STAT260: Robust Statistics Course Notes

Feynman Liang*
Department of Statistics, UC Berkeley

Last updated: November 14, 2019

Contents

*feynman@berkeley.edu

1 9/3/2019

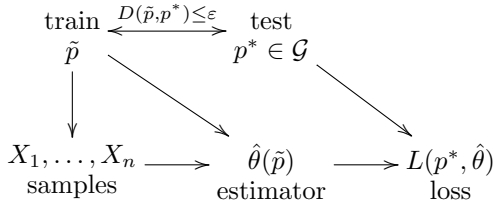


Figure 1: Overview of the framework. Training distribution $\tilde{p}$ differs from test distribution p^* by some discrepancy $D(\tilde{p}, p^*) \leq \epsilon$. We constrain $p^* \in \mathcal{G}$ to encode distributional assumptions. Given an estimator $\hat{\theta}$ trained using samples $X_1, \ldots, X_n \sim \tilde{p}$, we want to control the loss $L(p^*, \hat{\theta})$ incurred at test time. {fig:robust-statistics-framework}

1.1 Minimum distance functional

Introduced in donoho1988automatic Donoho et al. (1988), the minimum distance functional is one way to produce robust estimators which easily generalizes and also leverages distributional assumptions in $\mathcal{G}$.

Definition 1 (*Minimum distance functional*) {def:mdf}

The ***minimum distance functional*** (MDF) is

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

In other words, q is the projection (under D) of $\tilde{p}$ onto $\mathcal{G}$ and $\hat{\theta}$ is the estimator obtained by using q as the training distribution.

One nice property of the MDF is that we can bound it using a supremum over nearby pairs $p, q \in \mathcal{G}$ satisfying $D(p, q) \leq 2\varepsilon$. This is useful because we eliminate $\tilde{p}$ and focus the theory around $\mathcal{G}$.

Proposition 2 (*Modulus of continuity bound*) {prop:mdf-modulus-error-bound}

If D is a pseudometric (metric without requirement $d(x, y) = 0 \implies x = y$), then the cost $L(p^*, \hat{\theta}(\tilde{p}))$ of the MDF (Definition 1) is bounded by:

$$\mathfrak{m}(\mathcal{G}, 2\varepsilon, D, L) = \sup_{\substack{p,q \in \mathcal{G} \\ D(p,q) \leq 2\epsilon}} L(p, \theta^*(q)) \tag{2}$$

$\mathfrak{m}$ *is called the* **modulus of continuity**.

Proof. First fix $p = p^* \in \mathcal{G}$

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

Next, let $q = \operatorname{argmin}_{g \in \mathcal{G}} D(g, \tilde{p})$ be the projection of $\tilde{p}$ onto $\mathcal{G}$ as in Definition 1. Then since $D(p^*, \tilde{p}) \leq \varepsilon$ by assumption and $p^* \in \mathcal{G}$, we have

$$D(q, \tilde{p}) = \min_{g \in \mathcal{G}} D(g, \tilde{p}) \leq D(p^*, \tilde{p}) \leq \varepsilon \tag{4}$$

The following drawing visualizes the argument.

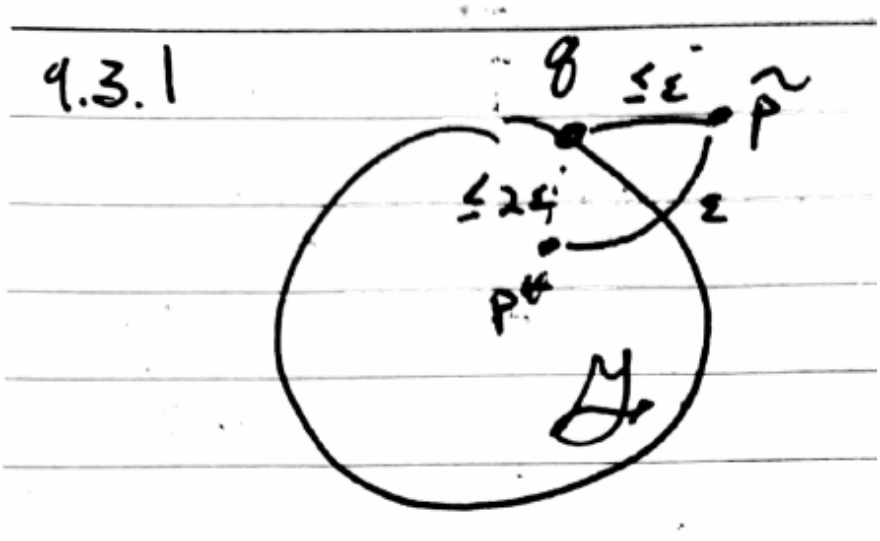


Figure 2: Given $D(p^*, \tilde{p}) \leq \varepsilon$, $p^* \in \mathcal{G}$, and q is the projection of $\tilde{p}$ onto $\mathcal{G}$ under D, we must have $D(\tilde{p}, q) \leq \varepsilon$ and by triangle inequality $D(p^*, q) \leq 2\varepsilon$

So $D(p^*, q) \leq 2\varepsilon$ and we can conclude

$$\mathfrak{m} \geq L(p^*, \theta^*(q)) \tag{5}$$

□

For now, we will specialize to the case $D = \mathrm{TV}$ and $L(p, \theta) = \|\theta - \mu(p^*)\|_2$. Consider a Gaussian distributional assumption $\mathcal{G}_{gauss} = \{\mathcal{N}(\mu, I) : \mu \in \mathbb{R}^d\}$.

Lemma 3

{lem:gauss-tv}

$\mathrm{TV}(\mathcal{N}(\mu, I), \mathcal{N}(\mu', I)) \asymp \Theta(\min(\|\mu - \mu'\|_2, 1))$
Therefore

$$\mathfrak{m}(\mathcal{G}_{gauss}, \varepsilon) = \sup_{\substack{p,q \in \mathcal{G} \\ \mathrm{TV}(p,q) \leq 2\varepsilon}} \|\mu(p) - \mu(q)\|_2 = \Theta(\varepsilon) \tag{6}$$

for sufficiently small ε.

Proof. We first prove the 1D case. By translational symmetry, we can translate both distributions while preserving $\|\mu - \mu'\|_2 =: u$ so that wlog we may assume the two distributions are $p = \mathcal{N}(\frac{u}{2}, 1)$ and $q =$

$\mathcal{N}(-\frac{u}{2}, 1)$. Then

$$\mathrm{TV}(p,q) = \frac{1}{2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} |e^{(t+u/2)^2/2} - e^{(t-u/2)^2/2}|dt \tag{7}$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-u/2}^{u/2} e^{-t^2/2}dt \tag{8}$$

{{eq:9-3-int}
}

where the last equality follows by cancelling the probability mass in the following picture:

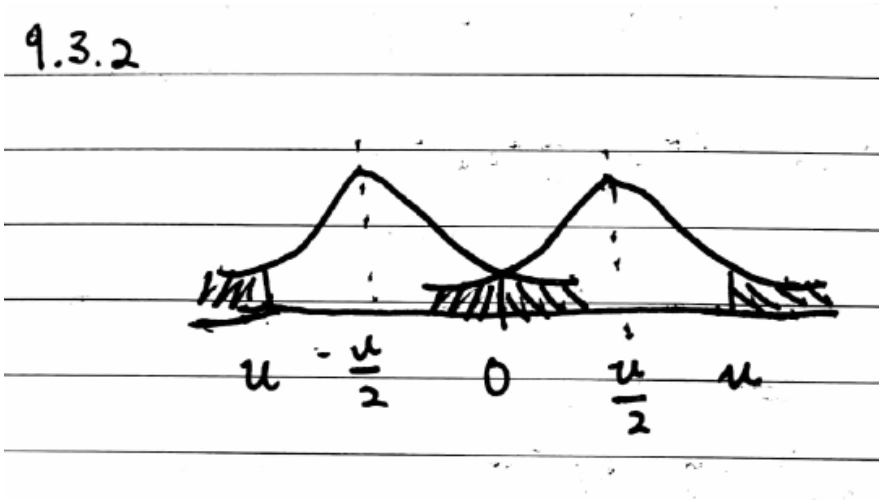


Figure 3: Both Gaussians exhibit identical $\pm\frac{u}{2}$ tails with opposite signs in the expression for TV, so the TV is equivalent to the area in $[-u/2, u/2]$ drawn out by the pointwise max between the two PDFs. By symmetry, this is just twice the area inside $[-u/2, u/2]$ which after cutting and pasting integration areas (and cancelling the 1/2 in definition of TV) is equal to the probability mass between $[-u/2, u/2]$ for a Gaussian.

Note that $e^{-t^2/2} \geq \frac{1}{2}$ if $|t| < 1$, so $\mathrm{TV} = \Omega(\min(u, 1))$ which can be seen by splitting the integral and examining the two cases where $\frac{u}{2} > 1$ (which yields the 1) and $\frac{u}{2} < 1$ (which yields the u).

Similarly, $e^{-t^2/2} \leq 1$ for all $t > 0$ so $\mathrm{TV} = O(\min(u, 1))$.

To generalize to higher dimensions, note identity covariance implies rotational invariance so we can rotate and translate such that the two means are on the first coordinate axis and separated by $\|\mu - \mu'\| = |\mu_1 - \mu_1'|$. In particular, $\mu_i = 0$ for $i \neq 1$ hence in the TV expression they can be factored out and integrated to 1 to reduce to the 1D case. □

1.2 Midpoint lemma and resilience

As a less restrictive family, consider distributions with bounded covariance:

$$\mathcal{G}_{cov}(\sigma) = \{p : \mathbb{E}_p[(X-u)(X-u)'] \preceq \sigma^2 I\} \tag{9}$$

We begin with an important lemma which will be used to prove the modulus of continuity for $\mathcal{G}_{cov}$ and generalized in the following section.

Lemma 4 (*Midpoint lemma*)

{lem:midpoint
}

If $\mathrm{TV}(p,q) \leq \varepsilon$ *then exists a* **midpoint** *distribution* r *such that* $r \leq \min\{\frac{p}{1-\varepsilon}, \frac{q}{1-\varepsilon}\}$ *and*

1. $r(x) \leq \frac{p(x)}{1-\varepsilon}$ *for all* x
2. r *is an* ϵ**-deletion of** p *(obtained by deleting* ϵ *mass from* p*)*
3. $r = p|_E$ *for* $p(E) \geq 1-\varepsilon$ *where* $E \mid X$ *has probability* 1 *if* $p(x) \leq q(x)$ *and* $\frac{q(x)}{p(x)}$ *if* $p(x) > q(x)$

Proof. The midpoint distribution is given by $r = \frac{\min(p,q)}{1-\mathrm{TV}(p,q)}$ and is obtained from p by deleting probability mass from q and renormalizing.

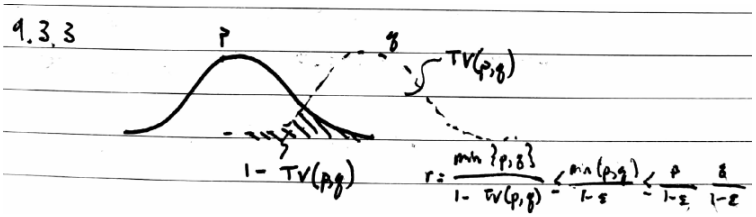


Figure 4: The midpoint distribution $r = \frac{\min(p,q)}{1-\mathrm{TV}(p,q)}$ can be reached from both p and q by deleting ϵ-mass and renormalizing.

Specifically, we delete $q(x) - p(x)$ mass from all points in $\{x : q(x) > p(x)\}$, the integral of which is precisely equal to the total variation distance. This means that we must renormalize by $1 - \varepsilon$ to ensure r is a proper distribution. □

Corollary 5

{corr:mod-con t-cov}

$\mathfrak{m}(\mathcal{G}_{cov}(\sigma), \varepsilon) = O(\sigma\sqrt{\varepsilon})$

Proof. Take $p, q \in \mathcal{G}_{cov}$ such that $\mathrm{TV}(p,q) \le \varepsilon$. By Lemma 4, there exists a midpoint distribution $r = p\,|_E$ for which

$$\mathbb{E}_r[X - \mu(p)] = \mathbb{E}_p[X - \mu(p) \mid \underbrace{E}_{1-\varepsilon}] = \frac{-\varepsilon}{1-\varepsilon}\mathbb{E}_p[X - \mu(p) \mid \underbrace{E^c}_{\varepsilon}] \tag{10}$$

where the last equality follows from

$$0 = \mathbb{E}_p[X - \mu(p)] = \underbrace{p(E)}_{1-\epsilon}\mathbb{E}_p[X - \mu \mid E] + \underbrace{p(E^c)}_{\epsilon}\mathbb{E}_p[X - \mu \mid E^c] \tag{11}$$

(This is a common trick for moving from conditioning on an event to conditioning on its complement in zero mean functionals).

(Chebyshev in $\mathbb{R}^d$) By linearity of expectation and Jensen's inequality

$$\|\mathbb{E}_p[X - \mu(p) \mid E^c]\|_2 = \sup_{\|v\|_2 \le 1} \langle \mathbb{E}_p[X - \mu(p) \mid E^c], v\rangle \tag{12}$$

$$= \sup_{\|v\|_2 \le 1} \mathbb{E}_p[\langle X - \mu(p), v\rangle \mid E^c] \tag{13}$$

$$\le \sup_{\|v\|_2 \le 1} \sqrt{\mathbb{E}_p[\langle X - \mu(p), v\rangle^2 \mid E^c]} \tag{14}$$

Note $\mathbb{E}_p[\langle X - \mu(p), v\rangle^2] = \mathrm{Var}_p[\langle X - \mu(p), v\rangle] = v^\top \mathrm{Cov}_p(X) v \le \sigma^2$ so

$$\|\mathbb{E}_p[X - \mu(p) \mid E^c]\|_2 \le \sqrt{\frac{\sigma^2}{\Pr[E^c]}} = \frac{\sigma}{\sqrt{\varepsilon}} \tag{15}$$

{{eq:9-3-cheb -cov-control}}

As a result, we have

$$\|\mu(r) - \mu(p)\|_2 = \|\mathbb{E}_r[X - \mu(p)]\|_2 \le \frac{\varepsilon}{1-\varepsilon}\frac{\sigma}{\sqrt{\varepsilon}} \le 2\sigma\sqrt{\varepsilon} \tag{16}$$

for $\varepsilon < 1/2$. A similar argument involving q gives $\|\mu(r) - \mu(q)\|_2 \le 2\sigma\sqrt{\varepsilon}$ so by triangle inequality $\|\mu(p) - \mu(q)\|_2 \le 4\sigma\sqrt{\varepsilon}$. □

Remark 6. Unlike the trimmed mean, there is no dependence on d here. This means that the MDF remains a good robust estimator even in high dimensions!

The above proof utilizes two key ingredients:

- The midpoint property of TV; both p and q are close to some ε-deletion r
- The bounded tails (second moment) of $\mathcal{G}_{cov}$, which is used to control how close $\mu(r)$ and $\mu(p)$ are in Eq. (15)

The previous proof can be suitably generalized to yield a modulus of continuity bound for other families:

{def:resilience}

Definition 7 (*Resilient distribution*)

A distribution is (ρ, ε)-resilient if

$$r \leq \frac{p}{1-\varepsilon} \implies \|\mathbb{E}_r[X] - \mathbb{E}_p[X]\|_2 \leq \rho \tag{17}$$

{eq:resilience-property}

In other words, for any (not just midpoint) ε-deletion r the mean does not change in norm by more than ρ. Equivalently (e.g. when p does not have a density) we can view $r = p\,|_E$ for an event E and use the definition

$$p(E) \geq 1 - \varepsilon \implies \|\mathbb{E}_p[X \mid E] - \mathbb{E}_p[X]\| \leq \rho \tag{18}$$

We let $\mathcal{G}_{TV}(\rho, \epsilon)$ be the set of all (ρ, ϵ)-resilient distributions.

