observe that the worst-case friendly perturbation will either move all of the $\langle x,v\rangle^2$ upwards, or all of the $\langle x,v\rangle^2$ downwards, since otherwise we could take just the upwards part or just the downwards part and perturb the mean further. In other words, we either have (i) $\langle x,v\rangle^2 \ge \langle y,v\rangle^2$ for all $(x,y) \in \operatorname{supp}(\pi)$ with $x \ne y$, or (ii) $\langle x,v\rangle^2 \le \langle y,v\rangle^2$ for all $(x,y) \in \operatorname{supp}(\pi)$ with $x \ne y$. We analyze each case in turn.

Case (i): $y$ moves downwards. In this case we can use the bounds $|\langle x-y,v\rangle| \le 2|\langle x,v\rangle|$ and $|\langle x+y,v\rangle| \le 2|\langle x,v\rangle|$ together with (236) to conclude that

$$|\mathbb{E}_p[\langle x,v\rangle^2] - \mathbb{E}_r[\langle x,v\rangle^2]| \le \sigma\epsilon\psi^{-1}\Big(\mathbb{E}_\pi\Big[2|\langle x,v\rangle|\psi\Big(\frac{2|\langle x,v\rangle|}{\sigma}\Big)\Big]/\epsilon\Big) \tag{237}$$

$$= \sigma\epsilon\psi^{-1}\Big(\mathbb{E}_p\Big[2\sigma\tilde{\psi}\Big(\frac{|\langle x,r\rangle|}{\sigma}\Big)\Big]/\epsilon\Big) \tag{238}$$

$$\le \sigma\epsilon\psi^{-1}(2\sigma/\epsilon), \tag{239}$$

where the final inequality is by bounded Orlicz norm of $p$.

Case (ii): $y$ moved upwards. In this case by friendliness we have that $|\langle y,v\rangle|^2 \le v^\top S_r v$ whenever $(x,y) \in \operatorname{supp}(\pi)$ and $y \ne x$. Thus

$$|\langle x-y,v\rangle|\psi(|\langle x,v\rangle + \langle y,v\rangle|/\sigma) \le |\langle x-y,v\rangle|\psi(2|\langle y,v\rangle|/\sigma) \le |\langle x-y,v\rangle|\psi(2\sqrt{v^\top S_r v}/\sigma). \tag{240}$$

for all $(x,y) \in \operatorname{supp}(\pi)$. Plugging back into (236) yields

$$|\mathbb{E}_p[\langle x,v\rangle^2] - \mathbb{E}_r[\langle x,v\rangle^2]| \le \sigma\epsilon\psi^{-1}(\mathbb{E}_\pi[|\langle x-y,v\rangle|\psi(2\sqrt{v^\top S_r v}/\sigma)]/\epsilon) \tag{241}$$

$$\le \sigma\epsilon\psi^{-1}(\epsilon\cdot\psi(2\sqrt{v^\top S_r v}/\sigma)/\epsilon) \tag{242}$$

$$= \sigma\epsilon\cdot 2v^\top S_r v/\sigma = 2\epsilon\sqrt{v^\top S_r v}. \tag{243}$$

Here the final inequality is because $\mathbb{E}_\pi[|\langle x-y, v\rangle|] \le \mathbb{E}_\pi[c(x,y)] \le \epsilon$ under the coupling. Comparing the left-hand-side to the final right-hand-side yields $|v^\top S_p v - v^\top S_r v| \le 2\epsilon\sqrt{v^\top S_r v}$. Thus defining $\Delta = |v^\top S_p v - v^\top S_r v|$ and using the fact that $v^\top S_p v \le \sigma^2$, we obtain $\Delta \le 2\epsilon\sqrt{\Delta + \sigma^2}$, which implies (after solving the quadratic) that $\Delta \le 4\epsilon^2 + 2\epsilon\sigma$.

Thus overall we have $|\mathbb{E}_p[\langle x, v\rangle^2] - \mathbb{E}_r[\langle x, v\rangle^2]| \le \max(\sigma\epsilon\psi^{-1}(2\sigma/\epsilon), 4\epsilon^2 + 2\epsilon\sigma)$, as was to be shown. □

## 4.2 Other Results

Our understanding of robust estimation under $W_c$ distances is still rudimentary. Below are a couple of known results, but many of these may be improved or extended in the near future (perhaps by you!).

The most straightforward extension is from second moment estimation to $k$th moment estimation. In that case instead of using $\tilde{\psi}(x) = x\psi(2x)$, we use $\tilde{\psi}(x) = x\psi(kx^{k-1})$. Essentially the same proof goes through.

We can also extend to more general loss functions $L(p, \theta)$, as long as $L$ is a convex function of $p$ for fixed $\theta$ (this holds e.g. for any $L(p, \theta) = \mathbb{E}_{x\sim p}[\ell(\theta; x)]$, since these loss functions are linear in $p$ and hence also convex). Here the main challenge is defining an appropriate family $\mathcal{F}$ of functions for which to consider friendly perturbations. For second moment estimation our family $\mathcal{F}$ was motivated by the obsevation that $L(p, S) = \sup\{|\mathbb{E}_p[f_v(x)] - \mathbb{E}_q[f_v(x)]| \mid f_v(x) = \langle x, v\rangle^2, \|v\|_2 = 1\}$, but such linear structure need not hold in general. But we can still exploit linear structure by looking at subgradients of the loss. In particular, we can take the Fenchel-Moreau representation

$$L(p, \theta) = \sup_{f \in \mathcal{F}_\theta} \mathbb{E}_{x\sim p}[f(x)] - L^*(f, \theta), \tag{244}$$

which exists for some $\mathcal{F}_\theta$ and $L^*$ whenever $L(p, \theta)$ is convex in $p$. The family $\mathcal{F}_\theta$ is roughly the family of subgradients of $L(p, \theta)$ as $p$ varies for fixed $\theta$. In this case we obtain conditions $G_\downarrow$ and $G_\uparrow$ as before, asking that

$$\mathbb{E}_r[f(x)] - L^*(f, \theta^*(p)) \le \rho_1 \text{ for all } f \in \mathcal{F}_{\theta^*(p)} \text{ and } \epsilon\text{-friendly } r, \tag{$\downarrow$}$$

and furthermore

$$L(p, \theta) \le \rho_2 \text{ if for every } f \in \mathcal{F}_\theta \text{ there is an } \epsilon\text{-friendly } r \text{ such that } \mathbb{E}_r[f(x)] - L^*(f, \theta) \le \rho_1. \tag{$\uparrow$}$$

Note that for the second condition ($\mathcal{G}_\downarrow$), we allow the perturbation $r$ to depend on the current function $f$. If $r$ was fixed this would closely match the old definition, but we can only do that for deletions since in general even the set of feasible $r$ depends on $f$.

Using this, we can (after sufficient algebra) derive sufficient conditions for robust linear regression under $W_1$, for conditions similar to the hypercontractivity condition from before. This will be a challenge problem on the homework.

Finally, we can define a $\tilde{W}_1$ similar to $\widetilde{\mathsf{TV}}$, but our understanding of it is far more rudimentary. In particular, known analyses do not seem to yield the correct finite-sample rates (for instance, the rate of convergence includes an $n^{-1/3}$ term that seems unlikely to actually exist).

# 5 Inference under Model Mis-specification

## 5.1 Model Mis-specification in Generalized Linear Models

## 5.2 Robust Inference via the Bootstrap

[Lecture 17]

## 5.3 Robust Inference via Partial Specification

In the previous section we saw how using a non-parametric inference method–the bootstrap–allowed us to avoid the pitfalls of mis-specified parametric models. Next we will explore a different idea, called *partial specification* or *robust standard errors*. Here we stay within a parametric model, but we derive algebraic formulas that hold even when the particular parametric model is wrong, as long as certain "orthogonality assumptions" are true.

Specifically, we will consider linear regression, deriving standard error estimates via typical parametric confidence regions as with GLMs. We will see that these are brittle, but that they are primarily brittle because they implicitly assume that certain equations hold. If we instead explicitly subtitute the right-hand side of those equations, we get better confidence intervals that hold under fewer assumptions. As a bonus, we'll be able to study how linear regression performs under distribution shift.

**Starting point: linear response with Gaussian errors.** In the simplest setting, suppose that we completely believe our model:

$$Y = \langle \beta, X \rangle + Z, \text{ where } Z \sim \mathcal{N}(0, \sigma^2 I). \tag{245}$$

We observe samples $(x_1, y_1), \dots, (x_n, y_n) \sim p$. Suppose that we estimate $\beta$ using the ordinary least squares estimator:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \beta, x_i \rangle)^2 = (\sum_{i=1}^{n} x_i x_i^\top)^{-1} \sum_{i=1}^{n} x_i y_i. \tag{246}$$

Define $S = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top$. Then since $y_i = x_i^\top \beta + z_i$, we can further write

$$\hat{\beta} = (\sum_{i=1}^{n} x_i x_i^\top)^{-1} (\sum_{i=1}^{n} x_i x_i^\top \beta + x_i z_i) \tag{247}$$

$$= (nS)^{-1}(nS\beta + \sum_{i=1}^{n} x_i z_i) \tag{248}$$

$$= \beta + \frac{1}{n} S^{-1} \sum_{i=1}^{n} x_i z_i. \tag{249}$$

From this we see that, conditional on the $x_i$, $\hat{\beta} - \beta$ is a zero-mean Gaussian distribution. Its covariance matrix is given by

$$\frac{1}{n^2} S^{-1} \sum_{i=1}^{n} \mathbb{E}[z_i^2 \mid x_i] x_i x_i^\top S^{-1} = \frac{\sigma^2}{n} S^{-1}. \tag{250}$$

**Confidence regions.** The above calculation shows that the error $\hat{\beta} - \beta$ is *exactly* Gaussian with covariance matrix $\frac{\sigma^2}{n} S^{-1}$ (at least assuming the errors $z_i$ are i.i.d. Gaussian). Thus the (parametric) confidence region for $\hat{\beta} - \beta$ would be an ellipsoid with shape $S^{-1}$ and radius depending on $\sigma$, $n$, and the significance level $\alpha$ of the test. As a specific consequence, the standard error for $\beta_i$ is $\sigma\sqrt{(S^{-1})_{ii}/n}$. This is the standard error estimate returned by default in most software packages.

Of course, this all so far rests on the assumption of Gaussian error. Can we do better?

**Calculation from moment assumptions.** It turns out that our calculation above relied only on conditional moments of the errors, rather than Gaussianity. We will show this explicitly by doing the calculations more carefully. Re-using steps above, we have that

$$\hat{\beta} - \beta = \frac{1}{n} S^{-1} \sum_{i=1}^{n} x_i z_i. \tag{251}$$

In particular, assuming that the $(x_i, y_i)$ are i.i.d., we have

$$\mathbb{E}[\hat{\beta} - \beta \mid x_1, \dots, x_n] = \frac{1}{n} S^{-1} \sum_{i=1}^{n} x_i \mathbb{E}[z_i \mid x_i] = S^{-1} b, \tag{252}$$

where $b \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} x_i \mathbb{E}[z_i \mid x_i]$.

In particular, as long as $\mathbb{E}[Z \mid X] = 0$ for all $X$, $\hat{\beta}$ is an unbiased estimator for $X$. In fact, since this only needs to hold on average, as long as $\mathbb{E}[ZX] = 0$ (covariates uncorrelated with noise) then $\mathbb{E}[\hat{\beta} - \beta] = 0$, and $\mathbb{E}[\hat{\beta} - \beta \mid x_{1:n}]$ converges to zero as $n \to \infty$. This yields an insight that is important more generally:

> *Orindary least squares yields an unbiased estimate of $\beta$ whenever the covariates $X$ and noise $Z$ are uncorrelated.*

This partly explains the success of OLS compared to other alternatives (e.g. penalizing the absolute error or fourth power of the error). While OLS might initially look like the maximum likelihood estimator under Gaussian errors, it yields consistent estimates of $\beta$ under much weaker assumptions. Minimizing the fourth power of the error requires stronger assumptions for consistency, while minimizing the absolute error would yield a different condition in terms of medians rather than expectations.

Next we turn to the covariance of $\hat{\beta}$. Assuming again that the $(x_i, y_i)$ are independent across $i$, we have

$$\begin{aligned}\mathsf{Cov}[\hat{\beta} \mid x_{1:n}] &= \mathsf{Cov}[\frac{1}{n}S^{-1}\sum_{i=1}^{n} x_i z_i \mid x_{1:n}] &(253)\\ &= \frac{1}{n^2}S^{-1}\sum_{i,j=1}^{n} x_i \mathsf{Cov}[z_i, z_j \mid x_i, x_j] x_j^\top S^{-1} &(254)\\ &= \frac{1}{n^2}S^{-1}\sum_{i=1}^{n} x_i \mathsf{Var}[z_i \mid x_i] x_i^\top S^{-1}, &(255)\end{aligned}$$

where the final line is because $z_i, z_j$ are independent for $i \neq j$. If we define $\Omega = \frac{1}{n}\sum_{i=1}^{n} x_i \mathsf{Var}[z_i \mid x_i] x_i^\top$, then the final term becomes $\frac{1}{n}S^{-1}\Omega S^{-1}$.

This quantity can be upper-bounded under much weaker assumptions than Gaussianity. If we, for instance, merely assume that $\mathsf{Var}[z_i \mid x_i] \leq \sigma^2$ for all $i$, then we have that $\Omega \preceq \sigma^2 S$ and hence $\mathsf{Cov}[\hat{\beta} \mid x_{1:n}] \preceq \frac{\sigma^2}{n}S^{-1}$.

Even better, this quantity can be estimated from data. Let $u_i^2 = (y_i - \hat{\beta}^\top x_i)^2$. This is a downward-biased, but asymptotically unbiased, estimate for $\mathsf{Var}[z_i \mid x_i]$ (it would be unbiased if we used $\beta$ instead of $\hat{\beta}$). Therefore, form the matrix

$$\hat{\Omega}_n = \frac{1}{n}\sum_{i=1}^{n} x_i u_i^2 x_i^\top. \tag{256}$$

Then $\frac{1}{n}S^{-1}\hat{\Omega}_n S^{-1}$ can be used to generate confidence regions and standard errors. In particular, the standard error estimate for $\beta_i$ is $\sqrt{(S^{-1}\hat{\Omega}_n S^{-1})_{ii}/n}$. This is called the *robust standard error* or *heteroskedacity-consistent standard error*.

There are a couple of simple improvements on this estimate. The first is a “degrees of freedom” correction: we know that $u_i^2$ is downward-biased, and it is more downward-biased the larger the dimension $d$ (because then $\hat{\beta}$ can more easily overfit). We often instead use $\frac{1}{n-d}S^{-1}\hat{\Omega}_n S^{-1}$, which corrects for this.

A fancier correction, based on the jacknnife, first corrects the errors $u_i$, via

$$u_i' = u_i/(1 - \kappa_i), \text{ with } \kappa_i = \frac{1}{n}x_i^\top S^{-1} x_i.$$

We obtain a corresponding $\Omega_n' = \frac{1}{n}\sum_{i=1}^{n} x_i (u_i')^2 x_i^\top$, and the matrix for the standard errors is

$$\frac{1}{n}S^{-1}(\Omega_n' - \zeta\zeta^\top)S^{-1}, \text{ where } \zeta = \frac{1}{n}\sum_{i=1}^{n} x_i u_i'.$$

The main difference is that each $u_i$ gets a different correction factor $\frac{1}{1-\kappa_i}$ (which is however roughly equal to $\frac{n}{n-d}$) and also that we subtract off the mean $\zeta$. There is some evidence that this more complicated estimator works better when the sample size is small, see for instance MacKinnon and White (1985).

**Out-of-distribution error.** Now suppose that we wish to estimate the error on test samples $\bar{x}_{1:m}$ drawn from a distribution $\bar{p} \neq p$. Start again with the Gaussian assumption that $y = \beta^\top x + z$.