Remark 8. This definition is only applicable for mean estimation under squared error loss.

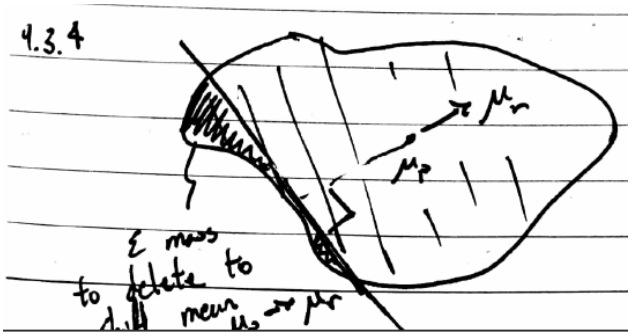


Figure 5: Deleting ε mass from a resilient distribution p shifts the mean by a controlled amount $\|\mu_p - \mu_r\|_2 \leq \rho$.

{eg:bdd-cov-resilient}

Example 9

Corollary 5 shows $\mathcal{G}_{cov}(\sigma) \subset \mathcal{G}_{TV}(2\sigma\sqrt{\varepsilon}, \varepsilon)$

Example 10

Lemma 3 shows $\mathcal{G}_{gauss}(\sigma) \subset \mathcal{G}_{TV}(\varepsilon\sqrt{\log \frac{1}{\varepsilon}}, \varepsilon)$

Combining with Proposition 2, for squared error loss we can say

{corr:mod-bound-resilient}

Corollary 11 (*Modulus of continuity bound for resilient distributions*)

$$\mathfrak{m}(\mathcal{G}_{TV}(\rho, \varepsilon), \varepsilon) \leq 2\rho \tag{19}$$

Proof. For any $p, q \in \mathcal{G}_{\text{TV}}$, use Lemma 4 to get a midpoint distribution and then Eq. (17) with triangle inequality to control the squared error loss. □

So we can always project onto the family of resilient distributions to get a MDF estimator which has loss independent of d.

1.3 Orlicz norms

Definition 12 (*Orlitz function / norm*)

An ***Orlicz function*** $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is

1. Convex
2. Non-decreasing
3. $\psi(0) = 0$, $\psi(x) \to \infty$ as $x \to \infty$

Given an Orlicz function ψ, the ***Orlicz norm*** or ψ-norm of a random variable X is

$$\|X\|_\psi = \inf \left\{ t : \mathbb{E}\psi\left(\frac{|X|}{t}\right) \leq 1 \right\} \tag{20}$$

For multivariate $X \in \mathbb{R}^d$, define

$$\|X\|_\psi = \inf \left\{ t > 0 : \sup_{v \in \mathcal{S}^{d-1}} \| \langle X, v \rangle \|_\psi \leq t \right\} \tag{21}$$

{{def:orlicz-norm-multivar}}

In other words, X has bounded ψ-norm if all of its one dimensional projections do.

Let $\mathcal{G}_\psi(\sigma) = \{X : \|X - \mathbb{E}[X]\|_\psi \leq \sigma\}$.

Example 13

$\psi(x) = x^k$ gives $\|X\|_\psi = \left(\mathbb{E}[|X|^k]\right)^{1/k}$, which looks like L_p norms. In fact, these are precisely distributions with bounded kth moments.

For $\psi(x) = x^2$, we have $\mathcal{G}_\psi(\sigma) = \mathcal{G}_{cov}(\sigma)$.

Definition 14 (*Sub-Gaussian/Sub-Exponential*)

{def:sub-gaussian-sub-exp-orlicz}

For $\psi_2(x) = e^{x^2} - 1$, $\mathcal{G}_{\psi_2}(\sigma)$ are called the σ-sub-Gaussian random variables.

For $\psi_1(x) = e^x - 1$, $\mathcal{G}_{\psi_1}(\sigma)$ are called the σ-sub-exponential random variables.

The next proposition shows that any distribution with bounded Orlicz norm is resilient.

Proposition 15 (*Bounded Orlicz norm implies resilience*)

{lem:orlicz-norm-resilient}

$\mathcal{G}_\psi(\sigma) \subset \mathcal{G}_{TV}(2\sigma\varepsilon\psi^{-1}(\frac{1}{\varepsilon}), \varepsilon)$ *if* $\varepsilon < 1/2$.

$\psi(x) \to \psi^{-1}(x) = \sqrt{x} \to \varepsilon\psi^{-1}(1/\varepsilon) = \sqrt{\varepsilon}$

Proof.

$$\|\mathbb{E}_r[X] - \mathbb{E}_p[X]\|_2 = \|\mathbb{E}_p[X - \mu \mid \underbrace{E}_{p(E)=1-\varepsilon}]\|_2 = \frac{\varepsilon}{1-\varepsilon}\|\mathbb{E}_p[X - \mu \mid E^c]\| \tag{22}$$

Focusing in on the expectation term

$$\|\mathbb{E}_p[X - \mu \mid E^c]\|_2 = \sup_{\|v\|_2 = 1} \mathbb{E}_p[\langle X - \mu, v \rangle \mid E^c] \tag{23}$$

By Jensen's inequality, convexity of ψ (equivalently concavity of ψ^{-1}), definition of multivariate Orlicz norm (Eq. (21)), and monotonicity of ψ, we have

$$\|\mathbb{E}_p[X-\mu \mid E^c]\|_2 = \sup_{\|v\|_2=1} \sigma\left(\mathbb{E}_p\left[(\sigma\psi^{-1}\circ\psi)\left(\frac{|\langle X-\mu, v\rangle|}{\sigma}\right) \mid E^c\right]\right) \tag{24}$$

$$\leq \sup_{\|v\|_2=1} \sigma\psi^{-1}\left(\mathbb{E}_p\left[\psi\left(\frac{|\langle X-\mu, v\rangle|}{\sigma}\right) \mid E^c\right]\right) \tag{25}$$

$$\leq \sup_{\|v\|_2=1} \sigma\psi^{-1}\left(\underbrace{\mathbb{E}_p\left[\psi\left(\frac{|\langle X-\mu, v\rangle|}{\sigma}\right)\right]}_{\leq 1}\underbrace{\frac{1}{\Pr[E^c]}}_{\frac{1}{\varepsilon}}\right) \tag{26}$$

$$\leq \sigma\psi^{-1}(\frac{1}{\varepsilon}) \tag{27}$$

□

2 9/5/2019

2.1 Recap

- Minimum distiance functionals: good error, bounded by modulus of continuity $\mathfrak{m}$
- Resilience $\implies$ bounded $\mathfrak{m}$
- Bounded Orlicz ψ-norm $\implies$ resilience

This lecture:

- Want to analyze $X_1, \ldots, X_n$
- "The empirical average converges to the mean if n is large"
- Two steps:
  1. Show ***concentration inequality***: bound variation of p in terms of σ
  2. Show ***composition property***: σ gets smaller as we take more independent samples

2.2 Concentration Inequalities and Composition

Example 16

A slot machine has expected payout of \$5 and always pays out positive.
Question: What is the maximum probability of $\geq$ \$100?
Answer: 5%, by letting $P(X = \$0) = 0.95$ and letting $P(X = \$100) = 5\%$.

The preceding example is an instance of Markov's Inequality:

Theorem 17 (*Markov's Inequality*)

{thm:markov}

If $X \geq 0$ has bounded first moment, then

$$\Pr[X \geq t\mathbb{E}[X]] \leq \frac{1}{t} \tag{28}$$

Proof.

$$t\mathbb{E}[X]\mathbb{1}\{X \geq t\mathbb{E}[X]\} \leq X \tag{29}$$

Take expectation of both sides and rearrange. □

Markov's Inequality has a nice composition property:

Theorem 18 (*Composition of Markov for sums*)

{thm:markov-composition}

Let $X_1, X_2 \sim p$ with mean μ.

$$\Pr\left[\frac{X_1 + X_2}{2} \geq t\mu\right] \leq \frac{1}{t} \tag{30}$$

This is because $\mathbb{E}[(X_1 + X_2)/2] = \mu = \mathbb{E}[X_1] = \mathbb{E}[X_2]$.

We can apply Markov's Inequality to $Z = f(X)$ for $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ and get a family of inequalities (provided $\mathbb{E}[f(X)] < \infty$). For example, taking $Z = (X - \mu)^2$ and assuming $\mathbb{E}[Z] = \mathrm{Var}[X] = \sigma^2 < \infty$ yields

Theorem 19 (*Chebyshev's inequality*)

{thm:chebyshev}

$$\Pr[|X - \mu| \geq t\sigma] \leq \frac{1}{t^2} \tag{31}$$

Analogous to Theorem 18 (Composition of Markov for sums), a composition property for Chebyshev's inequality would require a composition property involving variances:

Theorem 20 (*Variances add for independent RVs*)

If $\{X_i\}_{i=1}^n$ are independent, then

$$\mathrm{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathrm{Var}[X_i] \tag{32}$$

Example 21 (*Concentration of empirical average*)

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p$ with mean μ and variance σ^2. Let $S = \sum_i^n X_i$ and $\frac{S}{n}$ the empirical average. Then

$$\mathrm{Var}[S/n] = n\,\mathrm{Var}[X/n] = n\frac{\sigma^2}{n^2} = \frac{\sigma^2}{n} \tag{33}$$

Hence, the standard deviation of the empirical average $\frac{S}{n}$ is $\sigma_{avg} = \frac{\sigma}{\sqrt{n}}$. Chebyshev's inequality then yields

Corollary 22

{cor:chebyshev-empirical-avg-concentration}

$$\Pr\left[\left|\frac{S}{n} - \mu\right| \geq t\frac{\sigma}{\sqrt{n}}\right] \leq \frac{1}{t^2} \tag{34}$$

The t^{-2} quadratic decay in Corollary 22 is tight, as the following proposition shows:

Proposition 23 (*Lower bounds for Chebyshev*)

There exists $X_1, \ldots, X_n$ pairwise independent, bounded in $[-1, 1]$, mean zero, variance one, such that

$$\Pr\left[\sum_{i=1}^n X_i = n\right] = \frac{1}{n} \tag{35}$$

Consequently, Corollary 22 (with $\mu = 0$, $\sigma = 1$, and $t = \sqrt{n}$) is tight.

Proof. Flip k independent coins and let $n = 2^k$. For any subset $\emptyset \subsetneq S \subset [k]$, define the random variable

$$X_S = \begin{cases} 1 & \text{\# heads in } S \text{ is even} \\ -1 & \text{\# heads in } S \text{ is odd} \end{cases} \tag{36}$$

X_S is mean zero, variance one, bounded $[-1, 1]$, and pairwise independent (since the coin flips are). The event $\{\sum_{i=1}^n X_i = n\}$ occurs iff all of the coins land tails, which occurs with probability $2^{-k} = \frac{1}{n}$. □

2.3 Failure of composition of higher moments and Rosenthal's inequality

{ssec:rosenthals-inequality}

To try to extend Chebyshev's inequality, we can consider applying Markov's Inequality to $Z = f(X) = (X - \mu)^4$ to get:

Theorem 24

$$\Pr[|X - \mu| \geq t\mathbb{E}[Z]^{1/4}] \leq \frac{1}{t^4} \tag{37}$$

However, the composition property fails here since for $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$ we find

$$\mathbb{E}[(X_1 + X_2)^4] = \mathbb{E}[X_1^4] + \mathbb{E}[X_2^4] + 4\cancelto{0}{\mathbb{E}[X_1]}\mathbb{E}[X_2^3] + 4\cancelto{0}{\mathbb{E}[X_2]}\mathbb{E}[X_1^3] + \underbrace{6\mathbb{E}[X_1^2 X_2^2]}_{\geq 0} \tag{38}$$

{eq:9-5-4th-moment-nonadditive}

Thus, the fourth moment of a sum can be larger than the sum of the fourth moments.

In general, higher moments don't add. One method to work around this is to work with cumulants (see Section 2.6). An alternative method is through Rosenthal's inequality:

Lemma 25 (*Rosenthal's inequality*)

{lem:rosenthal-ineq}

If $X_1, \ldots, X_n$ are independent mean zero random variables with finite pth moments, then

$$\mathbb{E}\left[\left|\sum_{i=1}^n X_i\right|^p\right] \leq O(p)^p \sum_{i=1}^n \mathbb{E}[|X_i|^p] + O(\sqrt{p})^p \left(\sum_{i=1}^n \mathbb{E}[X_i^2]\right)^{p/2} \tag{39}$$

How can we use Rosenthal's inequality? Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \pi$ with $\mathbb{E}[|X|^p] = k^p$ and $\mathbb{E}[X^2] = \sigma^2$. Let $S = \sum_{i=1}^n X_i$. Then

$$\mathbb{E}[|S|^p] \leq O(p)^p nk^p + O(\sqrt{p})^p (n\sigma^2)^{p/2} \tag{40}$$

$$\mathbb{E}[|S|^p]^{1/p} \leq O(pkn^{1/p} + \sqrt{p}\sigma n^{1/2}) \tag{41}$$

$$\mathbb{E}\left[\left|\frac{S}{n}\right|^p\right]^{1/p} \leq O(pkn^{-(1-\frac{1}{p})} + \sqrt{p}\sigma n^{-1/2}) \tag{42}$$

Hence, all of the pth moments of the empirical mean $\frac{S}{n}$ decay in n, so the empirical mean concentrates about the population mean as the number of samples $n \to \infty$.

2.4 Exponential tails and Chernoff bounds

Another approach which can yield exponential tail bounds is through the Moment Generating Function.

Definition 26 (*Moment Generating Function*)

{def:moment-generating-function}

Let X be a random variable with bounded moments. The ***moment generating function*** (MGF)

of X is

$$m_X(\lambda) = \mathbb{E}\exp(\lambda(X-\mu)) = 1 + \lambda\mathbb{E}[x] + \frac{\lambda^2}{2}\mathbb{E}[X^2]\frac{\lambda^3}{6}\mathbb{E}[X^3] + \cdots \tag{43}$$

MGFs satisfy a desirable composition property enabling us to easily compute the MGF of a sum in terms of the MGFs of the summands:

Lemma 27 (*Composition property for MGFs*)

{lem:mgf-sum-composition}

If $X_1, \ldots, X_n$ are independent, then

$$m_{\sum_i^n X_i}(\lambda) = \prod_i^n m_{X_i}(\lambda) \tag{44}$$

Proof. Exponential of sum is product of exponentials, independence of X_i allows splitting of $\mathbb{E}$. □

Another strong advantage of working with moment generating functions is that we can use them to get exponentially decaying tail bounds:

Theorem 28 (*Chernoff's bound*)

{thm:chernoff}

For $\lambda \geq 0$,

$$\Pr[X - \mu \geq t] \leq \inf_{\lambda \geq 0} m_X(\lambda)e^{-\lambda t} \tag{45}$$

Proof. $X - \mu \geq t$ implies $\exp(\lambda(X-\mu)) \geq e^{\lambda t}$. The same technique used to prove Chebyshev's inequality (with $f(x) = e^{\lambda x}$) gives

$$\Pr[\exp(\lambda(X-\mu)) \geq e^{\lambda t}] \leq e^{-\lambda t} m_X(\lambda) \tag{46}$$

□

Example 29 (*sub-exponential Chernoff bound*)

{eg:sub-exponential-chernoff}

Recall from Definition 14 (Sub-Gaussian/Sub-Exponential) that σ-sub-exponential means bounded Orlicz norm $\|X-\mu\|_\psi = \mathbb{E}\left[\psi\left(\frac{|X-\mu|}{\sigma}\right)\right] \leq 1$ for $\psi(x) = e^x - 1$. Chernoff's bound then implies

$$\mathbb{E}[\exp(|X-\mu|/\sigma) - 1] \leq 1 \tag{47}$$

$$\mathbb{E}[\exp(|X-\mu|/\sigma)] \leq 2 \tag{48}$$

$$m_X(1/\sigma) = \mathbb{E}\exp\left(\frac{x-\mu}{\sigma}\right) \leq \mathbb{E}\exp\left(\frac{|x-\mu|}{\sigma}\right) \leq 2 \tag{49}$$

$$\Pr[X - \mu \geq t] \leq 2\exp(-t/\sigma) \tag{50}$$

This explains the name "sub-exponential": the tail probabilities decay faster than an exponential.

Example 30 (*sub-Gaussian Chernoff bound*)

{eg:sub-gaussian-chernoff}

Recall from Definition 14 (Sub-Gaussian/Sub-Exponential) that σ-sub-Gaussian means bounded Orlicz norm $\|X-\mu\|_\psi$ with $\psi(x) = e^{x^2} - 1$. Hence, $\mathbb{E}[\exp((X-\mu)^2/\sigma^2)] \leq 2$ and

$$m_X(\lambda) = \mathbb{E}\exp(\lambda(X-\mu)) \leq \exp(\lambda^2\sigma^2/4)\mathbb{E}\exp((X-\mu)^2/\sigma^2)] \leq 2\exp(\lambda^2\sigma^2/4) \tag{51}$$

where we have used inequality $ab \leq \frac{a^2}{4} + b^2$ to conclude

$$\lambda(X - \mu) \leq \frac{\lambda^2\sigma^2}{4} + \frac{(X-\mu)^2}{\sigma^2} \tag{52}$$

Remark 31. We can also show

$$m_X(\lambda) \leq \exp\left(\frac{1}{2}\lambda^2(\sigma')^2\right) \tag{53}$$

where $\sigma' \leq \sqrt{3}\sigma$. This is sometimes taken as definition of σ'-sub-Gaussian.

Applying Chernoff's bound shows

$$\Pr[X - \mu \geq t] \leq \inf_{\lambda \geq 0} m_X(\lambda)e^{-\lambda t} \tag{54}$$

$$\leq \inf_{\lambda \geq 0} \exp\left(\frac{1}{2}\lambda^2(\sigma')^2 - \lambda t\right) \tag{55}$$

$$= \exp\left(-\frac{t^2}{2(\sigma')^2}\right) \tag{56}$$

This explains the name "σ'-sub-Gaussian": the tail probabilities are decaying faster than those of a Gaussian distribution with variance σ'.

By Lemma 27 (Composition property for MGFs), we have that the sum $S = \sum_i^n X_i$ of σ'-sub-Gaussian RVs is itself $\frac{\sigma'}{\sqrt{n}}$-sub-Gaussian and satisfies the tail bound

$$\Pr\left[\frac{S}{n} - \mu \geq t\right] \leq \exp\left(-\frac{nt^2}{2(\sigma')^2}\right) = \exp\left(-\frac{nt^2}{6\sigma^2}\right) \tag{57}$$

{{eq:sum-sub-gaussian-tail-bound}}

This yields our desired exponential rate of concentration.

2.5 Bounded random variables

Bounded RVs are sub-Gaussian, but we can get tighter bounds than the previous example. Let $X - \mu \in [-M, M]$. Then

$$\mathbb{E}\exp\frac{|X-\mu|}{M^2/\log 2} \leq \mathbb{E}\exp\log 2 = 2 \tag{58}$$

Hence X is sub-Gaussian with parameter $\sigma = \sqrt{\frac{M^2}{\log 2}}$ and we can use Eq. (66) to get tail bounds. More generally:

Corollary 32 (*Hoeffding's inequality*)

{corr:hoeffding-inequality}

Let $X_1, \ldots, X_n \in [a, b]$ be bounded independent mean zero random variables. Then

$$\Pr\left[\frac{S}{n} - \mu \geq t\right] \leq \exp\left(-\frac{2nt^2}{(a-b)^2}\right) \tag{59}$$

Proof. Bound MGF (tighter than what we are doing here) and apply Chernoff's bound. □

Hoeffding's inequality shows that an empirical average of independent bounded random variables converges to its mean at a rate of $\frac{1}{\sqrt{n}}$ with tails that decay at least as fast as Gaussians. Compare this against the $\frac{1}{n}$ rate for sub-exponentials we found in Example 29 and the quadratic $\frac{1}{t^2}$ tails from Chebyshev's inequality (which only required finite second moments).

2.6 Aside: Cumulants are additive

{ssec:cumulants-additive}

We saw in Section 2.3 that fourth moments are additive. While Lemma 27 (Composition property for MGFs) provides a convenient composition property for moment generating functions, the existence of MGFs requires all moments of the random variable to be bounded. In particular, this excludes random variables with fat tails.

To construct additive quantities, we can start with MGF (multiplicative) and take log (which is additive)

$$K_X(\lambda) = \log \mathbb{E} \exp(\lambda(X-\mu)) \tag{60}$$

$$= \log\left(1 + \mathbb{E}[(X-\mu)^2]\frac{\lambda^2}{2} + \mathbb{E}[(X-\mu)^3]\frac{\lambda^3}{6} + \cdots\right) \tag{61}$$

{{eq:cumulant-function-taylor-series}}

$$= 1 + \sum_{n=1}^{\infty} \frac{\kappa_n(X)}{n!}\lambda^n \tag{62}$$

This leads to the cumulant generating function:

Definition 33 (*Cumulants*)

The ***cumulant generating function*** for a random variable X is

$$K_X(\lambda) = \log \mathbb{E} \exp(\lambda X) = 1 + \sum_{n=1}^{\infty} \frac{\kappa_n(X)}{n!}\lambda^n \tag{63}$$

$\kappa_n(X)$ is called the nth cumulant of X.

Notice $K_{X+Y}(\lambda) = K_X(\lambda) + K_Y(\lambda)$ so we have additivity of the CGF and consequentially

$$\kappa_4(X+Y) = \kappa_4(X) + \kappa_4(Y) \tag{64}$$

Contrast this to Eq. (38).

However, computing the cumulants require Taylor expanding log using the infinite series in Eq. (61) as the argument and are laborious to work with. To handle heavy tails, it may be easier to use Rosenthal's inequality instead.

2.7 Max of n sub-Gaussians

Let $X_1, \ldots, X_n \sim p$, p is σ-sub-Gaussian. A simple union bound shows:

Theorem 34 (*Max of sub-gaussian bound*)

{thm:max-sub-gaussian}

$$\Pr[X_1 \vee \cdots \vee X_n \geq t] \leq \sum_{i=1}^{n} \Pr[X_i \geq t] \leq n e^{-\frac{t^2}{2\sigma^2}} \tag{65}$$

So in particular if $t \gg \sigma\sqrt{\log n}$, then its not likely the max will exceed t.

3 9/10/2019

3.1 Bounding suprema via concentration

The typical type of quantity we will focus on here is

$$\underbrace{\sup_{v \in V}}_{\text{bound by discretization}} \underbrace{\frac{1}{n}\sum_{i=1}^{n}(X_i(v) - \mathbb{E}[X(v)])}_{\text{bound for fixed } v \text{ via concentration}} \tag{66}$$

{{eq:typical-sup-via-concentration}}

When V is finite, a simple union bound can be applied. To deal with infinitely large $|V|$, we will need to first discretize V into a finite set.

3.2 Warmup: max of sub-Gaussian

Suppose $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} p$ where p is mean zero and σ-sub-Gaussian. How big is $\max_{i=1}^n X_i$?

Lemma 35

{lem:concentration-max-sub-gaussian}

With probability $\geq 1-\delta$

$$\max_{i=1}^n X_i \in O\left(\sigma\sqrt{\log n + \log\frac{1}{\delta}}\right) \tag{67}$$

Proof. By union bound, iid, and sub-Gaussian Chernoff bound

$$\Pr\left[\max_{i=1}^n X_i \geq t\right] \leq n\Pr[X_1 \geq t] \leq n\exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{68}$$

To ensure this failure event occurs with probability $\leq \delta$, we need

$$n\exp\left(-\frac{t^2}{2\sigma^2}\right) \leq \delta \tag{69}$$

$$\frac{t^2}{2\sigma^2} = \log n + \log\frac{1}{\delta} \tag{70}$$

$$t \leq \sigma\sqrt{2\left(\log n + \log\frac{1}{\delta}\right)} \tag{71}$$

□

If instead we were interested in $\max_{i=1}^n |X_i|$, then a union bound on the two tail events $\{X_i \geq t\}$ and $\{-X_i \geq t\}$ (note $-X_i$ is still sub-Gaussian) gives

$$\Pr\left[\max_{i=1}^n |X_i| \geq t\right] \leq 2n\exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{72}$$

{{eq:tail-bound-abs-value-two-factor}}

$$\max_{i=1}^n |X_i| \in O\left(\sigma\left(\sqrt{\log 2 + \log n + \log\frac{1}{\delta}}\right)\right) \tag{73}$$

In later the next section, we will see how we can "reduce" an infinitely large V into an exponentially large N after which we will use the same technique to bound the event $\{\max_{i\in N} X_i \geq t\}$. To get concentration, we will need the exponential tail bound to dominate the now exponentially large $n = |N|$ arising from the union bound over N.

3.3 Maximum eigenvalue of random matrix

Suppose $\{X_i \in \mathbb{R}^d\}_{i=1}^n \overset{\text{iid}}{\sim} p$ with p zero mean and σ-sub-Gaussian.

Recall from Eq. (21) (Orlitz function / norm) and Definition 14 (Sub-Gaussian/Sub-Exponential) that $X \in \mathbb{R}^d$ is σ-sub-Gaussian if all its one dimensional projections are, that is:

$$\|X\|_\psi + 1 = \sup_{v\in\mathcal{S}^{d-1}} \|\langle v, X\rangle\|_\psi + 1 = \sup_{v\in\mathcal{S}^{d-1}} \mathbb{E}\exp\left(\frac{\langle v, X\rangle^2}{\sigma^2}\right) \leq 2 \tag{74}$$

{{eq:9-10-recap-multivar-sub-gaussian}}

We are interested in the (random) empirical covariance matrix

$$M = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top \tag{75}$$

Specifically, we would like to understand how big $\|M\| = \lambda_{\max}(M)$ is.

Proposition 36

{prop:9-10-cov-operator-norm}

With probability $\geq 1-\delta$

$$\|M\| = O\left(\sigma^2\left(1+\frac{d}{n}+\frac{\log\frac{1}{\delta}}{n}\right)\right) \tag{76}$$

Remark 37. Proposition 36 shows that:

- As $n\to\infty$, $\|M\| = O(\sigma^2)$ and does not depend on d.
- The population covariance operator norm $\|\mathbb{E}X_iX_i^\top\| = O\left(\frac{\sigma^2}{n}\log\frac{1}{\delta}\right)$ is attained if $d=\Theta(n)$ (i.e. if the dimension grows at the same rate as n)

To relate back to the two-step strategy outlined in Eq. (66), note

$$\|M\| = \sup_{v\in\mathcal{S}^{d-1}} v^\top Mv = \sup_{v\in\mathcal{S}^{d-1}} \frac{1}{n}\sum_{i=1}^n \langle X_i, v\rangle^2 \tag{77}$$

This quantity looks promising as it is the sum of independent sub-Gaussian RVs.

Since $\langle X_i, v\rangle$ is σ-sub-Gaussian, $\langle X_i, v\rangle^2$ is σ^2-sub-exponential (Definition 14, or equivalently Eq. (74)) and for any fixed $v\in\mathcal{S}^{n-1}$

$$\mathbb{E}\exp\left(\frac{\langle X_i, v\rangle^2}{\sigma^2}\right) \leq 2 \tag{78}$$

$$\mathbb{E}\exp\left(\frac{n}{\sigma^2}\frac{1}{n}\sum_{i=1}^n\langle X_i, v\rangle^2\right) = \prod_{i=1}^n \mathbb{E}\exp\left(\frac{\langle X_i, v\rangle^2}{\sigma^2}\right) \leq 2^n \tag{79}$$

where we used Composition property for MGFs for the equality in the second line.

By Theorem 28, for fixed $v\in\mathcal{S}^{d-1}$

$$\Pr[v^\top Mv \geq t] \leq 2^n\exp\left(\frac{-nt}{2\sigma^2}\right) \tag{80}$$

{eq:9-10-conc-sum}

So we have accomplished the first step (showing the individual terms inside the sup concentrate for fixed v).

For the second step, we will take a sufficiently small discretization of the unit ball $\{\|v\|\leq 1\}$:

Lemma 38

{lem:9-10-unit-ball-packing}

There exists a finite set $N\subset\mathbb{R}^d$ *with* $|N|\leq 9^d$ *and*

$$\sup_{v\in\mathcal{S}^{d-1}} v^\top Mv \leq 2\sup_{v\in N} v^\top Mv \tag{81}$$

Applying Lemma 38, a union bound, Eq. (80), and bounding the failure probability by δ shows that

$$\Pr[\|M\|\geq t] = \Pr\left[\sup_{v\in\mathcal{S}^{d-1}} v^\top Mv \geq t\right] \leq 9^d 2^n\exp\left(\frac{-nt}{2\sigma^2}\right) = \delta \tag{82}$$

$$\frac{nt}{2\sigma^2} = d\log 9 + n\log 2 + \log\frac{1}{\delta} \tag{83}$$

$$t = O\left(\sigma^2\left(\frac{d}{n}+1+\frac{\log 1/\delta}{n}\right)\right) \tag{84}$$

Proof of Lemma 38. Let N be a maximal packing of $\mathrm{Ball}_1(0)$ in $\mathbb{R}^d$ such that $\|u-v\|_2\geq\frac{1}{4}$ for all $u\neq v\in N$.

As shown in Fig. 6, if we place a 1/8-radius ball at all the points in N then (1) all the balls are disjoint and (2) the union of all the balls is contained in $\text{Ball}_{9/8}(0)$ Therefore, by the (converse of the) pidgeonhole principle, $|N| \leq \frac{\text{Vol}(\text{Ball}_{9/8}(0))}{\text{Vol}(\text{Ball}_{1/8}(0))} = 9^d$.