The expected error on sample $\bar{x}_i$ (over test noise $\bar{z}_i$) is $\sigma^2 + \langle \hat{\beta} - \beta, \bar{x}_i \rangle^2$. If we let $\bar{S} = \frac{1}{m} \sum_{i=1}^m \bar{x}_i \bar{x}_i^\top$, and let $\mathbb{E}_Z$ denote the expectation with respect to the training noise $z_1, \ldots, z_n$, then the overall average expected error (conditional on $x_{1:n}, \bar{x}_{1:m}$) is

$$\sigma^2 + \mathbb{E}_Z[\frac{1}{m} \sum_{i=1}^m (\bar{x}_i^\top (\beta - \hat{\beta}))^2] = \sigma^2 + \langle \frac{1}{m} \sum_{i=1}^m \bar{x}_i \bar{x}_i^\top, \mathbb{E}_Z[(\beta - \hat{\beta})(\beta - \hat{\beta})^\top] \rangle \tag{257}$$

$$= \sigma^2 + \langle \bar{S}, \frac{\sigma^2}{n} S^{-1} \rangle \tag{258}$$

$$+ \sigma^2 \Big(1 + \frac{1}{n} \langle \bar{S}, S^{-1} \rangle \Big). \tag{259}$$

This shows that the error depends on the divergence between the second moment matrices of $p(x)$ and $\bar{p}(x)$:

- When $p(x) = \bar{p}(x)$, then $\langle \bar{S}, S^{-1} \rangle = \mathrm{tr}(\bar{S} S^{-1}) \approx \mathrm{tr}(I) = d$, so the error decays as $\frac{d}{n}$.
- If $S$ is low-rank and is missing any directions that appear in $\bar{S}$, then the error is infinite. This makes sense, as we have no way of estimating $\beta$ along the missing directions, and we need to be able to estimate $\beta$ in those directions to get good error under $\bar{p}$. We can get non-infinite bounds if we further assume some norm bound on $\beta$; e.g. if $\|\beta\|_2$ is bounded then the missing directions only contribute some finite error.
- On the other hand, if $S$ is full-rank but $\bar{S}$ is low-rank, then we still achieve finite error. For instance, suppose that $S = I$ is the identity, and $\bar{S} = \frac{d}{k} P$ is a projection matrix onto a $k$-dimensional subspace, scaled to have trace $d$. Then we get a sample complexity of $\frac{d}{n}$, although if we had observed samples with second moment matrix $\bar{S}$ at training time, we would have gotten a better sample complexity of $\frac{k}{n}$.
- In general it is always better for $S$ to be bigger. This is partially an artefact of the noise $\sigma^2$ being the same for all $X$, so we would always rather have $X$ be as far out as possible since it pins down $\beta$ more effectively. If the noise was proportional to $\|X\|_F$ (for instance) then the answer would be different.

**Robust OOD error estimate.** We can also estimate the OOD error even when the Gaussian assumption doesn't hold, using the same idea as for robust standard errors. Letting $\bar{z}_i$ be the noise for $\bar{x}_i$, the squared error is then $\frac{1}{m} \sum_{j=1}^m (\langle \beta - \hat{\beta}, \bar{x}_i \rangle + \bar{z}_i)^2$, and computing the expectation given $x_{1:n}, \bar{x}_{1:m}$ yields

$$\mathbb{E}[\frac{1}{m} \sum_{j=1}^m (\langle \beta - \hat{\beta}, \bar{x}_i \rangle + \bar{z}_i)^2 \mid x_{1:n}, \bar{x}_{1:m}] \tag{260}$$

$$= \frac{1}{m} \sum_{i=1}^m \bar{x}_i^\top \mathbb{E}[(\beta - \hat{\beta})(\beta - \hat{\beta})^\top \mid x_{1:n}] \bar{x}_i + 2 \bar{x}_i^\top \mathbb{E}[\beta - \hat{\beta} \mid x_{1:n}] \mathbb{E}[\bar{z}_i \mid x_i] + \mathbb{E}[\bar{z}_i^2 \mid \bar{x}_i] \tag{261}$$

$$= \Big\langle \bar{S}, S^{-1} \Big( \frac{1}{n} \Omega + b b^\top \Big) S^{-1} \Big\rangle + 2 \Big\langle \bar{b}, S^{-1} b \Big\rangle + \frac{1}{m} \sum_{j=1}^m \mathbb{E}[\bar{z}_i^2 \mid \bar{x}_i]. \tag{262}$$

To interpret this expression, first assume that the true model is "actually linear", meaning that $b = \bar{b} = 0$. Then the expression reduces to $\frac{1}{n} \langle \bar{S}, S^{-1} \Omega S^{-1} \rangle + \frac{1}{m} \sum_{j=1}^m \mathbb{E}[\bar{z}_i^2 \mid x_i]$. The second term is the intrinsic variance in the data, while the first term is similar to the $\frac{1}{n} \langle \bar{S}, S^{-1} \rangle$ term from before, but accounts for correlation between $X$ and the variation in $Z$.

If the model is not actually linear, then we need to decide how to define $\beta$ (since the optimal $\beta$ is then no longer independent of the distribution). In that case a natural choice is to let $\beta$ be the minimizer under the training distribution, in which case $b \to 0$ as $n \to \infty$ and thus the $\langle \bar{b}, S^{-1} b \rangle$ term conveniently becomes asymptotically negligible. The twist is that $\mathbb{E}[\bar{z}_i^2 \mid \bar{x}_i]$ now measures not just the intrinsic variance but also the departure from linearity, and could be quite large if the linear extrapolation away from the training points ends up being poor.

**Partial specification.** In general, we see that we can actually form good estimates of the mean-squared error on $\bar{p}$ making only certain moment assumptions (e.g. $b = \bar{b} = 0$) rather than needing to assume the Gaussian model is correct. This idea is called *partial specification*, where rather than making assumptions that are stringent enough to specify a parametric family, we make weaker assumptions that are typically insufficient to even yield a likelihood, but show that our estimates are still valid under those weaker assumptions. The weaker the assumptions, the more happy we are. Of course $b = \bar{b} = 0$ is still fairly strong, but much better than Gaussianity. The goal of partial specification aligns with our earlier desire to design estimators for the entire family of resilient distributions, rather than specific parametric classes. We will study other variants of partial specification later in the course, in the context of clustering algorithms.

[Lecture 18]

## 5.4 Partial Specification and Agnostic Clustering

We next study the idea of partial specification for clustering. Our setting for clustering will be the following:

- There are $k$ unknown distributions $p_1, \dots, p_k$.
- We observe points $x_1, \dots, x_n$, such that a fraction $\alpha_j$ of the points $x_i$ are drawn from $p_j$.

Generally the $\alpha_j$ are not known but we have a lower bound on $\alpha_{\min} = \min_{j=1}^{k} \alpha_j$. In clustering we have two goals:

- **Parameter recovery:** We wish to estimate some parameter of the $p_j$ (usually their means).
- **Cluster recovery:** We wish to determine for each point $x_i$ which cluster $p_j$ it was drawn from.

In the simplest setting, we assume that each of the $p_j$ has a known parametric form (for instance, each $p_j$ is a Gaussian with unknown mean and variance). In the *agnostic* setting, we do not assume a parametric form for the $p_j$ but instead only assume e.g. bounded moments. In the *robust* setting, we allow some fraction $\epsilon$ of the points to be arbitrary outliers (so $\alpha_1 + \cdots + \alpha_k = 1 - \epsilon$).

Partial specification thus corresponds to the agnostic setting. Clustering is a particularly interesting setting for studying partial specification because some algorithms that work in the simple setting fail completely in the agnostic setting. Below we will first study the simple setting and give an algorithm based on the method of moments, then turn our attention to the agnostic setting. In the agnostic setting, resilience will appear once again as an information-theoretically sufficient condition enabling clustering. Finally, we will turn our attention to efficient algorithms. In many cases the agnostic algorithms will work even in the robust agnostic setting.

### 5.4.1 Clustering Mixtures of Gaussians

Here we assume that each $p_j = \mathcal{N}(\mu_j, \Sigma_j)$. Thus we can treat each $x_i$ as being drawn from $p = \sum_{j=1}^{k} \alpha_j \mathcal{N}(\mu_j, \Sigma_j)$. This is a parametric model with parameters $(\alpha_j, \mu_j, \Sigma_j)$, so (at least in the limit of infinite data) a sufficient condition for exact parameter recovery is for the model to be identifiable, meaning that if $\sum_{j=1}^{k} \alpha_j \mathcal{N}(\mu_j, \Sigma_j) = \sum_{j=1}^{k} \alpha'_j \mathcal{N}(\mu'_j, \Sigma'_j)$, then $\alpha_j = \alpha'_j$, $\mu_j = \mu'_j$, and $\Sigma_j = \Sigma'_j$.[5]

As stated, the model is never identifiable because we can always permute the $(\alpha_j, \mu_j, \Sigma_j)$ and obtain an identical distribution. What we actually care about is *identifiability up to permutation*: if $p_{\alpha,\mu,\Sigma} = p_{\alpha',\mu',\Sigma'}$ then $\alpha_j = \alpha'_{\sigma(j)}$, $\mu_j = \mu'_{\sigma(j)}$, and $\Sigma_j = \Sigma'_{\sigma(j)}$ for some permutation $\sigma$.

We have the following result:

**Proposition 5.1.** *As long as the orders pairs $(\mu_j, \Sigma_j)$ are all distinct, the parameters $(\alpha_j, \mu_j, \Sigma_j)$ are identifiable up to permutation.*

[5] We also need to worry about the case where $k \neq k'$, but for simplicity we ignore this.

*Proof.* This is equivalent to showing that the functions $f_{\mu,\Sigma}(x)$ defining the pdf of a Gaussian are all linearly independent (i.e., there is no non-trivial finite combination that yields the zero function). We will start by showing this in one dimension. So, suppose for the sake of contradiction that

$$\sum_{j=1}^{m} c_j \exp(-(x-\mu_j)^2/2\sigma_j^2)/\sqrt{2\pi\sigma^2} = 0, \tag{263}$$

where the $c_j$ are all non-zero. Then integrating (263) against the function $\exp(\lambda x)$ and using the formula for the moment generating function of a Gaussian, we obtain

$$\sum_{j=1}^{m} c_j \exp(\frac{1}{2}\sigma_j^2\lambda^2 + \mu_j\lambda) = 0. \tag{264}$$

Let $\sigma_{\max} = \max_{j=1}^{m} \sigma_j$, then dividing the above equation by $\exp(\frac{1}{2}\sigma_{\max}^2\lambda^2)$ and taking $\lambda \to \infty$, we see that only those $j$ such that $\sigma_j = \sigma_{\max}$ affect the limit. If $S$ is the set of such indices $j$, we obtain

$$\sum_{j\in S} c_j \exp(\mu_j\lambda) = 0, \tag{265}$$

i.e. there is a linear relation between the functions $g_{\mu_j}(\lambda) = \exp(\mu_j\lambda)$. But this is impossible, because as long as the $\mu_j$ are distinct, the largest $\mu_j$ will always dominate the limit of the linear relation as $\lambda \to \infty$, and so we must have $c_j = 0$ for that $j$, a contradiction.

It remains to extend to the $n$-dimensional case. Suppose there was a linear relation among the PDFs of $n$-dimensional Gaussians with distinct parameters. Then if we project to a random 1-dimensional subspace, the corresponding marginals (which are linear functions of the $n$-dimensional PDFs) are also each Gaussian, and have distinct parameters with probability 1. This is again a contradiction since we already know that distinct 1-dimensional Gaussians cannot satisfy any non-trivial linear relation. □

Proposition 5.1 shows that we can recover the parameters exactly in the limit of infinite data, but it doesn't say anything about finite-sample rates. However, asymptotically, as long as the log-likelihood function is locally quadratic around the true parameters, we can use tools from asymptotic statistics to show that we approach the true parameters at a $1/\sqrt{n}$ rate.

**Recovery from moments.** Proposition 5.1 also leaves open the question of efficient computation. In practice we would probably use $k$-means or EM, but another algorithm is based on the *method of moments*. It has the virtue of being provably efficient, but is highly brittle to mis-specification.

The idea is that the first, second, and third moments give a system of equations that can be solved for the parameters $(\alpha, \mu, \Sigma)$: letting $p = \sum_j \alpha_j \mathcal{N}(\mu_j, \Sigma_j)$, we have

$$\mathbb{E}_p[X] = \sum_{j=1}^{k} \alpha_j\mu_j, \tag{266}$$

$$\mathbb{E}_p[X \otimes X] = \sum_{j=1}^{k} \alpha_j(\mu_j\mu_j^\top + \Sigma_j), \tag{267}$$

$$\mathbb{E}_p[X \otimes X \otimes X] = \sum_{j=1}^{k} \alpha_j(\mu_j^{\otimes 3} + 3\,\mathrm{Sym}(\mu_j \otimes \Sigma_j)), \tag{268}$$

where $\mathrm{Sym}(X)_{i_1i_2i_3} = \frac{1}{6}(X_{i_1i_2i_3} + X_{i_1i_3i_2} + X_{i_2i_1i_3} + X_{i_2i_3i_1} + X_{i_3i_1i_2} + X_{i_3i_2i_1})$.

In $d$ dimensions, this yields $d + \binom{d+1}{2} + \binom{d+2}{3} \approx d^3/6$ equations and $k(1 + d + \binom{d+1}{2}) \approx kd^2/2$ unknowns. Thus as long as $d > 3k$ we might hope that these equations have a unique (up to permutation) solution for $(\alpha, \mu, \Sigma)$. As an even more special case, if we assume that the covariance matrices are all diagonal, then we only have approximately $2kd$ unknowns, and the equations have a solution whenever the $\mu_j$ are linearly independent. We can moreover find this solution via an efficient algorithm called the *tensor power method*, which is a generalization of the power method for matrices, and the rate of convergence is polynomial in $k$, $d$, and the condition number of certain matrices (and decays as $1/\sqrt{n}$).

However, this method is very brittle—it relies on exact algebraic moment relations of Gaussians, so even small departures from the assumptions (like moving from Gaussian to sub-Gaussian) will likely break the algorithm. This is one nice thing about the agnostic clustering setting—it explicitly reveals the brittleness of algorithms like the one above, and (as we shall see) shows why other algorithms such as $k$-means are likely to perform better in practice.

**Cluster recovery.** An important point is that even in this favorable setting, exact cluster recovery is impossible. This is because even if the Gaussians are well-separated, there is some small probability that a sample ends up being near the center of a different Gaussian.

To measure this quantitatively, assume for simplicity that $\Sigma_j = \sigma^2 I$ for all $j$ (all Gaussians are isotropic with the same variance), and suppose also that the $\mu_j$ are known exactly and that we assign each point $x$ to the cluster that minimizes $\|x - \mu_j\|_2$.[6] Then the error in cluster recovery is exactly the probability that a sample from $\mu_j$ ends up closer to some other sample $\mu_{j'}$, which is

$$\sum_{j=1}^{k} \alpha_j \mathbb{P}_{x \sim \mathcal{N}(\mu_j, \sigma^2 I)}[\|x - \mu_j\|_2 > \|x - \mu_{j'}\|_2 \text{ for some } j' \neq j] \leq \sum_{j=1}^{k} \alpha_j \sum_{j' \neq j} \Phi(\|\mu_j - \mu_{j'}\|/\sigma) \tag{269}$$

$$\leq k\Phi(\Delta/\sigma), \tag{270}$$

where $\Delta = \min_{j' \neq j} \|\mu_j - \mu_{j'}\|_2$ and $\Phi$ is the normal CDF. As long as $\Delta \gg \sqrt{\log(k/\epsilon)}$, the cluster error will be at most $\epsilon$. Note that the cluster error depends on a *separation condition* stipulating that the cluster centers are all sufficiently far apart. Moreover, we need greater separation if there are more total clusters (albeit at a slowly-growing rate in the Gaussian case).