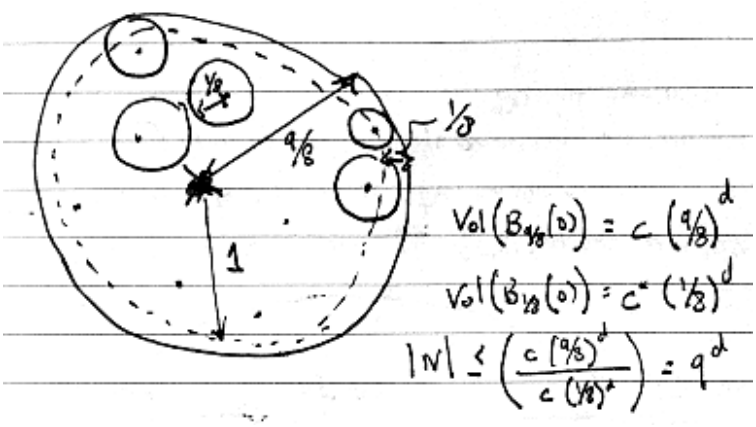


Figure 6: 1/8-radius balls centered at all packing points are disjoint, the union of all these balls is contained in $B_{9/8}(0)$, so the cardinality $|N| \leq \left(\frac{9/8}{1/8}\right)^d = 9^d$. {fig:9-10-1}

Let $v \in \mathcal{S}^{d-1}$ maximize $v^\top M v$ and $u \in N$ such that $\|u-v\|_2 \leq \frac{1}{4}$. Such a u must exist, otherwise $N \cup \{v\}$ is a larger 1/4-packing which contradicts maximality of N.

$$\begin{aligned}
\|M\| - |u^\top M u| &= |v^\top M v| - |u^\top M u| && (85)\\
&\leq |v^\top M v - u^\top M u| && (86)\\
&= |(u+v)^\top M (u-v)| && (87)\\
&\leq \underbrace{\|u+v\|_2}_{\leq 2} \|M\| \underbrace{\|u-v\|_2}_{\leq 1/4} && (88)\\
&\leq \frac{1}{2}\|M\| && (89)
\end{aligned}$$

Hence $\|M\| \leq 2u^\top M u \leq 2\sup_{u \in N} u^\top M u$ as desired. □

3.4 VC inequality and Symmetrization

In this section, we will see how a family of events with certain geometric structure (which we will quantify using VC-dimension) converges to its expectation at a rate dependent on the geometry. In the process, we will encounter the technique of ***symmetrization*** (Prof. Steinhardt calls it "bring your own randomness") used to add additional randomness which will be required to get concentration.

Let $\mathcal{H}$ be a collection of functions $f : \mathcal{X} \to \{0,1\}$ and $\{X_i \in \mathcal{X}\}_{i=1}^n \overset{\text{iid}}{\sim} p$. For $f \in \mathcal{H}$, let

$$\nu(f) = \mathbb{E}_{x \sim p}[f(x)] = \Pr_{x \sim p}[f(X) = 1] \tag{90}$$

$$\nu_n(f) = \frac{1}{n}\sum_{i=1}^n f(X_i) = \frac{1}{n}\#\{i : f(X_i) = 1\} \tag{91}$$

be the population and empirical averages respectively.

Question: How big is the discrepancy

$$D_n = \sup_{f \in \mathcal{H}} |\nu_n(f) - \nu(f)| \tag{92}$$

Easy case: $|\mathcal{H}| < \infty$. Since $f(X_i)$ is bounded, apply Hoeffding's inequality to the sum of independent bounded random variables to get:

$$D_n = \max_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - \mathbb{E}f(X)) \tag{93}$$

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \mathbb{E}f(X) \geq t\right] \leq \exp(-2nt^2) \tag{94}$$

A subsequent union bound over $|\mathcal{H}|$ reveals $t = O\left(\sqrt{\frac{1}{2n}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right)}\right)$

More common case: $|\mathcal{H}| = \infty$. In this situation, we will bound D_n using the geometry of $\mathcal{H}$. To do so, we will quantify the geometry using the following definitions:

Definition 39 (*Shattering number / VC dimension*)

The ***shattering number*** of $\mathcal{H}$ is

$$V_{\mathcal{H}}(\{x_i\}_{i=1}^n) = \#\text{ distinct}\{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{H}\} \tag{95}$$

$$V_{\mathcal{H}}(n) = \max_{|S|=n} V_{\mathcal{H}}(S) \tag{96}$$

It measures the number of possible ways to assign $\{0, 1\}$ labels to x_i which can be perfectly fit by $f \in \mathcal{H}$.

The ***VC dimension***

$$vc(\mathcal{H}) = \max\{n : V_{\mathcal{H}}(n) = 2^n\} \tag{97}$$

{eq:vc-dim-def}

It measures the largest cardinality n such that for any set of points S with cardinality $|S| = n$ and any $\{0, 1\}$ labelling of those points, some $f \in \mathcal{H}$ can perfectly fit it.

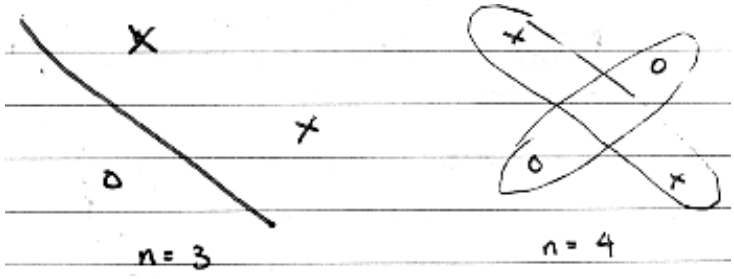
The shattering number is useful because instead of taking $\sup_{f \in \mathcal{H}}$ of a term involving f only through $\{f(X_i)\}_{i=1}^n$, we can instead take the supremum over $\{f(X_i)\}_{i=1}^n$ directly and only deal with $V_{\mathcal{H}}(n)$ terms.

Example 40 (*VC dimension of half spaces*)

{eg:half-spaces-vc}

Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H}$ = half spaces = $\{f(x) = \mathbb{1}[\langle v, x\rangle \geq \tau] : v \in \mathbb{R}^d, \tau \in \mathbb{R}\}$. Then $vc(\mathcal{H}) = d + 1$.

We will see a full proof later in Proposition 43, but for now consider an example where $d = 2$. We can always separate 3 points by drawing a line, so $vc(\mathcal{H}) \geq 3$. However, with 4 points there can be crossings (see Example 40) which cannot be shattered.



{fig:9-10-vc-hs}

Figure 7: $n = 3$ can always be shattered by a line, but the crossings possible when $n = 4$ prevent this.

Clearly by definition $V_{\mathcal{H}}(n) = 2^n$ for all $n \leq vc(\mathcal{H})$. When $n > vc(\mathcal{H})$, by Eq. (97) (Shattering number / VC dimension) we have $V_{\mathcal{H}}(n) < 2^n$. The following lemma quantifies this and shows that the shattering number is actually significantly smaller (growing polynomially in n rather than exponentially):

Lemma 41 (*Sauer-Shelah*)

{lem:sauer-shelah}

If $vc(\mathcal{H}) = d$, then $V_{\mathcal{H}}(n) \leq 2n^d$.

While we will use this without proof, Sauer-Shelah is the main reason why VC dimension is useful for us: it allows us to convert the infinite supremum over $f \in \mathcal{H}$ into a finite supremum over $O(n^{c(\mathcal{H})})$ many terms of the form $\{f(X_i)\}_{i=1}^d$.

Theorem 42 (*VC inequality*)

{thm:vc-inequality}

With probability $\geq 1-\delta$

$$D_n = O\left(\sqrt{\frac{vc(\mathcal{H}) + \log\frac{1}{\delta}}{n}}\right) \tag{98}$$

Proof. We will show something weaker, namely:

$$\mathbb{E}D_n \leq O\left(\frac{vc(\mathcal{H})\log n}{n}\right) \tag{99}$$

The $\log\frac{1}{\delta}$ tail bound follows from McDarmid's inequality, and removing the extra $\log n$ refines the argument we will give using a tool called chaining.

Incorrect proof path: Notice that

$$D_n = \sup_{f\in\mathcal{H}} \left| \underbrace{\frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}f(X))}_{\Pr[\cdot\geq t]\leq \exp(-2nt^2)} \right| \tag{100}$$

So Hoeffding's inequality can be used to control the term inside the supremum. Let $vc(\mathcal{H}) = d$. By Lemma 41, there are only $O(n^d)$ distinct $(f(X_1),\ldots,f(X_n))$ so a union bound implies $t = O\left(\sqrt{\frac{d\log n + \log\frac{1}{\delta}}{2n}}\right)$

This is incorrect because applying Sauer-Shelah requires us to condition on a specific realization of $\{X_i\}_{i=1}^n$ (after which we know there are at most $V_{\mathcal{H}}(n)$ distinct values of $(f(X_1),\ldots,f(X_n))$). After conditioning, there's no randomness left for applying Hoeffding's inequality to get concentration.

Solution: Introduce additional randomness using ***symmetrization***. Introduce independent copies X_i' and note

$$\mathbb{E}[D_n] = \mathbb{E}_{X_1,\ldots,X_n}\left[\sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)\right|\right] \tag{101}$$

$$= \mathbb{E}_{X_1,\ldots,X_n}\left[\sup_{f\in\mathcal{H}}\left|\mathbb{E}_{X_1',\ldots,X_n'}\left[\frac{1}{n}\sum_{i=1}^n (f(X_i) - f(X_i'))\right]\right|\right] \tag{102}$$

$|\cdot|$ is convex, so by Jensen's inequality

$$\mathbb{E}[D_n] \leq \mathbb{E}_X\left[\sup_{f\in\mathcal{H}}\mathbb{E}_{X'}\left[\left|\frac{1}{n}\sum_{i=1}^n (f(X_i) - f(X_i'))\right|\right]\right] \tag{103}$$

Also, $\sup_y \mathbb{E}f(X,y) \leq \mathbb{E}\sup_y f(X,y)$ for any function f (since $\mathbb{E}f(X,y) \leq \mathbb{E}\sup_y f(X,y)$ then take supremum on left-hand side, or see Fatou-Lebesgue theorem) hence we can move $\mathbb{E}_{X'}$ out of $\sup_{f\in\mathcal{H}}$ to get

$$\mathbb{E}[D_n] \leq \mathbb{E}_{X,X'}\left[\sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i) - f(X_i'))\right|\right] \tag{104}$$

Here is where the randomness from symmetrization is added: since $f(X_i) - f(X_i') \overset{d}{=} \varepsilon_i(f(X_i) - f(X_i'))$ for $\varepsilon_i \sim \text{Rad}$

$$\mathbb{E}[D_n] \leq \mathbb{E}_{X,X',\varepsilon}\left[\sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i) - f(X_i'))\right|\right] \tag{105}$$

{{eq:9-10-symmetrized-need-concentration}}

Condition on X, X' and let $f(X_i) = a \in V_{\mathcal{H}}(\{x_1, \ldots, x_n\})$ and $f(X_i') = b \in V_{\mathcal{H}}(\{x_1', \ldots, x_n'\})$. Then

$$\sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i) - f(X_i'))\right| = \sup_{a,b}\underbrace{\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(a_i - b_i)\right|}_{\Pr[|\cdot|\geq t]\leq 2\exp(\frac{-nt^2}{2})} \tag{106}$$

Now we can apply Hoeffding's inequality (picking up an extra factor of 2 because of the absolute value, see Eq. (72)) to the independent, zero-mean (since $\mathbb{E}\varepsilon_i = 0$), bounded (since a_i, b_i, and ε_i are all bounded) random (since ε_i is still random) variables and union bound over the $O(n^{2d})$ (by Sauer-Shelah, squared since there is both $f(X)$ and $f(X')$) distinct $f(X)$ and $f(X')$

$$\Pr\left[\sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i) - f(X_i'))\right| \geq t \middle| X, X'\right] \leq (2n^{2d})2\exp\left(\frac{-nt^2}{2}\right) \tag{107}$$

$$\tag{108}$$

This tail probability is small if $t \gg \sqrt{\frac{d\log n}{n}}$, so the expectation over ε in Eq. (105) is of the same order and we have

why?? Try $\mathbb{E}X = \int P(X \geq t)dt$ for $X \geq 0$

$$\mathbb{E}[D_n] \leq \mathbb{E}_{X,X'}\left[\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i) - f(X_i'))\right| X, X'\right]\right] = O\left(\sqrt{\frac{d\log n}{n}}\right) \tag{109}$$

□

Discretization to a representative set ("fingerprinting") is how previous sections worked. The complication here is that to apply Lemma 41 we had to condition on X_i and remove the randomness. The secret sauce was to add randomness back using the ε_i in symmetrization.

4 9/12/2019

4.1 Recap

- Bounded $\mathbb{E}\sup_{v\in V} X(v)$ where $X(v)$ concentrates and V is finite or could be well approximated by a finite set
  - Top eigenvalue of random covariance matrix
  - VC inequality and symmetrization
- Debt: VC-dim of halfspaces is $d+1$ (Example 40)

Today, we will:

- Pay off debt: prove the VC dimension of half spaces is $d+1$
- Give a finite-sample analysis of Definition 1 (Minimum distance functional)
  - Weaken TV to $\widetilde{\text{TV}}$
  - Bound Modulus of continuity bound via "mean crossing lemma"
  - $\widetilde{\text{TV}}(\tilde{p}, \tilde{p}_n) \to 0$ as $n \to \infty$

4.2 VC dimension of half spaces

In Example 40 (VC dimension of half spaces) we claimed that $vc(\mathcal{H}) = d + 1$ for the family of half spaces (i.e. linear decision boundaries)

$$\mathcal{H} = \{\mathbb{1}\{\langle v, x\rangle \geq \tau\} : v \in \mathbb{R}^d, \tau \in \mathbb{R}\} \tag{110}$$

We previously showed it geometrically for the case when $d = 2$. Here, we will generalize this to higher dimensions.

Proposition 43 (*VC dimension of half spaces*)

{prop:vc-dim-half-spaces}

No $d + 2$ set of points in $\mathbb{R}^d$ can be shattered by any $f \in \mathcal{H}$.

Proof. Fix $\{x_i\}_{i=1}^{d+2} \in \mathbb{R}^d$ distinct. We will find two sets $S_+, S_- \subset \{x_1, \ldots, x_{d+2}\}$ such that $S_+ \cap S_- = \emptyset$ but $\text{conv}(S_+) \cap \text{conv}(S_-) \neq \emptyset$. This is sufficient because every $f = \mathbb{1}\{\langle v, x\rangle \geq \tau\} \in \mathcal{H}$ can be identified with a half-space (of the points classified $+1$ by f)

$$H = f^{-1}(\{1\}) = \{x \in \mathbb{R}^d : \langle v, x\rangle \geq \tau\} \tag{111}$$

and by convexity of H

$$S_+ \subset H \implies \text{conv}(S_+) \subset H \tag{112}$$

Hence, if f correctly classifies all of S_+ then it must also misclassify some $x \in S_+ \cap S_- \subset S_-$.

Consider the linear system

$$\sum_{i=1}^{d+2} a_i x_i = 0, \qquad \sum_{i=1}^{d+2} a_i = 0 \tag{113}$$

{eq:9-12-lin-system}

or equivalently in matrix form

$$\underbrace{\begin{bmatrix} \vdots & \vdots & \ldots & \vdots \\ x_1 & x_2 & \ldots & x_{d+2} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 1 & \ldots & 1 \end{bmatrix}}_{(d+1)\times(d+2)} \begin{bmatrix} a_1 \\ \vdots \\ a_{d+2} \end{bmatrix} = \mathbf{0} \tag{114}$$

By the rank-nullity theorem, the null-space must have dimension ≥ 1 hence there exists at least one solution $\boldsymbol{a}$. Let

$$S_+ = \{i : a_i > 0\}, \qquad S_- = \{i : a_i < 0\} \tag{115}$$

Then by Eq. (113)

$$\underbrace{\sum_{i \in S_+} \underbrace{\frac{a_i}{A}}_{\in[0,1]} x_i}_{\in \text{conv}(S_+)} = \underbrace{\sum_{i \in S_-} \underbrace{\frac{a_i}{A}}_{\in[0,1]} x_i}_{\in \text{conv}(S_-)} \qquad \text{where} \quad A = \sum_{i \in S_+} a_i = \sum_{i \in S_-} (-a_i) \tag{116}$$

This gives us a point in $\text{conv}(S_+) \cap \text{conv}(S_-)$. □

Remark 44. The geometric result that "any set of $d + 2$ points in $\mathbb{R}^d$ can be partitioned into two disjoint sets whose convex hulls intersect" is known as ***Radon's theorem*** on convex sets.