### 5.4.2 Clustering Under Resilience

The mixture of Gaussians case is unsatisfying because data are unlikely to actually be Gaussian mixtures in practice, yet common algorithms like $k$-means still do a good job at clustering data. We therefore move to the agnostic setting, and show that we only need the distributions to be *resilient* in order to cluster successfully.

We will start by proving an even stronger result—that if a set of points contains a $(\rho, \alpha)$-resilient subset $S$ of size $\alpha n$, then it is possible to output an estimate $\hat{\mu}$ that is close to the true mean $\mu$ of $S$, regardless of the other $(1 - \alpha)n$ points. As stated, this is impossible, since there could be $\mathcal{O}(1/\alpha)$ identical clusters in the data. So what we will actually show is a *list-decoding* result—that it is possible to output $\mathcal{O}(1/\alpha)$ "candidates" $\hat{\mu}_l$ such that one of them is close to the mean of $S$:

**Proposition 5.2.** *Suppose that a set of points $\tilde{S} = \{x_1, \dots, x_n\}$ contains a $(\rho, \alpha/4)$-resilient set $S$ with mean $\mu$. Then if $|S| \geq \alpha n$ (even if $\alpha < \frac{1}{2}$), it is possible to output $m \leq \frac{2}{\alpha}$ candidates $\hat{\mu}_1, \dots, \hat{\mu}_m$ such that $\|\hat{\mu}_j - \mu\| \leq \frac{8\rho}{\alpha}$ for some $j$.*

*Proof.* The basic intuition is that we can cover the points in $\tilde{S}$ by resilient sets $S'_1, \dots, S'_{2/\alpha}$ of size $\frac{\alpha}{2}n$. Then by the pigeonhole principle, the resilient set $S$ must have large overlap with at least one of the $S'$, and hence have similar mean. This is captured in Figure 9 below.

The main difference is that $S$ and $S'$ may have relatively small overlap (in a roughly $\alpha$-fraction of elements). We thus need to care about resilience when the subset $T$ is small compared to $S$. The following lemma relates resilience on large sets to resilience on small sets:

**Lemma 5.3.** *For any $0 < \epsilon < 1$, a distribution/set is $(\rho, \epsilon)$-resilient if and only if it is $(\frac{1-\epsilon}{\epsilon}\rho, 1 - \epsilon)$-resilient.*

This was already proved in Appendix C as part of Lemma 2.14. Given Lemma 5.3, we can prove Proposition 5.2 with a similar triangle inequality argument to how we showed that resilient sets have small modulus of continuity. However, we now need to consider multiple resilient sets $S_i$ rather than a single $S'$.

[6] This is not quite optimal, in reality we would want to assign based on $\|x - \mu_j\|_2^2/\sigma^2 + \log \alpha_j$, but we consider this simpler assignment for simplicity.

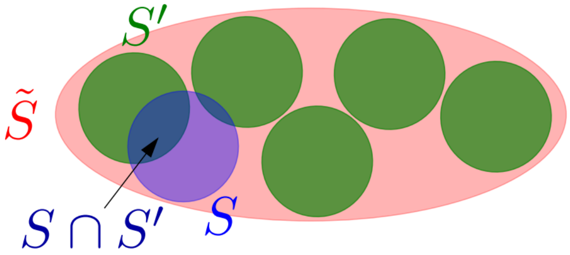


Figure 9: If we cover $\tilde{S}$ by resilient sets, at least one of the sets $S'$ has large intersection with $S$.

Suppose $S$ is $(\rho, \frac{\alpha}{4})$-resilient around $\mu$–and thus also $(\frac{4}{\alpha}\rho, 1-\frac{\alpha}{4})$-resilient by Lemma 5.3–and let $S_1, \dots, S_m$ be a maximal collection of subsets of $[n]$ such that:

1. $|S_j| \geq \frac{\alpha}{2}n$ for all $j$.
2. $S_j$ is $(\frac{4}{\alpha}\rho, 1-\frac{\alpha}{2})$-resilient (with mean $\mu_j$).
3. $S_j \cap S_{j'} = \emptyset$ for all $j \neq j'$.

Clearly $m \leq \frac{2}{\alpha}$. We claim that $S$ has large intersection with at least one of the $S_j$ and hence $\mu_j$ is close to $\mu$. By maximality of the collection $\{S_j\}_{j=1}^m$, it must be that $S_0 = S \backslash (S_1 \cup \cdots \cup S_m)$ cannot be added to the collection. First suppose that $|S_0| \geq \frac{\alpha}{2}n$. Then $S_0$ is $(\frac{4}{\alpha}\rho, 1-\frac{\alpha}{2})$-resilient (because any subset of $\frac{\alpha}{2}|S_0|$ points in $S_0$ is a subset of at least $\frac{\alpha}{4}|S|$ points in $S$). This contradicts the maximality of $\{S_j\}_{j=1}^m$, so we must have $|S_0| < \frac{\alpha}{2}n$.

Now, this implies that $|S \cap (S_1 \cup \cdots \cup S_m)| \geq \frac{\alpha}{2}n$, so by pigeonhole we must have $|S \cap S_j| \geq \frac{\alpha}{2}|S_j|$ for some $j$. Letting $T = S \cap S_j$ as before, we find that $|T| \geq \frac{\alpha}{2}|S_j| \geq \frac{\alpha}{4}|S|$ and hence by resilience of $S_j$ and $S$ we have $\|\mu - \mu_j\| \leq 2 \cdot (\frac{4}{\alpha}\rho) = \frac{8}{\alpha}\rho$ by the same triangle inequality argument as before. □

**Better bounds for well-separated clusters.** Proposition 5.2 is powerful because it holds under very minimal conditions (we do not need to assume anything about separation of clusters or even about any of the clusters other than the one we are estimating). However, its guarantees are also minimal—we only know that we get approximate parameter recovery in the list-decoding model, and cannot say anything about cluster recovery. We next obtain a stronger bound assuming that the data can actually be separated into clusters (with a small fraction of outliers) and that the means are well-separated. This stronger result both gives cluster recovery, and gives better bounds for parameter recovery:

**Proposition 5.4.** *Suppose that a set of points $\{x_1, \dots, x_n\}$ can be partitioned into $k$ sets $C_1, \dots, C_k$ of size $\alpha_1 n, \dots, \alpha_k n$, together with a fraction $\epsilon n$ of outliers ($\epsilon = 1 - (\alpha_1 + \cdots + \alpha_k)$), where $2\epsilon \leq \alpha = \min_{k=1}^k \alpha_j$. Further suppose that*

- *Each cluster is $(\rho_1, \epsilon)$-resilient and $(\rho_2, 2\epsilon/\alpha)$-resilient.*
- *The means are well-separated: $\Delta > \frac{4\rho_1}{\epsilon}$ where $\Delta = \min_{j \neq j'} \|\mu_j - \mu_{j'}\|_2$.*

*Then we can output clusters $\hat{C}_1, \dots, \hat{C}_k$ such that:*

- *$|C_j \triangle \hat{C}_j| \leq \mathcal{O}(\epsilon/\alpha)|C_j|$ (cluster recovery)*
- *The mean $\hat{\mu}_j$ of $\hat{C}_j$ satisfies $\|\hat{\mu}_j - \mu_j\|_2 \leq 2\rho_2$ (parameter recovery).*

*Proof.* We will construct a covering by resilient sets as before, but this time make use of the fact that we know the data can be approximately partitioned into clusters. Specifically, let $S_1, \dots, S_k$ be a collection of $k$ sets such that:

- $|S_l| \geq \alpha n$
- The $S_l$ are disjoint and contain all but $\epsilon n$ points.

- Each $S_l$ is $(\rho_1, \epsilon)$-resilient.

We know that such a collection exists because we can take the $C_j$ themselves. Now call a set $S$ "$j$-like" if it contains at least $\alpha_j(\epsilon/\alpha)|S|$ points from $C_j$. We claim that each $S_l$ is $j$-like for exactly one $j$. Indeed, by pigeonhole it must be $j$-like for at least one $j$ since $\epsilon/\alpha \leq 1/2 < 1$.

In the other direction, note that if $S$ if $j$-like then $S \cap C_j$ contains at least $(\alpha_j/\alpha)\epsilon$ of the points in $S$, and at least $(|S|/n)(\epsilon/\alpha) \geq \epsilon$ of the points in $C_j$. Thus by resilience of both sets, the means of both $S$ and $C_j$ are within $\frac{\rho_1}{\epsilon}$ of the mean of $S \cap C_j$ and hence within $\frac{2\rho_1}{\epsilon}$ of each other. In summary, $\|\mu_j - \mu_S\|_2 \leq \frac{2\rho_1}{\epsilon}$. Now if $S$ were $j$-like and also $j'$-like, then we would have $\|\mu_j - \mu_{j'}\|_2 \leq \frac{4\rho_1}{\epsilon}$, which contradicts the separation assumption.

Since $S_l$ is $j$-like for a unique $j$, it contains at most $(\epsilon/\alpha)|S_l|$ points from any of the other $C_{j'}$, together with at most $\epsilon n$ outliers. Moreover, since the other $S_{l'}$ are not $j$-like, $S_l$ is missing at most $\alpha_j(\epsilon/\alpha)n$ points from $C_j$. Thus $S_l \cap C_j$ is missing at most $2\epsilon/\alpha|S_l|$ points from $S_l$ and at most $\epsilon/\alpha|C_j|$ points from $C_j$. By resilience their means are thus within $2\rho_2$ of each other, as claimed. □

## 5.5 Efficient Clustering Under Bounded Covariance

We saw that resilience is information-theoretically sufficient for agnostic clustering, but we would also like to develop efficient algorithms for clustering. This is based on work in Kumar and Kannan (2010) and Awasthi and Sheffet (2012), although we will get a slightly slicker argument by using the machinery on resilience that we've developed so far.

As before, we will need a strong assumption than resilience. Specifically, we will assume that each cluster had bounded covariance and that the clusters are well-separated:

**Theorem 5.5.** *Suppose that the data points $x_1, \ldots, x_n$ can be split into $k$ clusters $C_1, \ldots, C_k$ with sizes $\alpha_n, \cdots, \alpha_k n$ and means $\mu_1, \ldots, \mu_k$, and moreover that the following covariance and separation conditions hold:*

- $\frac{1}{|C_j|}\sum_{i \in C_j}(x_i - \mu_j)(x_i - \mu_j)^\top \preceq \sigma^2 I$ *for each cluster* $C_j$,
- $\Delta \geq 36\sigma/\sqrt{\alpha}$, *where* $\Delta = \min_{j \neq j'} \|\mu_j - \mu_{j'}\|_2$.

*Then there is a polynomial-time algorithm outputting candidate clusters $\hat{C}_1, \ldots, \hat{C}_k$ and means $\hat{\mu}_1, \ldots, \hat{\mu}_k$ such that:*

- $|C_j \triangle \hat{C}_j| = \mathcal{O}(\sigma^2/\alpha\Delta^2)$ *(cluster recovery), and*
- $\|\mu_j - \hat{\mu}_j\|_2 = \mathcal{O}(\sigma^2/\alpha\Delta)$ *(parameter recovery).*

The basic idea behind the algorithm is to project each of the points $x_i$ onto the span of the top $k$ singular vectors of the data matrix $X = [x_1 \ \cdots \ x_n]$. Let $P_k$ be the projection operator onto this space. Then since the points $Px_i$ lie in only a $k$-dimensional space instead of a $d$-dimensional space, they are substantially easier to cluster. The algorithm itself has three core steps and an optional step:

1. Project points $x_i$ to $Px_i$.
2. Form initial clusters based on the $Px_i$.
3. Compute the means of each of these clusters.
4. Optionally run any number of steps of $k$-means in the original space of $x_i$, initialized with the computed means from the previous step.

We will provide more formal psuedocode later [NOTE: TBD]. For now, we focus on the analysis, which has two stages: (1) showing that the initial clustering from the first two steps is "nice enough", and (2) showing that this niceness is preserved by the $k$-means iterations in the second two steps.

**Analyzing the projection.** We start by analyzing the geometry of the points $P_k x_i$. The following lemma shows that the projected clusters are still well-separated and have small covariance:

**Lemma 5.6.** *The projected points $P_k x_i$ satisfy the covariance and separation conditions with parameters $\sigma$ and $\sqrt{\Delta^2 - 4\sigma^2/\alpha} \geq 35\sigma/\sqrt{\alpha}$:*

$$\frac{1}{|C_j|}\sum_{i \in C_j}(Px_i - P\mu_j)(Px_i - P\mu_j)^\top \preceq \sigma^2 I \text{ and } \|P\mu_j - P\mu_{j'}\|_2 \geq \sqrt{\Delta^2 - 4\sigma^2/\alpha}. \tag{271}$$

*In other words, the covariance condition is preserved, and separation is only decreased slightly.*

*Proof.* The covariance condition is preserved because the covariance matrix of the projected points for cluster $j$ is $P_k\Sigma_j P_k$, where $\Sigma_j$ is the un-projected covariance matrix. This evidently has smaller singular values than $\Sigma_k$.

The separation condition requires more detailed analysis. We start by showing that there is not much in the orthogonal component $(I - P_k)x_i$. Indeed, we have that the top singular value of $(I - P_k)x_i$ is at most $\sigma$:

$$S = \frac{1}{n}\sum_{i=1}^{n}((I - P_k)x_i)((I - P_k)x_i)^\top \preceq \sigma^2 I \tag{272}$$

This is because $P_k$ minimizes this top singular value among all $k$-dimensional projection matrices, and if we take the projection $Q_k$ onto the space spanned by the means $\mu_1, \ldots, \mu_k$, we have

$$\frac{1}{n}\sum_{i=1}^{n}((I - Q_k)x_i)((I - Q_k)x_j)^\top = \sum_{j=1}^{k}\frac{\alpha_j}{|C_j|}\sum_{i \in C_j}((I - Q_k)x_i)((I - Q_k)x_i)^\top \tag{273}$$

$$= \sum_{j=1}^{k}\frac{\alpha_j}{|C_j|}\sum_{i \in C_j}((I - Q_k)(x_i - \mu_j))((I - Q_k)(x_i - \mu_j))^\top \tag{274}$$

$$\preceq \sum_{j=1}^{k}\frac{\alpha_j}{|C_j|}\sum_{i \in C_j}(x_i - \mu_j)(x_i - \mu_j)^\top \preceq \sum_{j=1}^{k}\alpha_j\sigma^2 I = \sigma^2 I. \tag{275}$$

Given this, we know that the projections $(I - P_k)\mu_j$ must be small, since otherwise we have

$$v^\top S v = \frac{1}{n}\sum_{i=1}^{n}\langle (I - P_k)x_i, v\rangle^2 \tag{276}$$

$$\geq \frac{\alpha_j}{|C_j|}\sum_{i \in C_j}\langle (I - P_k)x_i, v\rangle^2 \tag{277}$$

$$\geq \alpha_j\Big\langle \frac{1}{|C_j|}\sum_{i \in C_j}(I - P_k)x_i, v\Big\rangle^2 \tag{278}$$

$$= \alpha_j\langle (I - P_k)\mu_j, v\rangle^2. \tag{279}$$

Consequently $\langle (I - P_k)\mu_j, v\rangle^2 \leq \sigma^2/\alpha_j$ and hence (taking $v$ to align with $(I - P_k)\mu_j$) we have $\|(I - P_k)\mu_j\|_2 \leq \sigma/\sqrt{\alpha_j}$. In particular $\|(I - P_k)(\mu_j - \mu_{j'})\|_2 \leq 2\sigma/\sqrt{\alpha}$.

Now, by the Pythagorean theorem we have

$$\|P_k(\mu_j - \mu_{j'})\|_2^2 = \|\mu_j - \mu_{j'}\|_2^2 - \|(I - P_k)(\mu_j - \mu_{j'})\|_2^2 \geq \Delta^2 - 4\sigma^2/\alpha, \tag{280}$$

and hence the projected means are separated by at least $\sqrt{\Delta^2 - 4\sigma^2/\alpha}$, as was to be shown. □

**Analyzing the initial clustering.** We now analyze the initial clustering. Call a point $i$ a *proto-center* if there are at least $\frac{\alpha}{2}n$ projected points within distance $3\sigma\sqrt{k}$ of $P_k x_i$, and call the set of these nearby points the associated *proto-cluster*.