4.3 Finite sample analysis of MDF via Generalized KS distance

Recall Definition 1 (Minimum distance functional) projects $\tilde{p}$ on to $\mathcal{G}$ under some discrepancy D. Previously we worked with $D = \mathrm{TV}$, which works fine if $\tilde{p}$ is a continuous distribution (e.g. $\tilde{p} = \mathcal{N}(\mu, I)$ in Lemma 3). However, when we only have a finite number of samples we can only form the empirical distribution

$$\tilde{p}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \qquad X_i \sim \tilde{p} \tag{117}$$

Here, TV is inadequate because $\mathrm{TV}(\tilde{p}_n, p) = 1$ almost surely for any continuous distribution p (this is because $\Pr_{X\sim p}[X = X_i] = 0$) so it's not clear how to project onto a continuous family such as $\mathcal{G}_{gauss}$. Moreover, in many cases $\mathrm{TV}(\tilde{p}_n, \tilde{p}) = 1$ even as $n \to \infty$.

To address this issue, we can consider relaxing TV to a weakening $\widetilde{\mathrm{TV}}$ which is more forgiving. We have two desidirata for $\widetilde{\mathrm{TV}}$:

1. The modulus $\mathfrak{m}(\mathcal{G}, \varepsilon, \widetilde{\mathrm{TV}})$ remains small, so that Proposition 2 (Modulus of continuity bound) still gives a good result
2. $\widetilde{\mathrm{TV}}(\tilde{p}, \tilde{p}_n) \to 0$ as $n \to \infty$, so that $\widetilde{\mathrm{TV}}$ detects convergence of (discrete) empirical distributions to a (possibly continuous) population distribution

Remark 45. The two desidirata are competing. We want $\widetilde{\mathrm{TV}}$ to be large in (1) so that $A = \{(p, q) \in \mathcal{G} : \widetilde{\mathrm{TV}}(p, q) \le \varepsilon\}$ is small and hence $\mathfrak{m} = \sup_{(p,q)\in A} L(p, \theta^*(q))$ is small. At the same time, in (2) we would like $\widetilde{\mathrm{TV}}$ to be small to avoid the failure of TV in detecting $\tilde{p_n} \to \tilde{p}$ (e.g. Glivenko-Cantelli ensures that the cumulative distribution functions converge uniformly).

Proposition 46

{prop:mdf-tilde-tv}

Suppose $\tilde{\mathrm{TV}}$ *is a pseudometric such that* $\widetilde{\mathrm{TV}} \le \mathrm{TV}$. *Let* $\hat{\theta}_{\widetilde{\mathrm{TV}}}(p) = \theta^*(q)$ *where* $q \in \operatorname{argmin}_{q\in\mathcal{G}} \widetilde{\mathrm{TV}}(p, q)$ *(the Minimum distance functional under* $\widetilde{\mathrm{TV}}$*). Then*

$$L(p^*, \hat{\theta}_{\widetilde{\mathrm{TV}}}(\tilde{p}_n)) \le \mathfrak{m}(\mathcal{G}, 2\varepsilon', \widetilde{\mathrm{TV}}) \tag{118}$$

where $\varepsilon' = \varepsilon + \widetilde{\mathrm{TV}}(\tilde{p}, \tilde{p}_n)$ *(and* $\widetilde{\mathrm{TV}}(p^*, \tilde{p}) \le \varepsilon$ *as per the conventions outlined in Fig. 1)*

Proof. By Proposition 2 (Modulus of continuity bound)

$$L(p^*, \hat{\theta}_{\widetilde{\mathrm{TV}}}(\tilde{p}_n)) \le \mathfrak{m}(\mathcal{G}, 2\widetilde{\mathrm{TV}}(p^*, \tilde{p}_n), \widetilde{\mathrm{TV}}, L) \tag{119}$$

Since $\widetilde{\mathrm{TV}}$ is a pseudometric, by the triangle inequality

$$\widetilde{\mathrm{TV}}(p^*, \tilde{p}_n) \le \underbrace{\widetilde{\mathrm{TV}}(p^*, \tilde{p})}_{\le\varepsilon} + \widetilde{\mathrm{TV}}(\tilde{p}, \tilde{p}_n) \tag{120}$$

□

Remark 47. The change from ε to ε' in Proposition 46 is why the second desidirata is relevant. We will see that in particular it will shift us to consider $\mathcal{G}_{\mathrm{TV}}(\rho, \varepsilon') \subsetneq \mathcal{G}_{\mathrm{TV}}(\rho, \varepsilon)$ when we later bound the modulus.

We will consider the following weakening of total variation distance:

Definition 48 (*Generalized Kolmogorov-Smirnov distance*)

{def:tilde-tv}

For a family of functions $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R}\}$, the ***generalized Kolmogorov-Smirnov distance*** induced by $\mathcal{H}$ is

$$\widetilde{\mathrm{TV}}_{\mathcal{H}}(p, q) = \sup_{f\in\mathcal{H}, \tau\in\mathbb{R}} \left| \Pr_p[f(X) \ge \tau] - \Pr_q[f(X) \ge \tau] \right| \tag{121}$$

Remark 49. For $f \in \mathcal{H}$ and $\tau \in \mathbb{R}$, if we define the event $E_{f,\tau} = \{f(X) \geq \tau\}$ then notice

$$\widetilde{\mathrm{TV}}_{\mathcal{H}}(p,q) = \sup_{E_{f,\tau}} \left| \Pr_p[E_{f,\tau}] - \Pr_q[E_{f,\tau}] \right| \leq \sup_{E \text{ meas}} \left| \Pr_p[E] - \Pr_q[E] \right| = \mathrm{TV}(p,q) \tag{122}$$

So $\widehat{\mathrm{TV}}$ is indeed dominated by TV as required by Proposition 46.

What $\mathcal{H}$ should we pick? The answer depends on what we are trying to estimate (i.e. choice of $L(p,\theta)$). For now, consider mean estimation (i.e. $L(p,\theta) = \|\theta - \mu(p)\|_2$). Motivated by the fact that knowledge of the one dimensional projections ($\mathbb{E}\langle v, x\rangle$ for all $v \in \mathbb{R}^d$) allows us to determine $\mathbb{E}[X]$, we consider

$$\mathcal{H} = \mathcal{H}_{lin} = \{x \mapsto \langle v, x\rangle : v \in \mathbb{R}^d\} \tag{123}$$

To bound the modulus, recall that resilient distributions were convenient for $D = \mathrm{TV}$ because if $p, q \in \mathcal{G}_{\mathrm{TV}}(\rho,\varepsilon)$ then Corollary 11 (Modulus of continuity bound for resilient distributions) gave us

$$\mathrm{TV}(p,q) \leq \varepsilon \implies \|\mu(p) - \mu(q)\|_2 \leq 2\rho \tag{124}$$

Here, we will also restrict attention to resilient distributions $\mathcal{G} \subset \mathcal{G}_{\mathrm{TV}}$. To satisfy desidirata 1, we generalize this result to $D = \widetilde{\mathrm{TV}}$ in the following way:

Proposition 50 (*Generalized modulus bound*)

{eq:9-12-desidirata-1}

If $p, q \in \mathcal{G} \subset \mathcal{G}_{\mathrm{TV}}(\rho,\varepsilon')$ and $\widetilde{\mathrm{TV}}(p,q) \leq \varepsilon'$, then

$$\|\mu(p) - \mu(q)\|_2 \leq 2\rho \tag{125}$$

In other words, $\mathfrak{m}(\mathcal{G}, \varepsilon', \widetilde{\mathrm{TV}}) \leq 2\rho$

However, as shown in Proposition 46, we have that $\varepsilon' = \varepsilon + \widetilde{\mathrm{TV}}(\tilde{p}, \tilde{p}_n)$ and our framework (Fig. 1) only assumes $D(\tilde{p}, p^*) = \widetilde{\mathrm{TV}}(\tilde{p}, p^*) \leq \varepsilon$. The requirement of being (ρ,ε')-resilient is stronger than that of (ρ,ε)-resilience, so to ensure broad applicability we would like desidirata 2 formalized as follows:

Proposition 51

{eq:9-12-desidirata-2}

$\widetilde{\mathrm{TV}}(\tilde{p}, \tilde{p}_n)$ is small, specifically:

$$\widetilde{\mathrm{TV}}(\tilde{p}, \tilde{p}_n) = O\left(\sqrt{d/n}\right) \tag{126}$$

Proof of Proposition 50. Previously we used Midpoint lemma to find an ε-deletion $r \leq \min\left\{\frac{p}{1-\varepsilon}, \frac{q}{1-\varepsilon}\right\}$ close to both p and q in the sense that

$$\|\mu(p) - \mu(r)\|_2 \leq \rho \text{ and } \|\mu(q) - \mu(r)\|_2 \leq \rho \tag{127}$$

After which a triangle inequality completed the proof.

Unfortunately, we don't know of a way to find a single midpoint distribution under $\widetilde{\mathrm{TV}}$. Instead, we will use the following key property:

Lemma 52 (*Mean crossing property*)

{lem:mean-crossing-property}

Suppose $\widetilde{\mathrm{TV}}(p,q) \leq \varepsilon$. For any $v \in \mathbb{R}^d$, there exists ε-deletions $r_p \leq \frac{p}{1-\varepsilon}$ and $r_q \leq \frac{q}{1-\varepsilon}$ such that

$$\mathbb{E}_{r_q}\langle v, x\rangle \leq \mathbb{E}_{r_p}\langle v, x\rangle \tag{128}$$

In other words, after deleting ε mass to create r_q and r_p, the means are shifted such that they cross.

If we have the ϵ deletions $r_p \le \frac{p}{1-\varepsilon}$ and $r_q \le \frac{q}{1-\varepsilon}$ from Lemma 52 (Mean crossing property), then

$$\underbrace{\mathbb{E}_p \langle v, x\rangle}_{=\langle v,\mu_p\rangle} \le \mathbb{E}_{r_p}[\langle v, x\rangle] + \rho \qquad \text{resilience of } p \tag{129}$$

$$\le \mathbb{E}_{r_q}[\langle v, x\rangle] + \rho \qquad \text{mean crossing} \tag{130}$$

$$\le \underbrace{\mathbb{E}_q[\langle v, x\rangle]}_{=\langle v,\mu_q\rangle} + 2\rho \qquad \text{resilience of } q \tag{131}$$

Hence

$$\langle v, \mu_p - \mu_q\rangle \le 2\rho \tag{132}$$

for all $\|v\|_2 = 1$. Therefore $\|\mu_p - \mu_q\|_2 \le 2\rho$. □

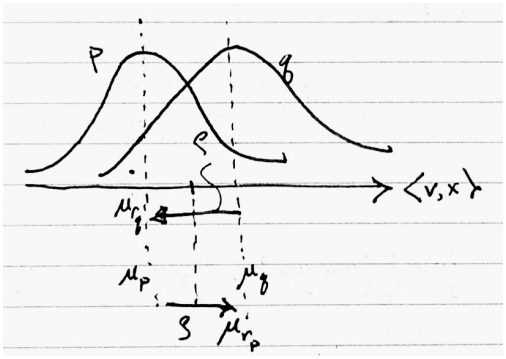


Figure 8: Resilience allows us to perform an ε-deletion to move from $\mu_p \to \mu_{r_p}$ and $\mu_q \to \mu_{r_q}$ and pick up a factor of $+2\rho$. Mean crossing allows us to relate μ_{r_p} and μ_{r_q}. {fig:resilience-mean-crossing}

Proof of Mean crossing property. Consider Fig. 8, which visualizes the 1D projections of p and q in the v direction. To make $\langle v, \mu_{r_q}\rangle$ cross over $\langle v, \mu_{r_p}\rangle$, we would like to shift the mean of q to the left and the mean of p to the right as much as possible. Thus, delete ε mass from the right tail of q (and delete the left tail of p). Then

$$\Pr_{r_p}[\langle v, x\rangle \ge \tau] \ge \frac{\Pr_p[\langle v, x\rangle \ge \tau]}{1-\varepsilon} \ge \frac{\Pr_q[\langle v, x\rangle \ge \tau] - \varepsilon}{1-\varepsilon} = \Pr_{r_q}[\langle v, x\rangle \ge \tau] \tag{133}$$

where the first inequality is because r_p is p with the left tail deleted and renormalized by $1-\varepsilon$, the second from $\Pr_q[\langle v, x\rangle \ge \tau] - \Pr_p[\langle v, x\rangle \ge \tau] \le \widehat{\mathrm{TV}}(p, q) \le \varepsilon$, and the third from r_q being formed by deleting ε from the right tail of q and renormalizing by $1-\varepsilon$.

We have shown that the right tail probabilities of r_p are always larger than those of r_q, i.e. r_p ***stochastically dominates*** r_q. As a consequence, $\mathbb{E}_{r_p}[\langle v, x\rangle] \ge \mathbb{E}_{r_q}[\langle v, x\rangle]$. □

Proof of Proposition 51. Notice

$$\widehat{\mathrm{TV}}_{\mathcal{H}_{lin}}(p, q) = \sup_{v\in\mathbb{R}^d, \tau\in\mathbb{R}} \left| \underbrace{\Pr_p[\langle v, x\rangle \ge \tau] - \Pr_q[\langle v, x\rangle \ge \tau]}_{\text{max discrepancy on halfspaces}} \right| \tag{134}$$

By VC inequality and Proposition 43 (VC dimension of half spaces)

$$\widehat{\mathrm{TV}}_{\mathcal{H}_{lin}}(\tilde{p}, \tilde{p}_n) \le O\left(\sqrt{\frac{vc(\text{half spaces})}{n}}\right) = O\left(\sqrt{\frac{d + \log\frac{1}{\delta}}{n}}\right) \tag{135}$$

with probability $\ge 1-\delta$. □

Consequences:

- Combining Proposition 50 and Proposition 51, we have that for $(\rho, \varepsilon' = \varepsilon + \widetilde{\mathrm{TV}}_{\mathcal{H}_{lin}}(\tilde{p}, \tilde{p}_n) = \varepsilon + O(\sqrt{d/n}))$-resilient distributions, we can estimate mean with error 2ρ
- For bounded covariance, Example 9 gave us $\rho(\varepsilon) = O(\sqrt{\varepsilon})$ hence

$$L(p^*, \tilde{\theta}_{\widetilde{\mathrm{TV}}}(\tilde{p}_n)) \leq O\left(\sqrt{\varepsilon + \sqrt{d/n}}\right) \tag{136}$$

  The lower bound $\sqrt{\varepsilon}$ is what we get in the infinite sample $n \to \infty$ limit, and $\sqrt{d/n}$ when $\varepsilon \to 0$, so we would like $\sqrt{\varepsilon} + \sqrt{d/n}$. The slack in the bound comes from requiring $n \gg d/\varepsilon^2$ for it to hold with high probability, whereas we would need $n \gg \frac{d}{\varepsilon}$ if we wanted to show the tighter bound

- For sub-Gaussians, $\rho(\varepsilon) = O(\varepsilon\sqrt{\log(1/\varepsilon)})$. When $n \gg \frac{d}{\varepsilon^2}$ we get $O(\varepsilon\sqrt{\log(1/\varepsilon)})$.

In general, this analysis holds for $n \gg d/\varepsilon^2$: whenever this holds, we can do as well ($O(\sqrt{\varepsilon + o(\varepsilon)}) = O(\sqrt{\varepsilon})$) as if we had infinite data. The analysis is tight in d but loose in ε.