We will show that the proto-clusters are nearly pure (have few points not from $C_j$) using a similar argument as when we analyzed resilient clustering. As before, call a proto-cluster *$j$-like* if there are at least $\frac{\alpha_j\alpha}{4}n$ points from $C_j$ in the proto-cluster.

**Lemma 5.7.** *Each proto-cluster is $j$-like for exactly one $j$.*

*Proof.* We know that it is $j$-like for at least one $j$ by the Pigeonhole principle (if not, then the proto-cluster has at most $\frac{\alpha}{4}n$ points in total, contradicting its size of at least $\frac{\alpha}{2}n$). So suppose for the sake of contradiction that it is both $j$-like and $j'$-like. By resilience, the mean of the points from $C_j$ is at most $2\sigma/\sqrt{\alpha}$ away from $P_k\mu_j$, and similarly the mean of the points from $C_{j'}$ is at most $2\sigma/\sqrt{\alpha}$ away from $P_k\mu_{j'}$. Since the cluster has radius $3\sigma\sqrt{k} \leq 3\sigma/\sqrt{\alpha}$, this implies that $\|P_k(\mu_j - \mu_{j'})\|_2 \leq 10\sigma/\sqrt{\alpha}$, contradicting the separation condition for the projected means. Thus no proto-cluster can be $j$-like for multiple $j$, which proves the lemma. □

Now since each proto-cluster is $j$-like for exactly one $j$, at least half of the points must come from that proto-cluster.

At this point we are essentially done if all we care about is constructing an efficient algorithm for cluster recovery (but not parameter recovery), since if we just extend each proto-cluster by $\mathcal{O}(\sigma)$ we are guaranteed to contain almost all of the points in its corresponding cluster, while still containing very few points from any other cluster (assuming the data are well-separated). However, parameter recovery is a bit trickier because we need to make sure that the small number of points from other clusters don't mess up the mean of the cluster. The difficulty is that while we have control over the projected distances, and can recover the projected centers $P_k\mu_j$ well, we need to somehow get back to the original centers $\mu_j$.

The key here is that for each proto-cluster, the $Px_i$ are all close to each other, and the missing component $(I - P_k)x_i$ has bounded covariance. Together, these imply that the proto-cluster is *resilient*—deleting an $\epsilon$-fraction of points can change the mean by at most $\mathcal{O}(\sigma\epsilon)$ in the $P_k$ component, and $\mathcal{O}(\sigma\sqrt{\epsilon})$ in the $(I - P_k)$ component. In fact, we have:

**Lemma 5.8.** *Let $B$ be a proto-cluster with mean $\nu$. Then*

$$\frac{1}{|B|}\sum_{i\in B}(x_i - \nu)(x_i - \nu)^\top \preceq 11\sigma^2/\alpha. \tag{281}$$

*In particular, if $B$ is $j$-like then we have $\|\mu_j - \nu\|_2 \leq 9\sigma/\sqrt{\alpha}$.*

*Proof.* The covariance bound is because the covariance of the $x_i$ are bounded in norm by at most $3\sigma\sqrt{k}$ in the $P_k$ component and hence can contribute at most $9\sigma^2 k \leq 9\sigma^2/\alpha$ to the covariance, while we get an additional $2\sigma^2/\alpha$ in an orthogonal direction because the overall second moment of the $(I - P_k)x_i$ is $\sigma^2$ and the $i \in B$ contribute to at least an $\frac{\alpha}{2}$ fraction of that.

Now, this implies that $B$ is resilient, while we already have that $C_j$ is resilient. Since $B \cap C_j$ contains at least half the points in both $B$ and $C_j$, this gives that their means are close–within distance $2(\sqrt{11}+1)\sigma/\sqrt{\alpha} < 9\sigma/\sqrt{\alpha}$. □

**Analyzing $k$-means.** We next show that $k$-means iterations preserve certain important invariants. We will call the assigned means $\hat{\mu}_j$ *$R$-close* if $\|\hat{\mu}_j - \mu_j\|_2 \leq R$ for all $j$, and we will call the assigned clusters $\hat{C}_j$ *$\epsilon$-close* if $|C_j \triangle \hat{C}_j| \leq \epsilon|C_j|$ for all $j$. We will show that if the means are $R$-close then the clusters are $\epsilon$-close for some $\epsilon = f(R)$, and that the resulting new means are then $g(R)$-close. If $R$ is small enough then we will also have $g(R) < R$ so that we obtain an invariant.

Let $\Delta_{jj'} = \|\mu_j - \mu_{j'}\|_2$, so $\Delta_{jj'} \geq \Delta$. We will show that if the $\hat{\mu}_j$ are $R$-close, then few points in $C_j$ can end up in $\hat{C}_{j'}$. Indeed, if $x_i$ ends up in $\hat{C}_{j'}$ then we must have $\|x_i - \hat{\mu}_{j'}\|_2^2 \leq \|x_i - \hat{\mu}_j\|_2^2$, which after some re-arrangement yields

$$\langle x_i - \hat{\mu}_j, \hat{\mu}_{j'} - \hat{\mu}_j\rangle \geq \frac{1}{4}\langle \hat{\mu}_{j'} - \hat{\mu}_j, \hat{\mu}_{j'} - \hat{\mu}_j\rangle. \tag{282}$$

Applying the covariance bound and Chebyshev's inequality along the vector $v = \hat{\mu}_{j'} - \hat{\mu}_j$, we see that the fraction of points in $C_j$ that end up in $\hat{C}_{j'}$ is at most $\frac{4\sigma^2}{\|\hat{\mu}_j - \hat{\mu}_{j'}\|_2^2} \leq \frac{4\sigma^2}{(\Delta_{jj'} - 2R)^2} \leq \frac{4\sigma^2}{(\Delta - 2R)^2}$. In total this means that at most $\frac{4\sigma^2 n}{(\Delta - 2R)^2}$ points from other clusters end up in $\hat{C}_j$, while at most $\frac{4k\sigma^2|C_j|}{(\Delta - 2R)^2}$ points from $C_j$ end up in other clusters. Thus we have $\epsilon \leq \frac{4k\sigma^2}{(\Delta - 2R)^2} + \frac{4\sigma^2}{\alpha(\Delta - 2R)^2} \leq \frac{8\sigma^2}{\alpha(\Delta - 2R)^2}$, so we can take

$$f(R) = \frac{8\sigma^2}{\alpha(\Delta - 2R)^2}. \tag{283}$$

Now suppose that $\gamma_{jj'}|C_j|$ points in $C_j$ are assigned to $\hat{C}_{j'}$, where we must have $\gamma_{jj'} \leq \frac{4\sigma^2}{(\Delta_{jj'}-2R)^2}$. By resilience, the mean of these points is within $\sigma/\sqrt{\gamma_{jj'}}$ of $\mu_j$ and hence within $\Delta_{jj'} + \sigma/\sqrt{\gamma_{jj'}}$ of $\mu_{j'}$. In total, then, these points can shift the mean $\hat{\mu}_{j'}$ by at most

$$\frac{\gamma_{jj'}\alpha_j n(\Delta_{jj'} + \sigma/\sqrt{\gamma_{jj'}})}{\frac{1}{2}\alpha n} \leq \frac{2\alpha_j}{\alpha}\Big(\frac{4\sigma^2\Delta_{jj'}}{(\Delta_{jj'}-2R)^2} + \frac{2\sigma^2}{\Delta_{jj'}-2R}\Big) \leq \frac{4\alpha_j}{\alpha}\Big(\frac{2\sigma^2\Delta}{(\Delta-2R)^2} + \frac{\sigma^2}{\Delta-2R}\Big). \tag{284}$$

At the same time, the $\frac{4k\sigma^2}{(\Delta-2R)^2}$ fraction of points that are missing from $C_{j'}$ can change its mean by at most $\frac{4\sigma^2\sqrt{k}}{\Delta-2R}$. Thus in total we have

$$\|\hat{\mu}_{j'} - \mu_{j'}\|_2 \leq \frac{4\sigma^2}{\Delta-2R} \cdot \Big(\sqrt{k} + \frac{1}{\alpha} + \frac{2\Delta}{\alpha(\Delta-2R)}\Big) \leq \frac{8\sigma^2(\Delta-R)}{\alpha(\Delta-2R)^2}. \tag{285}$$

In particular we can take $g(R) = \frac{8\sigma^2(\Delta-R)}{\alpha(\Delta-2R)^2}$.

As long as $R \leq \Delta/4$ we have $g(R) \leq \frac{24\sigma^2}{\alpha\Delta}$ and $f(R) \leq \frac{32\sigma^2}{\alpha\Delta^2}$, as claimed. Since our initial $R$ is $9\sigma/\sqrt{\alpha}$, this works as long as $\Delta \geq 36\sigma/\sqrt{\alpha}$, which completes the proof.

[Lecture 19]

# 6 Nonparametric Learning in Hilbert spaces

We will now shift focus to nonparametric learning, where we seek to estimate a function lying within some infinite-dimensional space. We will start by considering linear regression in reproducing kernel Hilbert spaces (RKHSs). This is the same as ordinary or ridge regression, except that all of the underlying objects are infinite-dimensional. It turns out that to handle this infinite dimensionality, it is helpful to consider a different basis representation for linear regression, in terms of the function itself rather than its parameters. This will be our starting point.

## 6.1 Preliminaries: Parameter vs. Function View

We will start by considering the usual task of ordinary linear regression: given $(x_1, y_1), \dots, (x_n, y_n)$, find $\beta$ such that $L(\beta) = \frac{1}{n}\sum_{i=1}^n (y_i - \langle \beta, x_i \rangle)^2$ is minimized.

Letting $X$ be the $n \times p$ matrix of covariates and $y$ the $n \times 1$ matrix of outputs, the solution $\hat{\beta}$ of this problem can be expressed as

$$\beta^* = (X^\top X)^{-1} X^\top y, \text{ or } \beta^* = (\underbrace{\frac{1}{n}\sum_{i=1}^n x_i x_i^\top}_{S})^{-1} \sum_{i=1}^n x_i y_i. \tag{286}$$

In addition, for any other value of $\beta$, the excess loss relative to $\beta^*$ can be expressed as

$$L(\beta) - L(\beta^*) = (\beta - \beta^*)^\top S (\beta - \beta^*), \tag{287}$$

where as above $S$ is the second moment matrix of the covariates.

Finally, we can consider instead ridge regression, where we minimize $L_\lambda(\beta) = \frac{\lambda}{n}\|\beta\|_2^2 + L(\beta)$. In this case the minimizer is equal to

$$\hat{\beta}_\lambda = (\lambda I + X^\top X)^{-1} X^\top y = (\frac{\lambda}{n} I + S)^{-1} (\frac{1}{n}\sum_{i=1}^n x_i y_i). \tag{288}$$

This is the standard view of linear regression and what we will call the *parameter view*, since it focuses on the parameters $\beta$. But it is also helpful to adopt a different view, which we'll the *function view*. The function view instead focuses on the learned function $f_\beta(x) = \beta^\top x$. Thus we instead have $L(f) = \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2$

(assume for now that $f$ is linear in $x$; by convention $L(f) = \infty$ for all other such $f$, but the derivations below will focus on the case where $L(f)$ is finite).

The excess loss is particularly easy to write down in the function view:

$$L(f) - L(f^*) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - f^*(x_i))^2 = \frac{1}{n}\|f - f^*\|_2^2, \tag{289}$$

where the $\ell_2$-norm treats $f$ as an $n$-dimensional vector of values on the $x_i$.

What is $f^* = f_{\beta^*}$ in this case? Again expressing $f^*$ as an $n$-dimensional vector, we have

$$f_{\beta^*} = X\beta^* = X(X^\top X)^{-1}X^\top y, \tag{290}$$

or in other words $f_{\beta^*}$ is the projection of $y$ onto the $p$-dimensional linear span of the $x_i$, which makes sense.

For ridge regression, we would also like to express $\|\beta\|_2^2$ in terms of $f$. This is a bit trickier, but noting that $\beta = (X^\top X)^{-1}X^\top f$, a formula that works is

$$\|\beta\|_2^2 = f^\top X(X^\top X)^{-2}X^\top f = f^\top K^\dagger f, \tag{291}$$

where $K = XX^\top$ is the $n \times n$ kernel matrix ($K_{ij} = \langle x_i, x_j\rangle$) and $\dagger$ denotes pseudoinverse. As this is also an important norm for $f$, we will give it a name, $\|f\|_K$:

$$\|f\|_K^2 = f^\top K^\dagger f \text{ if } f \text{ lies in the span of } K, \text{ and } \infty \text{ otherwise.} \tag{292}$$

Now we can write down ridge regression in the function view:

$$\hat{f}_\lambda = \arg\min_f \frac{\lambda}{n}\|f\|_K^2 + \frac{1}{n}\|f - y\|_2^2 \tag{293}$$

$$\implies \lambda K^\dagger \hat{f}_\lambda + (\hat{f}_\lambda - y) = 0 \tag{294}$$

$$\implies \hat{f}_\lambda = (\lambda K^\dagger + I)^{-1}y \tag{295}$$

$$\implies \hat{f}_\lambda = K(\lambda I + K)^{-1}y. \tag{296}$$

This is also called *kernel ridge regression*, since it depends only on the kernel matrix $K$.

A final issue is how to evaluate $f$ on a new data point $x$. It helps to first define $k_x = X^\top x$, the $n \times 1$ matrix of inner products between $x$ and the $x_i$. We then have $x = (X^\top)^\dagger k_x$ and $\beta = X^\dagger f$ and so

$$f(x) = \beta^\top x \tag{297}$$

$$= f^\top (X^\dagger)^\top X^\dagger k_x \tag{298}$$

$$= f^\top (XX^\top)^\dagger k_x = f^\top K^\dagger k_x. \tag{299}$$

Thus for instance $\hat{f}_\lambda(x) = k_x^\top K^\dagger K(\lambda I + K)^{-1}y$.

## 6.2 The infinite-dimensional case

We are interested in *nonparametric* regression, meaning regression over infinite-dimensional families of functions. The function view on regression is particularly useful in this case, because we can dispense with worrying about vectors $x_i$ and focus only on the kernel matrix $K$. In fact, we will simply posit that there exists a *kernel function* $k(x, x')$, such that for any finite set $T = \{x_1, \dots, x_n\}$ the matrix $K[T]$ defined by $K_{ij} = k(x_i, x_j)$ is positive semidefinite, and strictly positive definite whenever the $x_i$ are distinct. The corresponding function space, called the *reproducing kernel Hilbert space* of $k$, is defined as

$$\mathcal{H} = \{f \mid \sup_T \|f\|_{K[T]} < \infty\}, \tag{300}$$

with corresponding norm $\|f\|_{\mathcal{H}} = \sup_T \|f\|_{K[T]}$.

The strict positive definiteness actually implies that $k$ represents an infinite-dimensional space, since for a $p$-dimensional space $K[T]$ would have rank at most $p$ and thus eventually be only semidefinite.

To show that $\mathcal{H}$ is non-empty (and more generally that the definitions all make sense), we need two lemmas. First, evaluating $\|f\|_K$ on a larger subset $T$ always increases the norm:

**Lemma 6.1** (Restrictions are contractions)**.** *For any $f$ and any sets $T \subseteq T'$, we have $\|f\|_{K[T]} \leq \|f\|_{K[T']}$.*

Secondly, despite this, we can extend a function $f$ from any subset $T$ to all of $\mathcal{X}$ *without* increasing the norm:

**Lemma 6.2** (Isometric extension)**.** *Let $T = \{x_1, \dots, x_n\}$ and let $y_1, \dots, y_n \in \mathbb{R}$ be arbitrary target values. Then there exists an $f \in \mathcal{H}$ such that $f(x_i) = y_i$ for all $i$, and moreover $\|f\|_{\mathcal{H}}^2 = \|y\|_K^2 = y^\top K^{-1} y$.*

*Proof of Lemma 6.2.* Define $f(x) = k_x^\top K^{-1} y$. Since $k_{x_i}$ is the $i$th row of $K$, $k_{x_i}^\top K^{-1}$ is the $i$th standard basis vector, so $f(x_i) = y_i$.

It remains to show that $f$ has bounded norm. For sets $T = \{x_1, \dots, x_n\}$ and $T' = \{x'_1, \dots, x'_m\}$, let $K[T, T']$ be the (rectangular) kernel matrix where $K_{ij} = k(x_i, x'_j)$. Then $f$ evaluated on $T'$ is equal to $K[T, T']^\top K[T, T]^{-1} y$. For notational ease define $K_{11} = K[T, T]$, $K_{12} = K[T, T']$, etc. Thus

$$\|f\|_{K[T']}^2 = y^\top K[T,T]^{-1} K[T,T'] K[T',T']^{-1} K[T',T] K[T,T]^{-1} y \tag{301}$$

$$= y^\top K_{11}^{-1} K_{12} K_{22}^{-1} K_{21} K_{11}^{-1} y. \tag{302}$$

Recall we wish to show this is at most $y^\top K_{11}^{-1} y$. It suffices to show that $K_{11}^{-1} K_{12} K_{22}^{-1} K_{21} K_{11}^{-1} \preceq K_{11}^{-1}$, or $K_{12} K_{22}^{-1} K_{21} \preceq K_{11}$. But this follows from the Schur complement lemma, since $\begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}$ is positive definite. □

*Proof of Lemma 6.1.* Under the notation above we want to show that $y_1^\top K_{11}^{-1} y_1 \leq \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}^\top \begin{bmatrix} K_{11} & K_{12} \\ K_{12} & K_{22} \end{bmatrix}^{-1} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$. This follow by noting that for large $\lambda$ we have $\begin{bmatrix} K_{11} + \frac{1}{\lambda} I & 0 \\ 0 & \lambda^2 I \end{bmatrix} \succeq \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}$ and then inverting and taking limits. □

**Remark 6.3.** This route to defining $\mathcal{H}$ is different from the typical treatment which instead focuses on the representer theorem. That theorem starts with the infinite-dimensional space and then shows that the minimum-norm interpolator for a set $T$ lies within the span of $K[T]$. Here we take essentially the reverse approach, starting with an interpolating function and showing that it can be isometrically embedded in $\mathcal{H}$. I personally prefer the approach above, as it makes it much cleaner to define both $\|f\|_{\mathcal{H}}$ and $\mathcal{H}$ itself.

Note that Lemma 6.2 also shows that kernel ridge regression can be expressed intrinsically, without explicit reference to the set $T$: $\hat{f}_\lambda = \arg\min_f \frac{\lambda}{n} \|f\|_{\mathcal{H}}^2 + \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2$. This is because we can always find on $f$ on $T = \{x_1, \dots, x_n\}$ such that $\|f\|_{\mathcal{H}} = \|f\|_{K[T]}$.

Finally, since $K$ is invertible, kernel ridge regression has an even simpler form: $\hat{f}_\lambda(x) = k_x^\top (\lambda I + K)^{-1} y$. If we set $\lambda = 0$ then we recover the isometric extension $f$ of $y$ in Lemma 6.2.

[Lecture 20]

## 6.3 Bias-variance decomposition

Next we analyze the statistical error of kernel ridge regression. Let us suppose that for a fixed set $T = \{x_1, \dots, x_n\}$, we observe points $(x_i, y_i)$, where $y_i = f^*(x_i) + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Then as in the finite-dimensional case, we can analyze the bias and variance of an estimated function $\hat{f}$. For the ordinary least squares estimator $\hat{f}_0$, this is trivial, since $\hat{f}_0(x_i) = y_i$ and so the bias is zero while the variance is $\mathbb{E}[\frac{1}{n} \sum_{i=1}^n (y_i - f^*(x_i))^2] = \sigma^2$, independent of $n$. This makes sense, as $\hat{f}_0$ interpolates the data and is thus completely sensitive to noise.

Therefore consider instead the ridge estimator $\hat{f}_\lambda = K(\lambda I + K)^{-1} y$. We have $\mathbb{E}_\epsilon[y] = f^*$ while $\mathsf{Var}_\epsilon[y] = \sigma^2 I_{n \times n}$. Then $\mathbb{E}[\hat{f}_\lambda] = K(\lambda I + K)^{-1} f^*$ and so the bias-squared, $\frac{1}{n} \|\mathbb{E}[\hat{f}_\lambda] - f^*\|_2^2$, is

$$\text{Bias}^2 = \frac{1}{n} \|K(\lambda I + K)^{-1} f^* - f^*\|_2^2 = \frac{\lambda^2}{n} \|(\lambda I + K)^{-1} f^*\|_2^2. \tag{303}$$

A useful but loose upper bound can be obtained via $\|(\lambda I + K)^{-1} f^*\|_2^2 \leq \frac{1}{\lambda} (f^*)^\top K^{-1} f^* = \frac{1}{\lambda} \|f^*\|_{\mathcal{H}}^2$, so the bias is at most $\frac{\lambda}{n} \|f\|_{\mathcal{H}}^2$ assuming $f$ lies in $\mathcal{H}$ (we will later relax this assumption, and indeed show that it must be relaxed to get the optimal rates).

Next, the variance is $\frac{1}{n}\sum_{i=1}^{n} \mathsf{Var}_{\epsilon}[f(x_i)] = \mathbb{E}_{\epsilon}[\frac{1}{n}\|K(\lambda I + K)^{-1}\epsilon\|_2^2]$, or

$$\mathsf{Var} = \frac{\sigma^2}{n}\operatorname{tr}((\lambda I + K)^{-1}K^2(\lambda I + K)^{-1}). \tag{304}$$

Again, a somewhat loose upper bound is $\frac{\sigma^2}{\lambda}(\operatorname{tr}(K)/n)$.

Thus, assuming both $\operatorname{tr}(K)/n$ and $\|f\|_{\mathcal{H}}^2$ are bounded, if we optimize $\lambda$ the sum of the bias and variance can be bounded by $\frac{2\sigma}{\sqrt{n}}\|f\|_{\mathcal{H}}\sqrt{\operatorname{tr}(K)/n}$. One should generally think of $\operatorname{tr}(K)/n$ as being roughly constant as $n$ grows, so this gives a $1/\sqrt{n}$ rate of convergence for the squared error. This is slower than the usual parametric rate of $1/n$ (for the squared error), but that is perhaps not surprising given we are in infinite dimensions. However, more unusual is that the exponent on $n$ does not seem to depend at all on the geometry of $\mathcal{H}$. This is surprising, as we know for instance that it is much easier to estimate smooth functions than functions that are merely Lipschitz, and in fact the exponent in $n$ is different in both cases (as it turns out, $\frac{\log(n)}{n}$ versus $\frac{1}{n^{2/3}}$ for functions on $[0, 1]$).

**Prior on $f$.** The issue is with the "right prior" to place on $f$. From a Bayesian perspective, kernel ridge regression is optimal when $\beta$ is drawn from a Gaussian prior: $\beta \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda}I)$. From a function perspective this corresponds to the Gaussian process prior $f \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda}K)$. In this case the expected norm-squared is $\mathbb{E}[\|\beta\|_2^2] = \frac{\sigma^2}{\lambda}\operatorname{tr}(I) = \frac{\sigma^2}{\lambda}p$. This is a problem when $p = \infty$! A similar problem shows up with $f$: $\mathbb{E}[\|f\|_{\mathcal{H}}^2] = \mathbb{E}[f^{\top}K^{-1}f] = \frac{\sigma^2}{\lambda}\operatorname{tr}(K^{-1}K) = \frac{\sigma^2}{\lambda}n$.

We therefore need a better bound on the bias that doesn't rely on $\|f\|_{\mathcal{H}}^2$, which is unbounded. Returning to (303), under a Gaussian process prior $f \sim \mathcal{N}(0, \rho^2 K)$ we have

$$\mathbb{E}[\text{Bias}^2] = \frac{\lambda^2}{n}\mathbb{E}[f^{*}(\lambda I + K)^{-2}f^*] = \frac{\lambda^2\rho^2}{n}\operatorname{tr}((\lambda I + K)^{-2}K) = \frac{\lambda\rho^2}{n}\sum_{j=1}^{n}\frac{\lambda\mu_j}{(\lambda + \mu_j)^2}, \tag{305}$$

where $\mu_j$ are the eigenvalues of $K$. Similarly the variance is $\frac{\sigma^2}{n}\sum_{j=1}^{n}\frac{\mu_j^2}{(\lambda+\mu_j)^2}$. Thus the overall error is

$$\mathbb{E}[\text{Error}] = \frac{\lambda\rho^2}{n}\sum_{j=1}^{n}\frac{\lambda\mu_j}{(\lambda + \mu_j)^2} + \frac{\sigma^2}{n}\sum_{j=1}^{n}\frac{\mu_j^2}{(\lambda + \mu_j)^2}. \tag{306}$$

We can bound (306) using the inequalities $\frac{\lambda\mu_j}{(\lambda+\mu_j)^2} \leq \min(1, \frac{\mu_j}{\lambda})$ and $\frac{\mu_j^2}{(\lambda+\mu_j)^2} \leq \min(1, \frac{\mu_j^2}{\lambda^2})$. Sort the $\mu_j$ in descending order and let $J = \max\{j \mid \mu_j > \lambda\}$. Then these inequalities yields

$$\mathbb{E}[\text{Error}] \leq \frac{(\lambda\rho^2 + \sigma^2)J}{n} + \frac{\lambda\rho^2}{n}\sum_{j>J}\frac{\mu_j}{\lambda} + \frac{\sigma^2}{n}\sum_{j>J}\frac{\mu_j^2}{\lambda^2}. \tag{307}$$

**Interpretation for power law decay.** Suppose for concreteness that $\mu_j = nj^{-\alpha}$; this power law decay is common empirically and also shows up in theoretical examples such as learning Lipschitz functions. Then $J = (\lambda/n)^{-1/\alpha}$. Moreover, the sums can be approximated as

$$\sum_{j>J}\frac{\mu_j}{\lambda} \approx \frac{1}{\lambda}\int_J^{\infty} x^{-\alpha}dx = \frac{1}{\lambda(\alpha-1)}J^{1-\alpha} = \frac{J}{\alpha-1} = \frac{(\lambda/n)^{-1/\alpha}}{\alpha-1}, \tag{308}$$

$$\sum_{j>J}\frac{\mu_j^2}{\lambda^2} \approx \frac{1}{\lambda^2}\int_J^{\infty} x^{-2\alpha}dx = \frac{1}{\lambda^2(2\alpha-1)}J^{1-2\alpha} = \frac{(\lambda/n)^{-1/\alpha}}{2\alpha-1}. \tag{309}$$

Assuming $\alpha > 1$, the overall expression (307) is thus bounded by

$$\mathbb{E}[\text{Error}(\lambda)] \leq \mathcal{O}\Big(\frac{\alpha}{\alpha-1}(\lambda/n)^{1-1/\alpha}\rho^2 + \frac{\sigma^2(\lambda/n)^{-1/\alpha}}{n}\Big). \tag{310}$$

Now take the optimal $\lambda^* = \frac{\sigma^2}{\rho^2}$. Then we obtain the final bound of

$$\mathbb{E}[\mathrm{Error}(\lambda^*)] \leq \mathcal{O}\Big(\frac{\alpha}{\alpha-1} n^{\frac{1}{\alpha}-1} \big(\sigma^2\big)^{\frac{\alpha-1}{\alpha}} \big(\rho^2\big)^{\frac{1}{\alpha}}\Big). \tag{311}$$

Thus for instance when $\alpha = 2$ we get a $1/\sqrt{n}$ rate, but as $\alpha \to \infty$ we approach a $1/n$ rate.

Note that for sub-optimal choices of regularizer, we still get a bound but with a potentially worse dependence on $n$. For instance if we took $\lambda^* \propto n^\beta$, for $\beta > 0$, then our rate would be suboptimal by a $n^{\beta(\alpha-1)/\alpha}$ factor in $n$ due to the over-regularization.

## 6.4 Generalizing from train to test

So far our error bounds have all been on the original sample points $x_1, \dots, x_n$. But we generally care about how well we do on new points drawn from $p$. We will analyze this next, although in this case it is helpful to switch back to the parameter rather than the function view. Recall that there we had the matrix $S = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$, and we will also define $S^* = \mathbb{E}_{x \sim p}[xx^\top]$. Then letting $L^*$ be the population loss, we have $L^*(\beta) - L^*(\beta^*) = (\beta - \beta^*)^\top S^* (\beta - \beta^*)$.

The ridge regression estimator $\hat{\beta}_\lambda$ is still equal to $\frac{1}{n}(\frac{\lambda}{n}I + S)^{-1}X^\top(X\beta^* + \epsilon)$, so $\mathbb{E}_\epsilon[\hat{\beta}_\lambda] = (\frac{\lambda}{n}I + S)^{-1}S\beta^*$. Define the notation $\lambda_n = \lambda/n$. The bias is $\delta^\top S^* \delta$, where $\delta = \mathbb{E}[\hat{\beta}_\lambda] - \beta^*$, which works out to

$$\mathrm{Bias}^2 = \lambda_n^2 (\beta^*)^\top (\lambda_n I + S)^{-1} S^* (\lambda_n I + S)^{-1} \beta^*, \tag{312}$$

which under the Gaussian process prior is

$$\mathbb{E}[\mathrm{Bias}^2] = \lambda_n^2 \rho^2 \operatorname{tr}((\lambda_n I + S)^{-1} S^* (\lambda_n I + S)^{-1}). \tag{313}$$

Now, note that this is *almost* the expression we would get from running ridge regression directly on the test distribution–the only difference is that $(\lambda_n I + S^*)^{-1}$ is replaced with $(\lambda_n I + S)^{-1}$ in two places.

The variance can be computed similarly as

$$\begin{aligned} \mathsf{Var} &= \frac{1}{n^2}\mathbb{E}_\epsilon[\|(S^*)^{1/2}(\lambda_n I + S)^{-1}X^\top \epsilon\|_2^2] & (314)\\ &= \frac{\sigma^2}{n}\operatorname{tr}((S^*)^{1/2}(\lambda_n I + S)^{-1} S (\lambda_n I + S)^{-1}(S^*)^{1/2}). & (315) \end{aligned}$$

Here, the expression is trickier because we have both $S$ and $(\lambda_n I + S)^{-1}$. But since $S(\lambda_n I + S)^{-1} \preceq I$, we can apply the further bound

$$\mathsf{Var} \leq \frac{\sigma^2}{n}\operatorname{tr}((S^*)^{1/2}(\lambda_n I + S)^{-1}(S^*)^{1/2}). \tag{316}$$

Relative to this bound, again the only difference is replacing $(\lambda_n I + S^*)^{-1}$ with $(\lambda_n I + S)^{-1}$. Using these results, we can obtain the following generalization bound:

**Proposition 6.4.** *Suppose that the true and empirical covariance matrices satisfy $S^* \preceq 2S + \lambda_n I$. Then, letting $\mu_j^*$ denote the $j$th eigenvalues of $S^*$, the generalization error satisfies*

$$\begin{aligned} \mathbb{E}_\epsilon[L^*(\hat{\beta}_\lambda)] - L^*(\beta^*) &\leq 4\lambda_n\rho^2 \sum_j \frac{\lambda_n \mu_j^*}{(\lambda_n + \mu_j^*)^2} + \frac{2\sigma^2}{n}\sum_j \frac{\mu_j^*}{\lambda_n + \mu_j^*} & (317)\\ &\leq \Big(4\lambda_n\rho^2 + \frac{2\sigma^2}{n}\Big)\sum_j \min\big(1, \frac{\mu_j^*}{\lambda_n}\big). & (318) \end{aligned}$$

Note that relative to the kernel case, $\mu_j^*$ is $n$ times smaller, since it is an eigenvalue of $S$ rather than $K$.