5 9/17/2019

5.1 Outline

The Minimum distance functional enjoys strong robustness bounds such as Proposition 2 (Modulus of continuity bound). However, its definition involves performing a projection onto $\mathcal{G}$ (the set of distributions which p^* is assumed to be contained in):

$$\hat{\theta}(\tilde{p}) = \theta^*(q) = \min_{\theta} L(q, \theta) \text{ where } q = \mathrm{argmin}_{q \in \mathcal{G}} D(\tilde{p}, q) \tag{137}$$

We saw last time $D = \mathrm{TV}$ was not suitable if $\tilde{p} = \tilde{p}_n$ is discrete, motivating the use of Generalized Kolmogorov-Smirnov distance. For bounded kth moments, we have that $\rho = \mathcal{O}(\sigma_k \varepsilon^{1-1/k})$ which under our previous theory (Proposition 46 and Proposition 51) yields a guarantee

$$\|\mu(\tilde{p}_n) - \mu(p^*)\|_2 \leq \mathcal{O}\left(\left(\underbrace{\varepsilon + \sqrt{\frac{d}{n}}}_{\varepsilon'}\right)^{1-1/k}\right) \tag{138}$$

whenever $\widetilde{\mathrm{TV}}_{\mathcal{H}_{lin}}(\tilde{p}, p^*) \leq \varepsilon$.

Today, we consider an alternative solution where we expand $\mathcal{G}$ to some larger set $\mathcal{M}$ to perform the projection:

$$q = \mathrm{argmin}_{q \in \mathcal{M}} \tilde{D}(\tilde{p}, q) \tag{139}$$

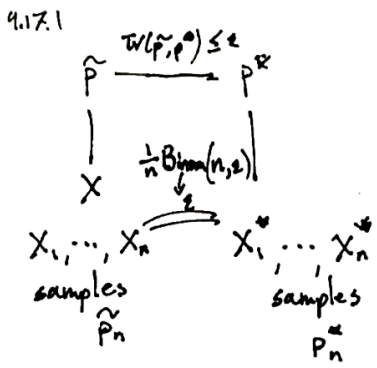
Under this analysis, we can achieve a tighter $\mathcal{O}\left(\varepsilon^{1-1/k} + \sqrt{d/n}\right)$ error.

Outline for today:

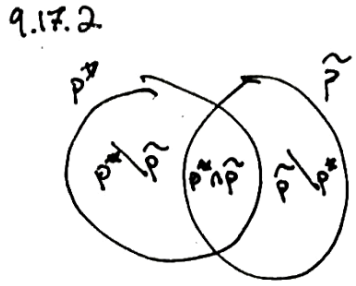
- True "empirical distribution"
- Expand the set idea
- Analyze concentration for bounded kth moments
  - symmetrization
  - truncated moments
  - ledoux-talagrand

5.2 True Empirical Distribution

Let p_n^* define an empirical distribution drawn from p^*.



Issue: No overlap between $\tilde{p}_n$, p_n^*
Solution: Define ***coupling*** between p_n^* and $\tilde{p}_n$.



- With probability $1-\varepsilon$:
  - Sample from $p^* \cap \tilde{p}$, $X_i = \tilde{X}_i$ = sample
- And with probability ε:
  - Sample X_i^* from $p^* \setminus \tilde{p}$
  - Sample X_i from $\tilde{p} \setminus p^*$

Takeaway: $\underbrace{\mathrm{TV}(\tilde{p}_n, p_n^*)}_{\tilde{\varepsilon}} \sim \frac{1}{n}\mathrm{Binom}(n, \varepsilon)$

Lemma 53 (*Tail bound for binomials*)

With probability $\geq 1-\delta$

$$\frac{1}{n}Binom(n,\varepsilon) \leq O\left(\sqrt{\varepsilon} + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right)^2 = O\left(\varepsilon + \frac{\log\frac{1}{\delta}}{2n}\right) \tag{140}$$

Remark 54. This is tighter than Hoeffding, which would have given $\exp(-\varepsilon^2 n/3)$. Need Bernstein's inequality and Chernoff bound for binomial random variables to prove this.

5.3 Finite-Sample Concentration via Expanding the Set

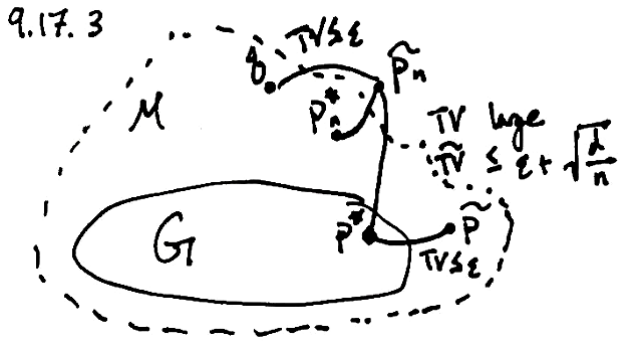


Figure 9: If we can expand $\mathcal{G} \subset \mathcal{M}$ so that $p_n^* \in \mathcal{M}$, then we can form $q = \min_{q \in \mathcal{M}} \mathrm{TV}(\tilde{p}_n, q)$ by projecting $\tilde{p}_n$ onto $\mathcal{M}$ and use the "true empirical distribution" p_n^* to connect q with p^*. This is made precise in Proposition 55

Need three properties for $\mathcal{M}$

- $\mathcal{M}$ large: $p_n^* \in \mathcal{M}$ whp.
- $\mathcal{M}$ small: modulus $\mathfrak{m}(\mathcal{M}_\varepsilon)$ small
- p_n^* good approx to p^*: $\|\mu(p^*) - \mu(p_n^*)\|_2$ bounded

Proposition 55

{prop:projection-bound-exp and-G-to-M}

Suppose

- $p_n^* \in \mathcal{M}$ *wp* $1 - \delta$
- $\mathrm{TV}(p_n^*, \tilde{p}_n) \le \tilde{\varepsilon}$ *wp* $1 - \delta$

Then projecting onto $\mathcal{M}$ *yields* q *where*

$$\|\mu(q) - \mu(p^*)\|_2 \le \mathfrak{m}(\mathcal{M}, 2\tilde{\varepsilon}) + \|\mu(p^*) - \mu(p_n^*)\|_2 \tag{141}$$

wp $1 - 2\delta$.

Proof. Since $p_n^* \in \mathcal{M}$, we have

$$\mathrm{TV}(\tilde{p}_n, q) = \min_{q \in \mathcal{M}} \mathrm{TV}(\tilde{p}_n, q) \le \tilde{\varepsilon} \tag{142}$$

Also by hypothesis $\mathrm{TV}(\tilde{p}_n, p_n^*) \le \tilde{\varepsilon}$, so by triangle inequality

$$\mathrm{TV}(p_n^*, q) \le 2\tilde{\varepsilon} \tag{143}$$

Together we have $\|\mu(p_n^*) - \mu(q)\|_2 \le \mathfrak{m}(\mathcal{M}, 2\tilde{\varepsilon})$ and by triangle inequality

$$\|\mu(p^*) - \mu(q)\|_2 \le \mathfrak{m}(\mathcal{M}, 2\tilde{\varepsilon}) + \|\mu(p^*) - \mu(p_n^*)\|_2 \tag{144}$$

□

5.4 Expanding bounded kth moments to set of resilient distributions

The following example will be our running example for this section. We will see how bounded kth moments may require n to be too large, and how we can expand to the larger set of resilient distributions.

Example 56 (*Bounded kth moments*)

Consider distributions with bounded kth moments, that is

$$\mathcal{G} = \mathcal{G}_k(\sigma) = \{p : |\mathbb{E}X|_{\psi_k} \le \sigma\} \tag{145}$$

$$= \{p : \mathbb{E}_p[|\langle X - \mu, v\rangle|^k] \le \sigma_k^k \ \forall \|v\|_2 \le 1\} \tag{146}$$

where $\psi_k(x) = x^k$. For example, $\mathcal{G}_2(\sigma)$ are the distributions with bounded covariance.

Isuse: $p_n^* \notin \mathcal{G}$ until $n \gg d^{k/2}$. For example, let $p^* = \mathcal{N}(\mu, I)$, $p_n^* = \sum_{i=1}^n \frac{1}{n}\delta_{x_i}$, and notice for $v = \frac{x_1-\mu}{\|x_1-\mu\|}$ we have $\|v\|_2 = 1$ but

$$\mathbb{E}_{p_n^*}[|\langle X - \mu, v\rangle|^k] \ge \frac{1}{n}|\langle x_1 - \mu, v\rangle|^k = \frac{1}{n}\underbrace{\|X_1 - \mu\|_2^k}_{=\mathcal{O}(\sqrt{d})} \asymp \frac{1}{n}d^{k/2} \tag{147}$$

$$\mathbb{E}_{p_n^*}[|\langle X - \mu, v\rangle|^k]^{1/k} \asymp \left(\frac{1}{n}d^{k/2}\right)^{1/k} = \frac{\sqrt{d}}{n^{1/k}} \tag{148}$$

Asymptotically, we see that we need $n \gg d^{k/2}$ for the kth moments to remain bounded.

Figure 10: The moment $\langle X - \mu, v\rangle$ along a single direction of a sample $v = \frac{x_1-\mu}{\|x_1-\mu\|}$ is large, need to average over many samples before it washes out.

Consider expanding bounded kth moments $\mathcal{G} = \mathcal{G}_k(\sigma)$ to the larger set of resilient distributions $\mathcal{M} = \mathcal{G}_{\mathrm{TV}}(\rho, \varepsilon)$ with $\rho = O(\varepsilon^{1-1/k})$. We already have a modulus bound $\mathfrak{m}(\mathcal{M}, \varepsilon) \le 2\rho = O(\varepsilon^{1-1/k})$ from Corollary 11 (Modulus of continuity bound for resilient distributions), so to make Proposition 55 meaningful it remains to show:

- Bound $\|\mu(p^*) - \mu(p_n^*)\|_2 = O\left(\sigma\sqrt{\frac{d}{n}\delta^{-1/k}}\right)$ We do this using Kintchine's inequality.
- $p_n^* \in \mathcal{M}$ whp. We do this using truncated moments.

Lemma 57

$$\|X\|_2 = \mathbb{E}_{v\sim\mathcal{N}(0,I)}[|\langle x, v\rangle|]\sqrt{\frac{\pi}{2}} \tag{149}$$

Proof.

$$\mathbb{E}[|\langle(\|x\|_2, 0, \ldots, 0), (v_1, \ldots, v_d)\rangle|] = \mathbb{E}[|v_1| \cdot \|x\|_2] \tag{150}$$

$$\mathbb{E}[|v_1|] = \sqrt{2/\pi} \tag{151}$$

□

Remark 58. There's a better version of the above called ***Khintchine's inequality***:

$$\|X\|_2 \le \sqrt{2}\mathbb{E}[|\langle X, \varepsilon\rangle|] \tag{152}$$

with $\varepsilon \sim$ Rad. So we can just test using Rademachers rather than Gaussian process.

We failed to prove the second easy thing in class using above, see next lecture for resolution

5.4.1 Truncated moments

Now we show $p_n^* \in \mathcal{M}$, introducing some new ideas along the way.

Problem: $\|\cdot\|_{\psi_k}$ is not small.

Solution: Truncate moments, replace $\psi_k(x) = x^k$ with

$$\tilde{\psi}_k(x) = \begin{cases} x^k & x \le x_0 \\ kx_0^{k-1}(x - x_0) + x_0^k & x > x_0 \end{cases} \tag{153}$$

This is equal to ψ_k for $x \le x_0$ and linearly interpolates beyond $x \ge x_0$.

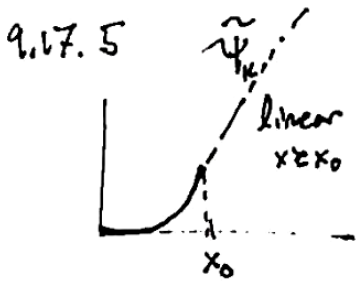


Figure 11: The Orlicz function $\tilde{\psi}_k$ used for truncating moments {fig:tilde-psi-k}

$\tilde{\psi}_k$ is L-Lipschitz with $L = kx_0^{k-1}$, so in particular if we choose $x_0 = \left(\frac{1}{\varepsilon}\right)^{1/k}$ then we have $L = k/\varepsilon^{1-1/k}$.

Proposition 59 {prop:truncated-moments-bound}

Let $X_1, \dots, X_n \sim p^*$, where $p^* \in \mathcal{G}_k(\sigma)$. Then

$$\mathbb{E}_{X_i \sim p^*}\left[\sup_{\|u\|_2=1} \frac{1}{n}\sum_{i=1}^n \tilde{\psi}_k\left(\left|\frac{\langle X_i - \mu, v\rangle}{\sigma}\right|\right)\right] \le 1 + 2L\sqrt{\frac{d}{n}} \tag{154}$$

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

Remark 60. When $n \ge 4L^2 d = 4k^2 d/\varepsilon^{2-2/k}$, we have

$$\sup_{\|u\|_2=1} \frac{1}{n}\sum_{i=1}^n \tilde{\psi}_k\left(\left|\frac{\langle X_i - \mu, v\rangle}{\sigma}\right|\right) \le 2 \tag{155}$$

This implies that p_n^* has bounded Orlicz norm $\|p_n^*\|_{\tilde{\psi}_k}$, so by Proposition 15 p_n^* is resilient with parameter $\sigma\varepsilon\tilde{\psi}^{-1}(2/\varepsilon)$. So we really need to control how fast $\sigma\varepsilon\tilde{\psi}^{-1}(2/\varepsilon)$ grows. From Fig. 12, we may conclude

$$\sigma\varepsilon\tilde{\psi}^{-1}(2/\varepsilon) \le 2\sigma\varepsilon \underbrace{\left(\frac{1}{\varepsilon}\right)^{1/k}}_{>0} = 2\sigma\varepsilon^{1-1/k} \tag{156}$$

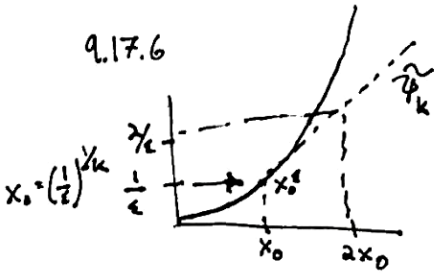


Figure 12: A proof by picture why $\psi^{-1}(2/\varepsilon) \leq 2x_0 = 2\varepsilon^{-1/k}$

{fig:bound-tilde-psi-k-inv}

5.4.2 Ledoux-Talagrand contraction

{ssec:ledoux-talagrand-first}

This result is used as part of symmetrization arguments. If I have already symmetrized and I have a Lipschitz function, than I can always repalce the function with just its arguments and make things bigger.

Theorem 61 (*Ledoux-Talagrand*)

$$\mathbb{E}_\varepsilon \left[\sup_{v \in V} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \phi(\langle x_i, v \rangle) \right] \leq \mathbb{E}_\varepsilon \left[\sup_{v \in V} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \langle x_i, v \rangle \right] \tag{157}$$

for ϕ 1-Lipschitz, i.e. $|\phi(x) - \phi(y)| \leq |x - y|$, V a symmetric set, and $\varepsilon \sim$ Rad.