This expression is similar to what we achieved in (307), and when $\mu_j^* = j^{-\alpha}$ setting $\lambda_n = \frac{\sigma^2}{n\rho^2}$ will yield the same power law rate as before of $n^{\frac{1}{\alpha}-1}$.

The key question is for what values of $\lambda_n$ the condition $S^* \preceq 2S + \lambda_n I$ is satisfied. This requires the empirical covariance matrix to concentrate to the population covariance. To analyze this, we will make use of the following result of Koltchinskii and Lounici (2017):

**Theorem 6.5** (Koltchinskii and Lounici)**.** *If the $x_i$ are Gaussian with covariance $\Sigma^*$, then the empirical covariance matrix $\hat{\Sigma}$ satisfies*

$$\|\Sigma^* - \hat{\Sigma}\| \lesssim \max\Big(\frac{\operatorname{tr}(\Sigma^*)}{n}, \sqrt{\frac{\operatorname{tr}(\Sigma^*)\|\Sigma^*\|}{n}}\Big) \tag{319}$$

*with high probability. In particular, if $S^*$ is $p \times p$ and $p \leq n$ then we have $\|S^* - S\| \leq \|S^*\|\sqrt{p/n}$ with high probability.*

Recall that we want to show that $2S + \lambda_n I \succeq S^*$, or $2S \succeq (S^* + \lambda_n I) - 2\lambda_n I$. By re-arrangement, this is equivalent to

$$(S^* + \lambda_n I)^{-1/2} S (S^* + \lambda_n I)^{-1/2} \succeq \frac{1}{2}I - \lambda_n (S^* + \lambda_n I)^{-1}. \tag{320}$$

But the left-hand-side is the empirical covariance matrix of a Gaussian with covariance $\Sigma^* = S^*(S^* + \lambda_n I)^{-1}$. We will apply Theorem 6.5 to this matrix, i.e. $\hat{\Sigma} = (S^* + \lambda_n I)^{-1/2} S (S^* + \lambda_n I)^{-1/2}$.

We have $\|\Sigma^*\| \leq 1$, and $\operatorname{tr}(\Sigma) = \sum_j \frac{\mu_j^*}{\mu_j^* + \lambda_n}$. Therefore, $\|\hat{\Sigma} - \Sigma^*\| \leq \frac{1}{2}$ as long as $n \gg \sum_j \frac{\mu_j^*}{\mu_j^* + \lambda_n}$. In this case, we have

$$\hat{\Sigma} = (S^* = \lambda_n I)^{-1/2} S (S^* + \lambda_n I)^{-1/2} \tag{321}$$

$$\succeq S^*(S^* + \lambda_n I)^{-1} - \frac{1}{2}I \tag{322}$$

$$= \frac{1}{2}I + (S^*(S^* + \lambda_n I)^{-1} - I) \tag{323}$$

$$= \frac{1}{2}I - \lambda_n (S^* + \lambda_n I)^{-1}, \tag{324}$$

as was to be shown. This yields the result:

**Corollary 6.6.** *Suppose $\lambda_n$ is such that $n \geq C \sum_j \frac{\mu_j^*}{\mu_j^* + \lambda_n}$, for some universal constant $C$. Then the generalization error of ridge regression satisfies*

$$\mathbb{E}_\epsilon[L^*(\hat{\beta}_\lambda)] - L^*(\beta^*) \leq \Big(4\lambda_n \rho^2 + \frac{2\sigma^2}{n}\Big) \sum_j \min\big(1, \frac{\mu_j^*}{\lambda_n}\big). \tag{325}$$

Consider again the case of power law decay, with $\mu_k^* = j^{-\alpha}$. We saw before that by taking the optimal value $\lambda_n = \frac{\sigma^2}{n\rho^2}$, we achieve generalization error $\mathcal{O}(\frac{\alpha}{\alpha-1} n^{\frac{1}{\alpha}-1} (\sigma^2)^{\frac{\alpha-1}{\alpha}} (\rho^2)^{\frac{1}{\alpha}})$. However, we also must have $\sum_j \min(1, \frac{\mu_j^*}{\lambda_n}) \ll n$, which reduces to $\frac{\alpha}{\alpha-1}\lambda_n^{-1/\alpha} \ll n$. Thus $\lambda_n \gg \big(\frac{\alpha}{\alpha-1}\big)^\alpha n^{-\alpha} \approx \frac{\alpha}{\alpha-1} n^{-\alpha}$. Since $\alpha > 1$, this usually does not matter except when $\sigma$ is very small. In this case, since $\lambda_n$ is forced to be larger than optimal, the $4\lambda_n \rho^2$ term dominates and we instead achieve error that is on the order of

$$\lambda_n \rho^2 \sum_j \min(1, \mu_j^*/\lambda_n) \approx \lambda_n \rho^2 \cdot \frac{\alpha}{\alpha - 1} \lambda_n^{-1/\alpha} \tag{326}$$

$$\approx \frac{\alpha}{\alpha-1} \Big(\frac{\alpha}{\alpha-1} n^{-\alpha}\Big)^{1-1/\alpha} \rho^2 \tag{327}$$

$$\approx \frac{\alpha}{\alpha-1} n^{1-\alpha} \rho^2. \tag{328}$$

Therefore, a general upper bound for the generalization error (assuming the best choice of lambda) is

$$\mathcal{O}\Big(\frac{\alpha}{\alpha-1} \max\big(n^{\frac{1}{\alpha}-1} (\sigma^2)^{\frac{\alpha-1}{\alpha}} (\rho^2)^{\frac{1}{\alpha}}, n^{1-\alpha}\rho^2\big)\Big). \tag{329}$$

Thus for instance when $\alpha = 2$, if $\sigma$ is large we achieve a rate $n^{-1/2}$ but as $\sigma \to 0$ the rate improves to $n^{-1}$.

Finally, the conditions in Theorem 6.5 can be relaxed somewhat, as discussed in Koltchinskii and Lounici (2017). In particular, we don't need perfect Gaussianity and can get by with appropriate sub-Gaussian assumptions.

[Lecture 21]

## 6.5 Eigenvalues and Mercer's theorem

It is often helpful to represent the above computations in terms of the eigenbasis of $k$. Since $k$ is infinite-dimensional, we need some conditions to ensure that it actually has an eigenbasis. These are given by Mercer's theorem:

**Theorem 6.7** (Mercer)**.** *Suppose that $k : \mathcal{X} \times \mathcal{X}$ is a positive definite kernel, that $k$ is continuous, and that $\mathcal{X}$ is compact and equipped with a finite measure $\nu$ that is supported on all of $\mathcal{X}$. Define*

$$T_k f(x) = \int_{\mathcal{X}} k(x, s) f(s) d\nu(s). \tag{330}$$

*Then there is an orthonormal basis of eigenfunctions of $T_k$, $e_1, e_2, \ldots$, with corresponding eigenvalues $\mu_1, \mu_2, \ldots$, such that*

$$k(x, x') = \sum_{m=1}^{\infty} \mu_m e_m(x) e_m(x'). \tag{331}$$

Note that the eigenfunctions will be different depending on the choice of $\nu$, and orthogonality is with respect to $\nu$: $\int e_l(x) e_m(x) d\nu(x) = \delta_{lm}$. There are thus three distinct inner products floating around: the one underlying the norm $\|f\|_{\mathcal{H}}$, the one underlying the function error $\|f - y\|_2$, and the one associated with $\nu$. This can get a bit confusing, but it helps to think of $\|f\|_{\mathcal{H}}$ and $\|f\|_2$ as the "important" norms, while $\nu$ specifies a non-intrinsic norm that is however useful for computational purposes. (We will typically take $\nu$ either to be $p$, so that $\|f\|_{L^2(\nu)} = \mathbb{E}_{x \sim p}[f(x)^2]^{1/2}$ measures generalization error, or pick a $\nu$ for which it is particularly easy to compute the eigenfunctions of $T_k$.)

Speaking of these computational purposes, let us see what Mercer's theorem buys us. Any function $f$ can now be represented as $f(x) = \sum_{m=1}^{\infty} c_m e_m(x)$, where $c_m = \int_{\mathcal{X}} f(x) e_m(x) d\nu(x)$. Under this representation, it turns out that $\|f\|_{\mathcal{H}}^2 = \sum_{m=1}^{\infty} \frac{c_m^2}{\mu_m}$. This might seem surprising, since the right-hand side depends on $\nu$ but the left-hand side does not. We explain this with the following result:

**Lemma 6.8.** *Let $\nu$ be any finite measure that is supported on all of $\mathcal{X}$. Then $\|f\|_{\mathcal{H}}^2 = \sum_{m=1}^{\infty} \frac{1}{\mu_m} (\int_{\mathcal{X}} f(x) e_m(x) d\nu(x))^2$.*

### Proving Lemma 6.8

*Heuristic proof of Lemma 6.8.* Consider any finite partition of $\mathcal{X}$ into sets $X_1, \ldots, X_n$, with a representative $x_i \in X_i$ for each set. We will take increasingly fine-grained partitions such that $\sum_{i=1}^n \int_{X_i} |f(x') - f(x_i)| d\nu(x') \to 0$, so in particular e.g. $\sum_{i=1}^n f(x_i) \nu(X_i) \to \int_{\mathcal{X}} f(x) d\nu(x)$.

Now we define $(f_\nu)_i = \nu(X_i)^{1/2} f(x_i)$ and $(K_\nu)_{ij} = \nu(X_i)^{1/2} \nu(X_j)^{1/2} k(x_i, x_j)$. We can check that for $T = \{x_1, \ldots, x_n\}$, $f^\top K^{-1} f = f_\nu^\top K_\nu^{-1} f_\nu$, independent of the choice of $\nu$. This is the first hint that the choice of $\nu$ doesn't really matter. Now note that

$$f_\nu^\top K_\nu^{-1} f_\nu = f_\nu^\top \Big( \sum_{m=1}^{\infty} \mu_m \begin{bmatrix} \nu(X_1)^{1/2} e_m(x_1) \\ \cdots \\ \nu(X_n)^{1/2} e_m(x_n) \end{bmatrix} \begin{bmatrix} \nu(X_1)^{1/2} e_m(x_1) \\ \cdots \\ \nu(X_n)^{1/2} e_m(x_n) \end{bmatrix}^\top \Big)^{-1} f_\nu \tag{332}$$

$$\overset{(i)}{\approx} f_\nu^\top \Big( \sum_{m=1}^{\infty} \frac{1}{\mu_m} \begin{bmatrix} \nu(X_1)^{1/2} e_m(x_1) \\ \cdots \\ \nu(X_n)^{1/2} e_m(x_n) \end{bmatrix} \begin{bmatrix} \nu(X_1)^{1/2} e_m(x_1) \\ \cdots \\ \nu(X_n)^{1/2} e_m(x_n) \end{bmatrix}^\top \Big) f_\nu \tag{333}$$

$$= \sum_{m=1}^{\infty} \frac{1}{\mu_m} \Big( \sum_{i=1}^{n} \nu(X_i) f(x_i) e_m(x_i) \Big)^2 \tag{334}$$

$$\overset{(ii)}{\approx} \sum_{m=1}^{\infty} \frac{1}{\mu_m} (\int_{\mathcal{X}} f(x) e_m(x) d\nu(x))^2. \tag{335}$$

Here (i) is the main "heuristic" step, where we are inverting the matrix on the premise that the different vectors $[\nu(X_i) e_m(x_i)]$ are nearly orthogonal, since $\langle [\nu(X_i)^{1/2} e_l(x_i)], [\nu(X_i)^{1/2} e_m(x_i)] \rangle = \sum_i \nu(X_i) e_l(x_i) e_m(x_i) \approx \int_X e_l(x) e_m(x) d\nu(x) = 0$. The step (ii) also needs to be made rigorous but is essentially just taking a limit.

Now, we by definition have $\|f\|_{\mathcal{H}}^2 \geq f^\top K^{-1} f = f_\nu^\top K_\nu^{-1} f_\nu$, which by the above can be made arbitrarily close to $A = \sum_{m=1}^\infty \frac{1}{\mu_m} (\int_{\mathcal{X}} f(x) e_m(x) d\nu(x))^2$ by taking sufficiently fine partitions of $\{X_i\}$. This proves one direction of the lemma, i.e. that $A \leq \|f\|_{\mathcal{H}}^2$. To prove the other direction, note that for any set $T = \{\hat{x}_1, \dots, \hat{x}_n\}$, we can build our partitions such that they contain the $\hat{x}_i$ as a subset of the representatives $x_i$ in the partition. In this case, $f[T]^\top K[T]^{-1} f[T] \leq f^\top K^{-1} f = f_\nu^\top K_\nu^{-1} f_\nu$ by Lemma 6.1. Taking the limit then yields $f[T]^\top K[T]^{-1} f[T] \leq A$. Since this holds for all $T$, we thus have $\|f\|_{\mathcal{H}}^2 \leq A$, which proves the other direction and completes the lemma. □

### Eigenvalue Decay for Smooth Kernels

We can use Mercer's theorem to obtain fairly general results on the eigenvalue decay of a kernel, assuming that the kernel satisfies certain smoothness properties. In particular, suppose that the underlying space $\mathcal{X}$ is $d$-dimensional, and that the kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is $t$ times differentiable. Then the eigenvalues of $T_k$ decay at a rate $\mathcal{O}(n^{-t/d-1})$, essentially independently of the measure $\mu$.

**Theorem 6.9** (Theorem 4 of Kühn (1987))**.** *Let $M$ be a compact $d$-dimensional smooth manifold with a finite Lesbegue-type measure $\mu$, and let $0 < t < \infty$. Then for every positive definite kernel $k$ with continuous $t$-th derivative, we have $\mu_m(T_k) = \mathcal{O}(m^{-t/d-1})$.*

In particular, we can take $\mu$ to be $p^*$. Then (combining with our previous results) we have that the eigenvalues of $S$ decay with power law exponent $\alpha = t/d + 1$, and thus the generalization error decays as $n^{-t/(t+d)}$ when $\sigma$ is large, and $n^{-t/d}$ when $\sigma$ is small.

We note these are slightly different than classical rates that may be found in the literature for smooth kernels, as in some classical settings it is assumed that a function deterministically lies within an RKHS, rather than being sampled from a Gaussian process with kernel $k$. (Recall that samples from a Gaussian process almost surely *do not* lie within the corresponding RKHS.)

## 6.6 Random Features: Applying Mercer's Theorem

Mercer's theorem allows us to draw a particularly tight connection between random features and kernel regression. To understand this connection, first consider a random feature distribution where $\phi(x) = [\phi_1(x); \cdots ; \phi_m(x)]/\sqrt{m}$ with $\phi_j \sim P$ (here $P$ is some probability distribution over functions $f : \mathcal{X} \to \mathbb{R}$).

As $m \to \infty$, the inner product $\langle \phi(x), \phi(y) \rangle$ converges to $\mathbb{E}_{\phi \sim P}[\phi(x)\phi(y)]$. Therefore, the random feature function approximates the kernel $k(x, y) = \mathbb{E}_\phi[\phi(x)\phi(y)]$.

On the other hand, Mercer's theorem lets us go in the opposite direction: given any kernel function, we can construct a corresponding random feature map. To see this, suppose that our kernel $k$ has the eigendecomposition

$$k(x, y) = \sum_{l=1}^\infty \mu_l e_l(x) e_l(y). \tag{336}$$

Then letting $Z = \sum_l \mu_l$, our measure $P$ samples the function $\sqrt{Z} \cdot e_l$ with probability $\mu_l/Z$. We can easily verify that

$$\begin{aligned} \mathbb{E}_\phi[\phi(x)\phi(y)] &= \sum_{l=1}^\infty \frac{\mu_l}{Z} \cdot \sqrt{Z} e_l(x) \cdot \sqrt{Z} e_l(y) && (337) \\ &= \sum_{l=1}^\infty \mu_l e_l(x) e_l(y) && (338) \\ &= k(x, y). && (339) \end{aligned}$$

The random features approach is particularly useful computationally, as kernel regression can be quite expensive (it involves inverting an $n \times n$ matrix), while with random features we reduce the problem to a standard regression task.