Proof of Proposition 59.

$$\mathbb{E}_{X_i \sim p^*} \left[\sup_{\|u\|_2 = 1} \frac{1}{n} \sum_{i=1}^n \tilde{\psi}_k \left(\left| \frac{\langle X_i - \mu, v \rangle}{\sigma} \right| \right) \right] = \underbrace{\mathbb{E}[\tilde{\psi}]}_{\leq \mathbb{E}\psi \leq 1} + \sup_{\|v\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^n \left(\tilde{\psi}_k(|\langle x_i - \mu, v \rangle / \sigma|) - \mathbb{E}[\tilde{\psi}] \right) \tag{158}$$

Symmetrize?

$$\mathbb{E}_{X_i, X_i' \sim p^*, \varepsilon \sim \text{Rad}} \left[\sup_{\|v\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \left(\tilde{\psi}_k \left(\frac{|\langle X_i - \mu, v \rangle|}{\sigma} \right) - \tilde{\psi}_k \left(\frac{|\langle X_i' - \mu, v \rangle|}{\sigma} \right) \right) \right] \tag{159}$$

□ See next lecture

6 9/19/2019

6.1 Recap

- Expand $\mathcal{G}$ to $\mathcal{M}$
  - Bound modulus of $\mathcal{M}$
  - Show $p_n^* \in \mathcal{M}$
  - Bound $\|\mu(p_n^*) - \mu(p^*)\|_2$
- $\mathcal{G} = \mathcal{G}_k(\sigma) =$ bounded kth moments (needed $n \geq d^{k/2}$ samples if $\mathcal{M} = \mathcal{G}$)
- $\mathcal{M} = \mathcal{G}_{\text{TV}}(\rho, \varepsilon) = (\rho, \varepsilon)$-resilient distributions with $\rho = \mathcal{O}(\sigma \varepsilon^{1-1/k})$

6.2 Truncated moments bounds

Our strategy to show $p_n^* \in \mathcal{M}$ is to consider the truncated (Orlicz) function

$$\tilde{\psi}_k = \begin{cases} x^k & \text{if } x \le x_0 \\ k x_0^{k-1}(x - x_0) + x_0^k & \text{if } x > x_0 \end{cases} \tag{160}$$

This function behaves as x^k until $x = x_0$, after which it is linear. See Fig. 11. Note that $\tilde{\psi}_k$ is L-Lipschitz with $L = k x_0^{k-1}$.

Todo this lecture:

- Ledoux-Talagrand
- Bound $\|\mu(p_n^*) - \mu(p^*)\|_2$ via Khintchine and Rosenthal
- Show truncated moments $\tilde{\psi}$ concentrate

Definition 62 (*Stochastic Dominance*)

Let Y, Z be RVs on $\mathbb{R}$. Z ***1st-order stochastically dominates*** Y, denoted by $Z \succeq_1 Y$, if

$$\mathbb{E}[f(Z)] \ge \mathbb{E}[f(Y)] \tag{161}$$

for all increasing f.

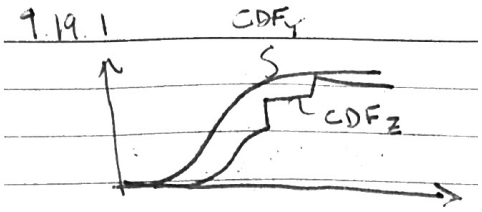


Figure 13: Intuition for $Z \succeq_1 Y$: going from Y to Z shifts cumulative distribution function (CDF) to the right

Lemma 63 (*Two-point first order stochastic dominance*)

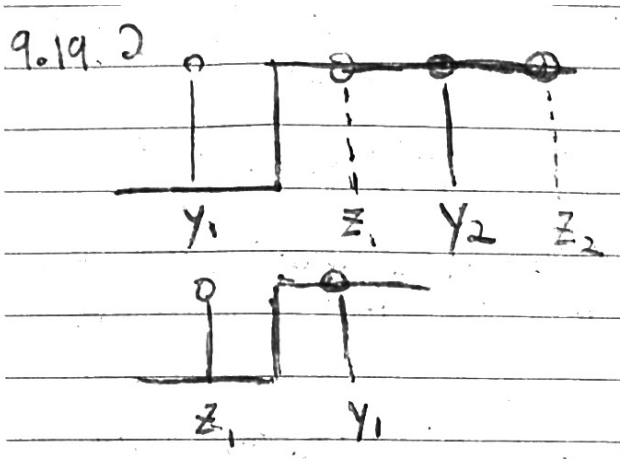
For $y_1 \le y_2$ and $z_1 \le z_2$, let

$$Y \sim \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2} \tag{162}$$

$$Z \sim \frac{1}{2}\delta_{z_1} + \frac{1}{2}\delta_{z_2} \tag{163}$$

Then $Z \succeq_1 Y$ iff $y_1 \le z_1$ and $y_2 \le z_2$.

Proof. For the necessity, consider $f_\tau(x) = \mathbb{1}\{x \ge \tau\}$ a step function.



A violation of $y_1 \leq z_1$ and $y_2 \leq z_2$ would imply for some τ both of the $y_i \leq \tau$ but only $z_1 \leq \tau$. The increasing function f_τ gives a contradiction to $Z \succeq_1 Y$, as

$$\mathbb{E}[f_\tau(Y)] = 1 \not\leq \frac{1}{2} = \mathbb{E}[f_\tau(Z)] \tag{164}$$

For the sufficiency,

$$\mathbb{E}[f(Z)] = \frac{f(z_1) + f(z_2)}{2} \geq \frac{f(y_1) + f(y_2)}{2} = \mathbb{E}[f(Y)] \tag{165}$$

□

Definition 64 (*Second order stochastic dominance*)

*Z **2nd-order stochasttically dominates** Y*, denoted $Z \succeq_2 Y$, if

$$\mathbb{E}[g(Y)] \leq \mathbb{E}[g(Z)] \tag{166}$$

for all convex, increasing g

Intuition: $Y \to Z$ by pushing CDF to right and spreading out

Lemma 65 (*Two-point second order stochastic dominance*)

{lem:two-point-2o-sd}

For $y_1 \leq y_2$ and $z_1 \leq z_2$, let

$$Y \sim \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2} \tag{167}$$

$$Z \sim \frac{1}{2}\delta_{z_1} + \frac{1}{2}\delta_{z_2} \tag{168}$$

If

$$\frac{1}{2}(y_1 + y_2) \leq \frac{1}{2}(z_1 + z_2) \tag{169}$$

$$z_2 \geq y_2 \tag{170}$$

then $Z \succeq_2 Y$.

Proof. Necessity follows from considering ramp functions such as:

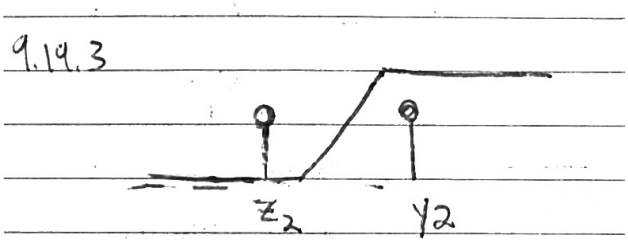


Figure 14: Example of a ramp function to show necessity that $y_2 \leq z_2$. The ramp function should be drawn to be convex (i.e. continue increasing)

g is convex and we see $\mathbb{E}[g(Z)] = 0 \not\geq \mathbb{E}[g(Y)]$ Sufficiency

□

6.3 Ledoux-Talagrand inequality

Theorem 66 (Ledoux-Talagrand) is a statement involving maxima of randomly signed sums of Lipschitz functions. We saw an incomplete presentation in Section 5.4.2 and today will give the complete proof.

Theorem 66 (*Ledoux-Talagrand*)

{thm:ledoux-talagrand}

Let $\phi : \mathbb{R} \to \mathbb{R}$ L-Lipschitz, $\phi(0) = 0$, $\{\varepsilon_i\}_{i=1}^n \overset{iid}{\sim}$ Rad, T = set of n-tuples $(t_1, \ldots, t_n)$ (think $t_i = \langle X_i - \mu, v\rangle$). Then

$$\mathbb{E}\left[g\left(\sup_{t\in T}\sum_{i=1}^n \varepsilon_i \phi(t_i)\right)\right] \leq \mathbb{E}\left[g\left(\sup_{t\in T} L\sum_{i=1}^n \varepsilon_i t_i\right)\right] \tag{171}$$

for all convex increasing g.

In terms of stochastic dominance, this is saying that the random variables

$$Y = \sup_{t\in T}\sum_{i=1}^n \varepsilon_i \phi(t_i) \tag{172}$$

$$Z = \sup_{t\in T} L\sum_{i=1}^n \varepsilon_i t_i \tag{173}$$

satisfy the second order stochastic dominance $Z \succeq Y$. This means that Z is more "spread out" than Y, which makes sense because $|\phi(s) - \phi(t)| \leq L|s - t|$.

Another way to get this intutiion is to notice that the term inside the supremum (for $L = 1$, if ε_i were Gaussian)

$$\mathrm{Var}\left(\sum_i \varepsilon_i \phi(t_i)\right) = \sum_i \phi(t_i)^2 \leq \sum_i |t_i|^2 = \mathrm{Var}\left(\sum_i \varepsilon_i t_i\right) \tag{174}$$

So we would expect Z to be more "spread out" because it has greater variance.

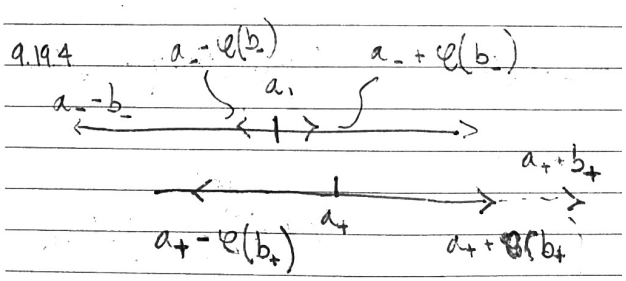
We will prove the theorem for $n = 2$ and then generalize by induction.

Proof for $n = 2$. Let T be the set of pairs (a, b), ϕ be 1-Lipschitz. Need to show

$$\mathbb{E}_\varepsilon\left[g\left(\underbrace{\sup_{(a,b)\in T} a + \varepsilon\phi(b)}_{=:Y}\right)\right] \leq \mathbb{E}_\varepsilon\left[g\left(\underbrace{\sup_{(a,b)\in T} a + \varepsilon b}_{=:Z}\right)\right] \tag{175}$$

Let (a_+, b_+) be the maximizer of $a + \phi(b)$, and (a_-, b_-) the maximizer of $a - \phi(b)$. Then

$$\Pr[Y = y_1 = a_+ + \phi(b_+)] = 1/2 \tag{176}$$
$$\Pr[Y = y_2 = a_- - \phi(b_-)] = 1/2 \tag{177}$$
$$\Pr[Z = z_1 = \max(a_+ + b_+, a_- + b_-)] = 1/2 \tag{178}$$
$$\Pr[Z = z_2 = \max(a_- - b_-, a_+ + b_-+)] = 1/2 \tag{179}$$



By Lipschitz condition and definition of y_i, z_i:

$$\max(y_1, y_2) \leq \max(z_1, z_2) \tag{180}$$
$$\max(a_+ + \phi(b_+), a_- - \phi(b_-)) \leq \max(a_+ + |b_+|, a_- + |b_-|) \tag{181}$$
$$y_1 + y_2 \leq z_1 + z_2 \tag{182}$$
$$a_+ + a_- + \phi(b_+) - \phi(b_-) \leq a_+ + a_- + |b_+ - b_-| \tag{183}$$

By Lemma 65, we are done. □

Extending to $n > 2$.

$$\mathbb{E}_{\varepsilon_{1:n}}\left[g\left(\sup_{t\in T}\sum_{i=1}^{n}\varepsilon_i\phi(t_i)\right)\right] = \mathbb{E}_{\varepsilon_{1:n-1}}\left[\mathbb{E}_{\varepsilon_n}\left[g\left(\sup_{t\in T}\underbrace{\sum_{i=1}^{n-1}\varepsilon_i\phi(t_i)}_{a} + \varepsilon_n\phi(\underbrace{t_n}_{b})\right)\middle|\varepsilon_1,\ldots,\varepsilon_{n-1}\right]\right] \tag{184}$$
$$\leq \mathbb{E}_{\varepsilon_{1:n-1}}\left[\mathbb{E}_{\varepsilon_n}\left[g\left(\sup_{t\in T}\sum_{i=1}^{n-1}\varepsilon_i\phi(t_i) + \varepsilon_n t_n\right)\middle|\varepsilon_1,\ldots,\varepsilon_{n-1}\right]\right] \tag{185}$$
$$= \mathbb{E}_{\varepsilon_{1:n}}\left[g\left(\sup_{t\in T}\sum_{i=1}^{n-1}\varepsilon_i\phi(t_i) + \varepsilon_n t_n\right)\right] \tag{186}$$
$$= \mathbb{E}_{\varepsilon_{[n]\setminus\{n-1\}}}\left[\mathbb{E}_{\varepsilon_{n-1}}\left[g\left(\sup_{t\in T}\sum_{i\in[n]\setminus\{n-1,n\}}\varepsilon_i\underbrace{\phi(t_i) + \varepsilon_n t_n}_{a} + \underbrace{\varepsilon_{n-1}\phi(t_{n-1})}_{b}\right)\middle|\varepsilon_{[n]\setminus\{n-1\}}\right]\right] \tag{187}$$

□

6.4 Bounding the empirical mean deviation

We now return to bounding $\|\mu(p_n^*) - \mu(p^*)\|_2$, which is the other step required before we can apply Proposition 55 for expanding the set.

The problem we encountered last time was the presence of a norm:

$$\mathbb{E}[\|\hat{\mu}_n - \mu\|_2^k] = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right\|_2^k\right] \tag{188}$$

One way to handle the norm is to take a supremum over inner products with $v \in \mathcal{S}^{d-1}$, since $\|w\|_2 = \sup_{v \in \mathcal{S}^{d-1}} \langle w, v \rangle$. Another is to use decoupling, which we will demonstrate today.

Decoupling technique: Use Khintchine's inequality to add in an $\mathbb{E}_\varepsilon$ with one ε_i per dimension d. Contrast this to symmetrization, which would have added random sign variables across n (one for each pair (X_i, X_i')).

Lemma 67 (*Khintchine's inequality*) {lem:khintchine}

Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \overset{iid}{\sim} Rad$.

$$A_k \|Z\|_2 \leq \mathbb{E}_\varepsilon[|\langle \varepsilon, z \rangle|^k]^{1/k} \leq B_k \|Z\|_2 \tag{189}$$

with $A_k = \Theta(1)$ *and* $B_k = \Theta(\sqrt{k})$ *if* $k \geq 1$.