Random features are also useful analytically, as neural networks in many ways behave like random features models, allowing parallels to be drawn between neural networks and kernel regression. One such modern analogy is the *neural tangent kernel*.

### Examples of Random Feature Maps

To make full use of random features, we will need a cousin of Mercer's theorem, called Bochner's theorem. The main difference is that Mercer's theorem requires everything to happen on a compact set and yields an eigenbasis that is at most countable in size. Bochner's theorem instead requires the kernel to be "shift-invariant", can handle non-compact sets, but the eigenbasis could be uncountably large:

**Theorem 6.10** (Bochner)**.** *Call a kernel shift-invariant if $k(x, y) = k(x - y)$ for some single-argument function $k(\cdot)$. Then $k$ is the Fourier transform of some positive measure, i.e.*

$$k(x, y) = \int e^{-i\langle \omega, x-y\rangle} d\mu(\omega) \tag{340}$$

*for some positive measure $\mu$.*

For instance, consider Gaussian kernel, $k(x, y) = \exp(-\|x - y\|_2^2/2\sigma^2)$. Thus $k(\Delta) = \exp(-\Delta^2/2\sigma^2)$. The Fourier transform is again a Gaussian–ignoring the normalization constant, it is $\hat{k}(\omega) = \exp(-\sigma^2/\omega^2/2)$. Thus the Gaussian kernel has the representation It has the representation (again up to the normalization constant)

$$k(x, y) \propto \int e^{-i\langle \omega, x-y\rangle} \exp(-\sigma^2\|\omega\|_2^2/2) d\omega \tag{341}$$

$$\propto \int \cos(\langle \omega, x - y\rangle) \exp(-\sigma^2\|\omega\|_2^2/2). \tag{342}$$

We can therefore construct a random feature map by sampling $\omega$ from a Gaussian with covariance $\frac{1}{\sigma^2}I$, and using the feature function $x \mapsto \exp(i\langle \omega, x\rangle)$. If we wish to avoid complex numbers, it turns out we can also use $x \mapsto \cos(\langle \omega, x\rangle + b)$, where $b$ is drawn at random from $[0, 2\pi)$. This is because

$$\int_0^{2\pi} \cos(x + b) \cos(y + b) db = \frac{1}{4}(2b \cos(x - y) + \sin(x + y + 2b))\Big|_0^{2\pi} = \pi \cos(x - y). \tag{343}$$

Similar results hold for other kernels. For instance, if $k(x, y) = \exp(-\|x-y\|_1/\lambda)$, then $k(\Delta) = \exp(-\|\Delta\|_1/\lambda)$, and the Fourier transform $\hat{k}(\omega)$ is, up to normalization constants, equal to $\prod_{j=1}^d \frac{1}{1+\lambda^2\omega_j^2}$. Here instead of sampling $\omega$ from a Gaussian distribution, we would construct random features by sampling $\omega$ from a Cauchy distribution.

These ideas are discussed in more detail in Rahimi et al. (2007).

### Neural Tangent Kernel

We claimed above that neural networks are somewhat similar to random features models. There are a number of attempts to make this precise. The most common is via the neural tangent kernel, where we can locally approximate a neural network (or any parameterized function) by a linear kernel. Specifically, suppose for simplicity that we have a parameterized function $f_\theta : \mathcal{X} \to \mathbb{R}$ (if we wanted a classifier we could take $\mathrm{sign}(f_\theta(x))$, or for $k$-class classification consider $k$ distinct $f_\theta$ and take the argmax).

Then, the tangent kernel is simply

$$k(x, y; \theta) = \langle \nabla f_\theta(x), \nabla f_\theta(y)\rangle \tag{344}$$

For a neural network, this will roughly be a sum over all the edges in the network of the derivative for that edge. If the number of edges is very large (larger than the number of data points), then $k$ effectively acts as if it is infinite-dimensional.

A number of papers starting with Jacot et al. (2018) have studied the case where the width is infinitely large, and the learning rate of SGD is small. In this case the neural network actually does converge to a random features model, with corresponding kernel $k$ as given above (where $k$ is evaluated at the random initial point $\theta = \theta_0$).

On the other hand, this regime is not necessarily realistic, as learning rates used in practice are larger than required by these infinite-width results. Thus in practice the tangent kernel actually varies over the course of training. Some recent work, such as Lewkowycz et al. (2020), tries to characterize this evolution.

[Lecture 22]

# 7 Domain Adaptation under Covariate Shift

We now shift focus again, to a type of perturbation called *covariate shift*. We work in a classification or regression setting where we wish to predict $y$ from $x$, and make the assumption that $\tilde{p}(y \mid x)$ and $p^*(y \mid x)$ are the same (the labeling function doesn't change between train and test):

**Assumption 7.1** (Covariate Shift)**.** *For a train distribution $\tilde{p}$ and test distribution $p^*$, we assume that $\tilde{p}(y \mid x) = p^*(y \mid x)$ for all $x$.*

Thus the only thing that changes between train and test is the distribution of the covariates $x$ (hence the name covariate shift). We furthermore assume that we observe labeled samples $(x_1, y_1), \dots, (x_n, y_n) \sim \tilde{p}$, together with *unlabeled* samples $\bar{x}_1, \dots, \bar{x}_m \sim p^*$. In the language of our previous setting, we could say that $D(\tilde{p}, p^*) = \|\tilde{p}(y \mid x) - p^*(y \mid x)\|_\infty$, $\epsilon = 0$, and $\mathcal{G} = \{p \mid p(x) = p_0(x)\}$ for some distribution $p_0$ (obtained via the unlabeled samples from $p^*$).

Beyond covariate shift, we will need to make some additional assumption, since if $\tilde{p}(x)$ and $p^*(x)$ have disjoint supports then the assumption that $\tilde{p}(y \mid x) = p^*(y \mid x)$ is meaningless. We will explore two different assumptions:

1. Either we assume the $\tilde{p}(x)$ and $p^*(x)$ are known and not too different from each other, or

2. We assume that the model family is realizable: $p^*(y \mid x) = p_\theta(y \mid x)$ for some $\theta$.

This will lead to two different techniques: importance weighting and uncertainty estimation. We will also see how to construct a "doubly robust" estimator that works as long as at least one of the assumptions holds.

## 7.1 Importance weighting

First assume that $\tilde{p}(x)$ and $p^*(x)$ are known. (We can generally at least attempt to estimate them from unlabeled data, although if our model family is misspecified then our estimates might be poor.)

In a traditional setting, to minimize the loss on $\tilde{p}(x)$ we would minimize

$$\mathbb{E}_{(x,y)\sim\tilde{p}}[\ell(\theta; x, y)], \tag{345}$$

where $\ell$ is the loss function for either classification or regression. We can approximate this via the samples from $\tilde{p}$ as

$$\frac{1}{n}\sum_{i=1}^{n} \ell(\theta; x_i, y_i). \tag{346}$$

To handle covariate shift we would like to instead minimize the expectation over $p^*$, but unfortunately we can't do this because we don't have any outputs $y$ drawn from $p^*$. The key insight that lets us get around this is the following identity:

$$\mathbb{E}_{(x,y)\sim p^*}\Big[\ell(\theta; x, y)\Big] = \mathbb{E}_{(x,y)\sim\tilde{p}}\Big[\frac{p^*(x)}{\tilde{p}(x)}\ell(\theta; x, y)\Big]. \tag{347}$$

Taking this identity as given for the moment, we can then approximate the expectation over $p^*$ via *samples from $\tilde{p}$* as follows:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{p^*(x_i)}{\tilde{p}(x_i)}\ell(\theta; x_i, y_i). \tag{348}$$

This quantity is called the *propensity-weighted training loss*[7], because each training sample is weighted by how much more it looks like a sample from $p^*$ than from $\tilde{p}$.

[7] Also sometimes called the importance-weighted loss.

To prove the identity, we make use of the covariate shift assumption:

$$\begin{aligned}\mathbb{E}_{(x,y)\sim p^*}[\ell(\theta;x,y)] &= \mathbb{E}_{(x,y)\sim\tilde{p}}\Big[\frac{p^*(x,y)}{\tilde{p}(x,y)}\ell(\theta;x,y)\Big] && (349)\\ &= \mathbb{E}_{(x,y)\sim\tilde{p}}\Big[\frac{p^*(x)}{\tilde{p}(x)}\frac{p^*(y\mid x)}{\tilde{p}(y\mid x)}\ell(\theta;x,y)\Big] && (350)\\ &= \mathbb{E}_{(x,y)\sim\tilde{p}}\Big[\frac{p^*(x)}{\tilde{p}(x)}\ell(\theta;x,y)\Big], && (351)\end{aligned}$$

where the final equality is by the covariate shift assumption.

**Variance of the estimator.** Even if $\frac{p^*(x)}{\tilde{p}(x)}$ can be computed, the importance weighted estimator could have high variance. This is because the weights $\frac{p^*(x_i)}{\tilde{p}(x_i)}$ could be large or potentially infinite.

For convenience assume that $\ell(\theta;x,y)\le B$ for all $\theta,x,y$. We can compute (or rather, bound) the variance as follows:

$$\begin{aligned}\mathsf{Var}_{\tilde{p}}[\frac{p^*(x)}{\tilde{p}(x)}\ell(\theta;x,y)] &= \mathbb{E}_{\tilde{p}}[(p^*(x)/\tilde{p}(x))^2\ell(\theta;x,y)^2] - \mathbb{E}_{p^*}[\ell(\theta;x,y)]^2 && (352)\\ &\le \mathbb{E}_{\tilde{p}}[(p^*(x)/\tilde{p}(x))^2]B^2 && (353)\\ &= (D_{\chi^2}(\tilde{p}\|p^*)+1)B^2, && (354)\end{aligned}$$

where $D_{\chi^2}$ is the $\chi^2$-divergence:

$$D_{\chi^2}(\tilde{p}\|p^*) \stackrel{\mathrm{def}}{=} \int \frac{(p^*(x)-\tilde{p}(x))^2}{\tilde{p}(x)}dx = \int \frac{p^*(x)^2}{\tilde{p}(x)}dx - 1. \qquad (355)$$

The variance of the propensity-weighted loss is thus more or less controlled by the $\chi^2$-divergence between source and target. To gain some intuition for how $\chi^2$ behaves, first note that is is always larger than KL divergence (in the reverse direction, though I'm not sure the order of arguments is canonical):

$$\begin{aligned}D_{\mathrm{kl}}(p^*\|\tilde{p}) &= \int p^*(x)\log\frac{p^*(x)}{\tilde{p}(x)}dx && (356)\\ &\le \int p^*(x)\frac{p^*(x)-\tilde{p}(x)}{\tilde{p}(x)}dx && (357)\\ &= \int \frac{p^*(x)^2}{\tilde{p}(x)}dx - 1 = D_{\chi^2}(\tilde{p}\|p^*). && (358)\end{aligned}$$

Additionally, the $\chi^2$-divergence between two Gaussians is exponential in the difference between their means. To see this, let $Z$ denote the normalization constant of an isotropic Gaussian, and write

$$\begin{aligned}D_{\chi^2}(\mathcal{N}(\mu,I),\mathcal{N}(\mu',I)) &= -1+\frac{1}{Z}\int\exp(\frac{1}{2}\|x-\mu'\|_2^2-\|x-\mu\|_2^2)dx && (359)\\ &= -1+\frac{1}{Z}\int\exp(\frac{1}{2}(-\|x\|_2^2-(2\mu'-4\mu)^\top x+\|\mu'\|_2^2-2\|\mu\|_2^2)) && (360)\\ &= -1+\frac{1}{Z}\int\exp(\frac{1}{2}(-\|x+(\mu'-2\mu)\|_2^2+\|\mu'-2\mu\|_2^2+\|\mu'\|_2^2-2\|\mu\|_2^2)) && (361)\\ &= -1+\exp(\|\mu'\|_2^2+\|\mu\|_2^2-2\mu^\top\mu') = -1+\exp(\|\mu-\mu'\|_2^2). && (362)\end{aligned}$$

This is bad news for propensity weighting, since the weights blow up exponentially as the distributions move apart.

**Connection to causal inference.** Propensity weighting can also be used in the context of causal inference. Here we have a patient with covariates $X$, with treatment condition $T$ (usually $T \in \{0,1\}$), and outcome $Y$. Our goal is to estimate the treatment effect, which, roughly speaking, is $\mathbb{E}[Y \mid T=1] - \mathbb{E}[Y \mid T=0]$ (this is wrong as stated and will be remedied below). We will see below how to do this by letting $p_0^*$ and $p_1^*$ be the distributions where $T=0$ and $T=1$, respectively. However, first we need to set up the problem more carefully.

To set the problem up more carefully, we use the *potential outcomes framework*. In this framework there are actually two variables, $Y(0)$ and $Y(1)$, which are what the outcome *would have been* if we had set $T=0$ or $T=1$, respectively. This is potentially different from the distribution of the outcome conditional on $T$, since there could be factors that correlate $T$ with $Y$ (for instance, if $T$ is smoking and $Y$ is lung cancer, there could be some gene that causes one to both be more likely to smoke and more likely to get lung cancer that accounts for the strong empirical correlation between $T$ and $Y$; this was an actual objection raised by Fisher!).

Of course, there are plenty of factors that create correlation between $T$ and $Y$ in an observational setting, for instance sicker patients are more likely to be treated aggressively. We are okay with this as long as these factors are observed as part of the covariates $X$. This leads us to the *unconfoundedness assumption*:

**Assumption 7.2** (Unconfoundedness)**.** *The distribution* $(X, T, Y(0), Y(1))$ *is said to be* unconfounded *if* $Y(0), Y(1) \perp\!\!\!\perp T \mid X$. *In other words, treatment and outcome should be independent conditional on the covariates* $X$.

The main challenge in the potential outcomes framework is that we only observe $(X, T, Y(T))$. In other words, we only observe the outcome for the treatment $T$ that was actually applied, which makes it difficult to estimate $\mathbb{E}[Y(1)]$ or $\mathbb{E}[Y(0)]$. We will deal with this by treating estimating $\mathbb{E}[Y(1)]$ as a domain adaptation problem, and using propensity weighting. First note that, by unconfoundedness, we have

$$\mathbb{E}_{\tilde{p}}[Y(1)] = \mathbb{E}_{X\sim\tilde{p}}[\mathbb{E}_{\tilde{p}}[Y(1) \mid X]] \tag{363}$$

$$= \mathbb{E}_{X\sim\tilde{p}}[\mathbb{E}_{\tilde{p}}[Y(1) \mid X, T=1]] \tag{364}$$

$$= \mathbb{E}_{p_1^*}[Y(T)], \tag{365}$$

where we define $p_1^*$ such that $p_1^*(x,t,y) = \tilde{p}(x)\mathbb{I}[t=1]\tilde{p}(y \mid x, t=1)$; this has the same distribution over $x$ as $\tilde{p}$, but the treatment $t=1$ is always applied. Since $\tilde{p}(y \mid x,t) = p^*(y \mid x,t)$ almost surely, the covariate shift assumption holds. We can thus estimate the expectation under $p_1^*$ via propensity weighting:

$$\mathbb{E}_{p_1^*}[Y(T)] = \mathbb{E}_{\tilde{p}}\Big[\frac{p_1^*(X,T)}{\tilde{p}(X,T)}Y(T)\Big] \tag{366}$$

$$= \mathbb{E}_{\tilde{p}}\Big[\frac{p_1^*(T \mid X)}{\tilde{p}(T \mid X)}Y(T)\Big] \tag{367}$$

$$= \mathbb{E}_{\tilde{p}}\Big[\frac{\mathbb{I}[T=1]}{\tilde{p}(T \mid X)}Y(T)\Big]. \tag{368}$$

A similar calculation holds for computing $\mathbb{E}_{\tilde{p}}[Y(0)]$, for the distribution $p_0^*(x,t,y) = \tilde{p}(x)\mathbb{I}[t=0]\tilde{p}(y \mid x, t=0)$. Together, we have that

$$\mathbb{E}_{\tilde{p}}[Y(1) - Y(0)] = \mathbb{E}_{\tilde{p}}\Big[\Big(\frac{\mathbb{I}[T=1]}{\tilde{p}(T \mid X)} - \frac{\mathbb{I}[T=0]}{\tilde{p}(T \mid X)}\Big)Y(T)\Big]. \tag{369}$$

Since the right-hand-side is in terms of $Y(T)$, it only involves observable quantities, and can be estimated from samples as long as $\tilde{p}(T \mid X)$ is known. This estimator is called *inverse propensity weighting* because it involves dividing by the propensity weights $\tilde{p}(T \mid X)$.

In the next section, we will explore an improvement on inverse propensity weighting called a *doubly-robust estimator.*

[Lecture 23]

## 7.2 Doubly-Robust Estimators

Recall that in the previous section we defined the inverse propensity weighted estimator

$$\mathbb{E}_{\tilde{p}}[Y(1) - Y(0)] = \mathbb{E}_{\tilde{p}}\Big[\Big(\frac{\mathbb{I}[T = 1]}{\tilde{p}(T \mid X)} - \frac{\mathbb{I}[T = 0]}{\tilde{p}(T \mid X)}\Big)Y(T)\Big]. \tag{370}$$

To actually estimate the left-hand-side, we take the empirical average over $n$ samples.

There are a couple of downsides of this estimator. One is that the variance of this estimator can be large. Specifically, we can compute it as

$$\frac{1}{n}\Big(\mathbb{E}_{\tilde{p}}\Big[\frac{1}{\tilde{p}(T = 1 \mid X)}Y(1)^2 + \frac{1}{\tilde{p}(T = 0 \mid X)}Y(0)^2\Big] - \mathbb{E}_{\tilde{p}}[Y(1) - Y(0)]^2\Big). \tag{371}$$

If the propensity weights are near zero then the variance explodes (similarly to the issue with $\chi^2$-divergence that we saw earlier).

Another issue is that estimating the propensity weights themselves is non-trivial, and if we use the wrong propensity weights, then the estimate could be arbitrarily wrong.

We will explore an idea that partially mitigates both issues; it reduces the variance when the propensity weights are correct (although doesn't avoid the exploding variance issue), and in some cases it produces a correct estimate even if the propensity weights are wrong.

The basic idea is as follows: suppose that we have some prediction $\bar{Y}(1, X)$, $\bar{Y}(0, X)$ of what will happen under $T = 1$, $T = 0$ conditioned on $X$. Since these predictions only require knowing $X$ and not $T$, an alternate estimate of the treatment effect can be obtained by adding and subtracting $\bar{Y}$:

$$\begin{aligned}\mathbb{E}_{\tilde{p}}[Y(1) - Y(0)] &= \mathbb{E}_{\tilde{p}}[\bar{Y}(1, X) - \bar{Y}(0, X)] + \mathbb{E}_{\tilde{p}}[(Y(1) - \bar{Y}(1, X)) - (Y(0) - \bar{Y}(0, X))] &(372)\\ &= \mathbb{E}_{\tilde{p}}[\bar{Y}(1, X) - \bar{Y}(0, X)] + \mathbb{E}_{\tilde{p}}\Big[\Big(\frac{\mathbb{I}[T = 1]}{\tilde{p}(T \mid X)} - \frac{\mathbb{I}[T = 0]}{\tilde{p}(T \mid X)}\Big)(Y(T) - \bar{Y}(T, X))\Big]. &(373)\end{aligned}$$

In other words, we first use our prediction $\bar{Y}$ to form a guess of the average treatment effect, then use inverse propensity weighting to correct the guess so as to obtain an unbiased estimate. This can yield substantial improvements when $Y(T) - \bar{Y}(T, X)$ is much smaller in magnitude than $Y(T)$. For instance, a patient's cholesterol after taking a cholesterol-reducing drug is still highly-correlated with their initial cholesterol, so in that case we can take $\bar{Y}(T, X)$ to be the pre-treatment cholesterol level. Even though this is independent of $T$ it can substantially reduce the variance of the estimate! (We will formally bound the variance below.)

**Bias of the estimate.** Call the prediction $\bar{Y}$ unbiased if $\mathbb{E}[Y \mid X, T] = \bar{Y}(T, X)$. The first key property of (373) is that it is unbiased as long as *either* $\bar{Y}$ is unbiased, or the propensity weights are correct. Indeed, if the prediction is unbiased then the first term is the average treatment effect while the second term is zero. Conversely, if the propensity weights are correct then the second term exactly estimates the difference between the predicted and true treatment effect. Correspondingly, (373) is called a *doubly-robust estimator*.

We can actually say more about the bias. Suppose that instead of the true propensity weights, we have an incorrect guess $\hat{p}(t \mid x)$. Then the bias of the estimate is the difference between $\mathbb{E}_{\tilde{p}}[Y(1) - Y(0)]$ and (373), which is

$$\begin{aligned}&\mathbb{E}_{\tilde{p}}[Y(1) - Y(0)] - \mathbb{E}[\bar{Y}(1, X) - \bar{Y}(0, X)] - \mathbb{E}_{\tilde{p}}\Big[\Big(\frac{\mathbb{I}[T = 1]}{\hat{p}(T \mid X)} - \frac{\mathbb{I}[T = 0]}{\hat{p}(T \mid X)}\Big)(Y(T) - \bar{Y}(T, X))\Big] &(374)\\ &= \mathbb{E}_{\tilde{p}}\Big[(Y(1) - \bar{Y}(1, X))\Big(1 - \frac{\mathbb{I}[T = 1]}{\hat{p}(t = 1 \mid X)}\Big) + (Y(0) - \bar{Y}(0, X))\Big(1 - \frac{\mathbb{I}[T = 0]}{\hat{p}(t = 0 \mid X)}\Big)\Big]. &(375)\end{aligned}$$

Focusing on the first term, and using the independence of $T$ and $Y$ conditioned on $X$, we have

$$\begin{aligned}&\mathbb{E}_{\tilde{p}}\Big[(Y(1) - \bar{Y}(1, X))\Big(1 - \frac{\mathbb{I}[T = 1]}{\hat{p}(t = 1 \mid X)}\Big)\Big] &(376)\\ &= \mathbb{E}_{\tilde{p}}\Big[(\mathbb{E}[Y(1) \mid X] - \bar{Y}(1, X))\Big(1 - \frac{\tilde{p}(t = 1 \mid X)}{\hat{p}(t = 1 \mid X)}\Big) \mid X\Big] &(377)\\ &\leq \mathbb{E}_{\tilde{p}}[(\mathbb{E}[Y(1) \mid X] - \bar{Y}(1, X))^2]^{1/2}\mathbb{E}_{\tilde{p}}\Big[\Big(1 - \frac{\tilde{p}(t = 1 \mid X)}{\hat{p}(t = 1 \mid X)}\Big)^2\Big]^{1/2}, &(378)\end{aligned}$$

meaning that the bias of the estimator is the *product* of the biases of $\bar{Y}$ and $\hat{p}$ (measured as the expected squared errors in (378)).

**Variance of the estimate.** We can obtain a somewhat similar relation for the variance. Usually the variance of $\bar{Y}(1, X) - \bar{Y}(0, X)$ is small compared to the propensity-weighted term, so again focusing on the $T = 1$ case we have

$$\mathsf{Var}\Big[(Y(1) - \bar{Y}(1, X))\frac{\mathbb{I}[T = 1]}{\hat{p}(t = 1 \mid X)}\Big] \leq \mathbb{E}\Big[\mathbb{E}_Y[(Y(1) - \bar{Y}(1, X))^2 \mid X]\frac{\tilde{p}(t = 1 \mid X)}{\hat{p}(t = 1 \mid X)^2}\Big]. \tag{379}$$

The variance is substantially reduced when $Y(1)$ is close to $\bar{Y}(1, X)$. We cannot always hope for this, e.g. if $Y$ has a large amount of intrinsic variance even conditioned on $X$. But in many cases even trivial $\bar{Y}$ can predict most of the variance in $Y$—for instance, for any chronic disease the patient's post-treatment status is well-predicted by their pre-treatment status. And the value of a stock tomorrow is well-predicted by its value today.

**Semi-parametric estimation.** It may seem difficult to estimate $\bar{Y}(\cdot, X)$ and $\tilde{p}(t = 1 \mid X)$, since any parametric model could be mis-specified and lead to biased estimates. One idea is to estimate these both non-parametrically, and then apply the doubly-robust estimator above. This is an instance of *semi-parametric estimation*, because while we estimate $\bar{Y}$ and $\tilde{p}(t \mid X)$ non-parametrically, the doubly-robust estimator itself is parametric (i.e a simple sample estimate of the mean), and in some cases we obtain non-parametric rates. This is explored in detail in Nie and Wager (2017) for estimating conditional average treatment effects; we describe the basic idea here. Since the squared error in an estimate is $\mathsf{Bias}^2 + \mathsf{Variance}/n$, the bias will dominate the error in the doubly-robust estimator as long as the variance doesn't explode (of course, the variance can often explode if the propensity weights are too close to 0 or 1, and the following idea won't help in that case).

We saw above that the bias of the doubly-robust estimator is the product of the biases in $\bar{Y}$ and $\hat{p}$, which are both given as expected squared errors between the true and estimated value. In non-parametric estimation, we typically get convergence rates of $\mathcal{O}(n^{-\alpha})$ for some $\alpha < 1/2$ (note that $\alpha = 1/2$ is what we typically get for parametric estimation). The parameter $\alpha$ typically depends on the dimension of the problem and the smoothness of the function class we wish to estimate. Since the doubly-robust bias is the product of the biases, we end up with a bias of $\mathcal{O}(n^{-2\alpha})$ as long as $\bar{Y}$ and $\hat{p}$ each converge at a $n^{-\alpha}$ rate. As long as $\alpha > 1/4$, this yields a parametric rate (the variance term will then asymptotically dominate as it only converges at $1/\sqrt{n}$).

# References

Radosław Adamczak and Paweł Wolff. Concentration inequalities for non-lipschitz functions with bounded derivatives of higher order. *Probability Theory and Related Fields*, 162(3-4):531–586, 2015.

Noga Alon and Assaf Naor. Approximating the cut-norm via grothendieck's inequality. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 72–80. ACM, 2004.

Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 37–49. Springer, 2012.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84*, pages 177–206. Springer, 1985.

Jean-Baptiste Bardet, Nathaël Gozlan, Florent Malrieu, Pierre-André Zitt, et al. Functional inequalities for gaussian convolutions of compactly supported measures: explicit bounds and dimension dependence. *Bernoulli*, 24(1): 333–353, 2018.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer School on Machine Learning*, pages 208–240. Springer, 2003.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford university press, 2013.

David L Donoho and Richard C Liu. The" automatic" robustness of minimum distance functionals. *The Annals of Statistics*, 16(2):552–586, 1988.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017.

Thomas Kühn. Eigenvalues of integral operators with smooth positive definite kernels. *Archiv der Mathematik*, 49(6): 525–534, 1987.

Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 299–308. IEEE, 2010.

Rafał Latała. Estimates of moments and tails of gaussian chaoses. *The Annals of Probability*, 34(6):2315–2331, 2006.

Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

James G MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, 29(3):305–325, 1985.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

Ali Rahimi, Benjamin Recht, et al. Random features for large-scale kernel machines. In *NIPS*, volume 3, page 5. Citeseer, 2007.

Omar Rivasplata. Subgaussian random variables: An expository note. 2012.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

# 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

$$\begin{aligned} |\mathbb{E}_{X\sim p}[X \mid E] - \mu| &= |\mathbb{E}_{X\sim p}[(X-\mu)\mathbb{I}[E]]|/\mathbb{P}[E] && (380) \\ &\leq \sqrt{\mathbb{E}_{X\sim p}[(X-\mu)^2] \cdot \mathbb{E}_{X\sim p}[\mathbb{I}[E]^2]}/\mathbb{P}[E] && (381) \\ &\leq \sqrt{\sigma^2 \cdot \mathbb{P}[E]}/\mathbb{P}[E] = \sigma/\sqrt{\mathbb{P}[E]}. && (382) \end{aligned}$$

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 (382) 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

$$\begin{aligned} |\mathbb{E}_{X\sim p}[X \mid E] - \mu| &= \frac{1-\mathbb{P}[E]}{\mathbb{P}[E]}|\mathbb{E}_{X\sim p}[X \mid \neg E] - \mu| && (383)\\ &\leq \frac{1-\mathbb{P}[E]}{\mathbb{P}[E]} \cdot \sigma/\sqrt{1-\mathbb{P}[E]}. && (384) \end{aligned}$$

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

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

# C Proof of Lemma 2.14

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

$$\begin{aligned} \|\mathbb{E}[X-\mu \mid E]\| &= \langle \mathbb{E}[X-\mu \mid E], v\rangle && (385)\\ &= \langle \mathbb{E}[\langle X-\mu, v\rangle \mid E] && (386)\\ &\overset{(i)}{\leq} \mathbb{E}[\langle X-\mu, v\rangle \mid \langle X-\mu, v\rangle \geq \tau_\epsilon(v)] && (387)\\ &\overset{(18)}{\leq} \frac{1-\epsilon}{\epsilon}\rho. && (388) \end{aligned}$$

Here (i) is because $\langle X-\mu, v\rangle$ is at least as large for the $\epsilon$-quantile as for any other event $E$ of probability $\epsilon$. This shows that (18) implies $(1-\epsilon, \frac{1-\epsilon}{\epsilon}\rho)$-resilience. For the other direction, given any $v$ let $E_v$ denote the event that $\langle X-\mu, v\rangle \geq \tau_\epsilon(v)$. Then $E_v$ has probability $\epsilon$ and hence

$$\begin{aligned} \mathbb{E}[\langle X-\mu, v\rangle \mid \langle X-\mu, v\rangle \geq \tau_\epsilon(v)] &= \mathbb{E}[\langle X-\mu, v\rangle \mid E_v] && (389)\\ &= \langle \mathbb{E}[X-\mu \mid E_v], v\rangle && (390)\\ &\overset{(ii)}{\leq} \|\mathbb{E}[X-\mu \mid E_v]\| && (391)\\ &\overset{(iii)}{\leq} \frac{1-\epsilon}{\epsilon}\rho, && (392) \end{aligned}$$

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 probablity at least $\frac{1}{2}$. Without loss of generality assume it is $E_+$. Then we have

$$\begin{aligned} \mathbb{E}_{x\sim p}[|\langle x-\mu, v\rangle|] &= 2\mathbb{E}_{x\sim p}[\max(\langle x-\mu, v\rangle, 0)] && (393)\\ &= 2\mathbb{P}[E_+]\mathbb{E}_{x\sim p}[\langle x-\mu, v\rangle \mid E_+] && (394)\\ &\leq 2\mathbb{P}[E_+]\|\mathbb{E}_{x\sim p}[x-\mu \mid E_+]\| \leq 2\rho, && (395) \end{aligned}$$

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

$$\begin{aligned} \mathbb{E}[\langle X-\mu, v\rangle \mid \langle X-\mu, v\rangle \geq \tau_{1/2}(v)] &\leq \mathbb{E}[|\langle X-\mu, v\rangle|]/\mathbb{P}[\langle X-\mu, v\rangle \geq \tau_{1/2}(v)] && (396)\\ &= 2\mathbb{E}[|\langle X-\mu, v\rangle|] \leq 2\rho, && (397) \end{aligned}$$

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