Applying the lower $\Theta(1)$ bound from Khintchine's inequality:

$$\mathbb{E}_X[\|\hat{\mu}_n - \mu\|_2^k] = \mathbb{E}_X \left[\left\| \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \right\|_2^k \right] \tag{190}$$

$$\leq \mathcal{O}(1)^k \mathbb{E}_{X,\varepsilon} \left[\left| \langle \frac{1}{n} \sum_{i=1}^n (X_i - \mu), \varepsilon \rangle \right|^k \right] \tag{191}$$

$$= \mathcal{O}(1)^k \mathbb{E}_{X,\varepsilon} \left[\left| \frac{1}{n} \sum_{i=1}^n \langle X_i - \mu, \varepsilon \rangle \right|^k \right] \tag{192}$$

Now pulling out the n^{-k} and applying Lem:rosenthal-ineq 25 with $Z_i = \langle X_i - \mu, \varepsilon \rangle$

$$\mathbb{E}[|\sum_i z_i|^k] \leq \mathcal{O}(k)^k \sum_i \mathbb{E}[|Z_i|^k] + \mathcal{O}(\sqrt{k})^k \left(\sum_i \mathbb{E}[|Z_i|^2] \right)^{k/2} \tag{193}$$

Under bounded kth moments hypothesis, $\mathbb{E}[|\langle X - \mu, v \rangle|^k] \leq \sigma^k \|v\|_2^k$ so

$$\mathbb{E}[|Z_i|^k] = \mathbb{E}_{X,\varepsilon}[|\langle X_i - \mu, \varepsilon \rangle|^k] \leq \mathbb{E}_\varepsilon[\|\varepsilon\|_2^k \sigma^k] = d^{k/2} \sigma^k \tag{194}$$

To record a tighter bound (since typically $\sigma_k \approx \sqrt{k} \sigma_2$), let

$$\mathbb{E}[|\langle X - \mu, v \rangle|^k] \leq \sigma_k^k \|v\|_2^k \tag{195}$$

$$\mathbb{E}[|\langle X - \mu, v \rangle|^2] \leq \sigma_2^k \|v\|_2^2 \tag{196}$$

So Rosenthal's inequality (and adding back n^{-k}) becomes

$$\mathbb{E}[\|\hat{\mu}_n - \mu\|_2^k] \leq \mathcal{O}(1)^k \mathbb{E}_{X,\varepsilon}[|\frac{1}{n} \sum_{i=1}^n \langle X_i - \mu, \varepsilon \rangle|^k] \tag{197}$$

$$\leq \mathcal{O}(1/n)^k \left[\mathcal{O}(k)^k n d^{k/2} \sigma_k^k + \mathcal{O}(\sqrt{k})^k (n d^{1/2} \sigma_2^2)^{k/2} \right] \tag{198}$$

$$= \mathcal{O}\left(\left(\frac{k\sqrt{d}}{n} \sigma_k \right)^k n + \left(\sqrt{\frac{kd}{n}} \sigma_2 \right)^k \right) \tag{199}$$

In the case $\sigma_k = \sqrt{k} \sigma_2$, the second term dominates as long as $n \geq k^{2k/(k-2)}$.

Takeaway: The average deviation $\mathbb{E}[\|\hat{\mu}_n - \mu\|_2^k]^{1/k} \approx \mathcal{O}(\sqrt{kd/n} \sigma_2)$.

6.5 Zooming out

Goal: we want the following to concentrate

$$\mathbb{E}\left[\left|\sup_{\|v\|_2\le 1}\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\psi}_k\left(\left|\frac{\langle X_i-\mu,v\rangle}{\sigma}\right|\right)\right)-\mu_{\tilde{\psi}_k}(v)\right|^k\right] \tag{200}$$

By symmetrization

$$\mathbb{E}\left[\left|\sup_{\|v\|_2\le 1}\frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon_i\tilde{\psi}_k\left(\left|\frac{\langle X_i-\mu,v\rangle}{\sigma}\right|\right)\right)\right|^k\right] \tag{201}$$

By Ledoux with $g(x)=x^k$

$$\mathbb{E}\left[\left|\sup_{\|v\|_2\le 1}\frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon_i\frac{(X_i-\mu)}{\sigma}\right)\right|^k\right] \tag{202}$$

By a stronger version of the mean deviation inequality we just proved

7 9/24/2019

- PSet 2 posted by Friday, due Tuesday 10/8

7.1 Recap

Wanted to bound truncated moments $\tilde{\psi}_k$

- Ledoux-Talagrand for bounding deviation of $\tilde{\psi}_k$
- Khintchin and Rosenthal to bound $\mathbb{E}[\|\hat{\mu}_n-\mu\|_2^k]$

Today

- Finish up proof
- Efficient algorithms for $\mathcal{G}_{cov}(\sigma)$

Goal: Bound

$$\sup_{\|v\|\le 1}\frac{1}{n}\sum_{i=1}^{n}\tilde{\psi}_k\left(\left|\frac{\langle x_i-\mu,v\rangle}{\sigma}\right|\right),\quad \text{where } \tilde{\psi}_k(x)=\begin{cases}x^k, & \text{if } x\le x_0\\ \text{linear}, & \text{ow}\end{cases} \text{ is } L\text{-Lipschitz.} \tag{203}$$

For convenience, take $\mu=0$ and $\sigma=1$.
Consider symmetrizing:

$$\underbrace{\sup_{\|v\|_2\le 1}\frac{1}{n}\sum_{i=1}^{n}\tilde{\psi}_k(|\langle x_i,v\rangle|)-\mathbb{E}_{X'}[\tilde{\psi}_k(|\langle x_i',v\rangle|)]}_{Z(X)}+1 \tag{204}$$

We're going to bound $Z(X)$ using Chebyshev's inequality.

Let g be convex increasing.

$$\mathbb{E}_X[g(Z(X))] = \mathbb{E}_X g\left(\sup_{\|v\|_2\le 1}\frac{1}{n}\sum_{i=1}^n \tilde{\psi}_k(\langle x_i, v\rangle) - \underbrace{\mathbb{E}_{X'}\tilde{\psi}_k(\langle x_i', v\rangle)}_{\sup \mathbb{E}\le \mathbb{E}\sup}\right) \tag{205}$$

$$\le \mathbb{E}_X g\left(\mathbb{E}_{X'}\sup_{\|v\|_2\le 1}\frac{1}{n}\sum_{i=1}^n \tilde{\psi}_k(\langle x_i, v\rangle) - \tilde{\psi}_k(\langle x_i', v\rangle)\right) \tag{206}$$

$$\le \mathbb{E}_{X,X'} g\left(\sup_{\|v\|_2\le 1}\frac{1}{n}\sum_{i=1}^n \tilde{\psi}_k(\langle x_i, v\rangle) - \tilde{\psi}_k(\langle x_i', v\rangle)\right) \tag{207}$$

$$= \mathbb{E}_{X,X',\varepsilon} g\left(\sup_{\|v\|_2\le 1}\frac{1}{n}\sum_{i=1}^n \varepsilon(\tilde{\psi}_k(\langle x_i, v\rangle) - \tilde{\psi}_k(\langle x_i', v\rangle))\right) \tag{208}$$

$$\le \mathbb{E}_{X,X',\varepsilon} g\left(\underbrace{\sup_{\|v\|_2\le 1}\frac{1}{n}\sum_{i=1}^n \varepsilon\tilde{\psi}_k(\langle x_i, v\rangle)}_{A} + \underbrace{\sup_{\|v\|_2\le 1}\frac{1}{n}\sum_{i=1}^n \varepsilon\tilde{\psi}_k(\langle x_i', v'\rangle)}_{B}\right) \tag{209}$$

Applying Jensen's on g gigves $\mathbb{E}[\cdot] \le \mathbb{E}[g(2A)]$ so

$$\mathbb{E}_X[g(Z(X))] \le \mathbb{E}_{X,\varepsilon} g\left(\sup_{\|v\|_2\le 1}\frac{2}{n}\sum_{i=1}^n \varepsilon\tilde{\psi}_k(\langle x_i, v\rangle)\right) \tag{210}$$

Since $\tilde{\psi}_k$ is L-Lipschitz, applying Ledoux-Talagrand gives

$$\mathbb{E}_X[g(Z(X))] \le \mathbb{E}_{X,\varepsilon} g\left(\sup_{\|v\|_2\le 1}\frac{2L}{n}\sum_{i=1}^n \varepsilon_i \langle x_i, v\rangle\right) \tag{211}$$

$$= \mathbb{E}_{X,\varepsilon} g\left(\sup_{\|v\|_2\le 1}\langle x_i, \frac{2L}{n}\sum_{i=1}^n \varepsilon_i v\rangle\right) \tag{212}$$

$$= \mathbb{E}_{X,\varepsilon} g\left(\left\|\frac{2L}{n}\sum_{i=1}^n \varepsilon_i v\right\|_2\right) \tag{213}$$

So far this has been for generic convex increasing g. For kth moments, $g(x) = x^k$ and

$$\mathbb{E}_X[g(Z(X))] \le \left(\frac{2L}{n}\right)^k \mathbb{E}_{X,\varepsilon}\left[\left\|\sum_{i=1}^n \varepsilon_i X_i\|_2^k\right\|\right] \tag{214}$$

The remainder is handled using Khintchine and Rosenthal's inequality.

7.2 Efficient algorithms via eigenvector projection

Let the true distribution $p^* \in \mathcal{G}_{cov}(\sigma)$, so $\|\operatorname{Cov}_{p^*}[X]\| \le \sigma^2$. Let the corrupted distribution $\tilde{p}$ be such that $\operatorname{TV}(p^*, \tilde{p}) \le \varepsilon$.

Goal: Estimate $\mu = \mathbb{E}_{p^*}[X]$ with error $\mathcal{O}(\sigma\sqrt{\varepsilon})$ in ℓ_2-norm.

Will show: There exists an efficient algorithm that outputs q such that:

- $\operatorname{TV}(q, p^*) = \mathcal{O}(\varepsilon)$, which yields a modulus of continuity bound
- $\|\operatorname{Cov}_q(X)\| = \mathcal{O}(\sigma^2)$, which yields an error $\mathcal{O}(\mathcal{O}(\sigma)\sqrt{\mathcal{O}(\varepsilon)}) = \mathcal{O}(\sigma\sqrt{\varepsilon})$.

7.2.1 Representation

Let $\tilde{p}$ be the empirical distribution over n points $\{x_i\}_{i=1}^n$.
Let p^* be the empirical distribution over a subset $\{X_i\}_{i\in S}$ of points from $\tilde{p}$ with $|S| \geq (1-\varepsilon)n$.

- Empirical distribution, so we can store on computer
- Subset condition, which is equivalent to p^* being an ε-deletion of $\tilde{p}$, hence $\mathrm{TV}(p^*, \tilde{p}) \leq \varepsilon$
- Deleting points can't make $\mathrm{Cov}[X]$ large: for any q an ε-deletion of p

$$\mathrm{Cov}_q[X] \preceq \frac{1}{1-\varepsilon}\mathrm{Cov}_p[X] \tag{215}$$

Figure 9.24.1: We had bad examples previously by putting all the points in a single direction.
Algorithm:

- Initialize $c_i = 1$ for all i
- Let $q(c)$ be the weight $\frac{c_i}{\sum_j c_j}$ on point x_i
- Repeat:
  - Compute $\hat{\mu}_= \mathbb{E}_{q(c)}[X]$
  - Compute $\hat{\Sigma}_c = \mathrm{Cov}_{q(c)}[X]$
  - Let $\hat{\sigma}_c^2 = \sup_{\|v\|\leq 1} v^\top \hat{\Sigma}_c v = \sup_{\|v\|_2 \leq 1} \frac{\sum_{i=1}^n c_i \langle x_i - \hat{\mu}_c, v\rangle^2}{\sum_i c_i}$,
  - If $\hat{\sigma}_c^2 \leq 20\sigma^2$, output $q(c)$
  - Else, $c_i \leftarrow c_i\left(1 - \frac{\tau_i}{\tau_{\max}}\right)$, $\tau_i = \langle x_i - \hat{\mu}_c, v^*\rangle^2$, $\tau_{\max} = \max_i \tau_i$.

The algorithm is based on the intuition that the maximizing v^* for $v^\top \Sigma v$ is precisely the top eigenvector for Σ.
Intuition: If $\mu \approx \hat{\mu}$, then $\tau_i \approx \langle x_i - \mu, v^*\rangle^2$ is the projection onto v^*.
Figure 9.24.2
However this intuition is flawed in two ways:

- Assuming $\mu \approx \hat{\mu}_c$, which is the goal to begin with
- Only holds in expectation (c.f. downweighting)
- Not many bad points

Proposition 68

Suppose $\|\mathrm{Cov}_{p^*}[X]\| \leq \sigma^2$. *Then Algorithm outputs* q *such that*

- $\mathrm{TV}(p^*, q) \leq \frac{\varepsilon}{1-\varepsilon}$
- $\|\mathrm{Cov}_q[X]\| \leq 20\sigma^2$

Hence by modulus bound, $\|\mu(q) - \mu\|_2 = O(\sigma\sqrt{\varepsilon})$

Sketch of proof. Invariant 1: $\mathrm{TV}(p^*, q(c)) \leq \frac{\varepsilon}{1-\varepsilon}$ always.
Invariant 2: $\sum_{i\in S}(1-c_i) \leq \sum_{i\notin S}(1-c_i)$
The second implies the first.

Let $c_i' = c_i\left(1 - \frac{\tau_i}{\tau_{\max}}\right)$.

$$\sum_{i\in S}(1-c_i') = \sum_{i\in S}(1-c_i) + \underbrace{\sum_{i\in S}(c_i - c_i')}_{=\frac{1}{\tau_{\max}}\sum_i c_i\tau_i} \tag{216}$$

To show Invariant 2, want

$$\sum_{i\in S} c_i\tau_i \le \sum_{i\notin S} c_i\tau_i \tag{217}$$

Note

$$\sum_{i\in S} c_i\tau_i = \sum_{i\in S} c_i\left\langle x_i - \hat{\mu}_c, v^*\right\rangle^2 \le \underbrace{\left(\frac{1}{|S|}\sum_{i\in S}\left\langle x_i - \hat{\mu}_c, v^*\right\rangle^2\right)}_{\text{looks like covariance}}\cdot(1-\varepsilon)n \tag{218}$$

Expanding this term

$$\frac{1}{|S|}\sum_{i\in S}\left\langle x_i - \hat{\mu}_c, v^*\right\rangle^2 = (v^*)^\top \mathbb{E}_{p^*}[(X-\hat{\mu}_c)(X-\hat{\mu}_c)^\top]v^* \tag{219}$$

$$= (v^*)^\top\left(\mathrm{Cov}_{p^*}[X] + (\mu-\hat{\mu}_c)(\mu-\hat{\mu}_C)^\top\right)v^* \tag{220}$$

$$= \underbrace{(v^*)^\top \mathrm{Cov}_{p^*}[X]v^*}_{\le\sigma^2} + \underbrace{\left\langle v^*, \mu-\hat{\mu}_c\right\rangle^2}_{\le\|\mu-\hat{\mu}_c\|_2^2\le\mathcal{O}(\hat{\sigma}_c^2\underbrace{\mathrm{TV}(p^*,q(c))}_{\le\varepsilon/(1-\varepsilon)})} \tag{221}$$

Therefore

$$\sum_{i\in S} c_i\tau_i \le (1-\varepsilon)n(\sigma^2 + \underbrace{\mathcal{O}(\hat{\sigma}_c^2\varepsilon)}_{\text{small as }\varepsilon\to 0}) \tag{222}$$

$$\sum_{i=1}^n c_i\tau_i = \hat{\sigma}_c^2\underbrace{\left(\sum_{i=1}^n c_i\right)}_{\ge(1-2\varepsilon)n} \tag{223}$$

Want: $\sum_{i=1}^n c_i\tau_i \ge 2\sum_{i\in S}\underbrace{c_i\tau_i}_{\approx\sigma^2} \approx 20\sigma^2$ where 20 works if $\varepsilon \le 1/12$. □

Generic proof outline for efficient algorithms: τ_i measures how bad points are, downweight on the τ_i.

8 9/26/2019

8.1 Recap

- Efficient algorithms for $\mathcal{G}_{cov}(\sigma)$
- Project onto top eigenvectors
  - Revealed bad points
- Invariant 1: remove more bad than good points

- **TODO** Invariant 2: $\mathrm{TV}(q(c), p^*) \le \frac{\varepsilon}{1-\varepsilon}$
- Today:
  - Other norms
    - Dual norm
    - Approximate eigenvector via SDP
    - Grothendieck's inequality

Recall our algorithm, which $\tilde{p}$ is an empirical distribution over n points $\{x_1, \dots, x_n\}$ and p^* an empirical distribution over a subset $\{x_i\}_{i \in S}$ from $\tilde{p}$ with $|S| \ge (1-\varepsilon)n$. Our algorithm represented the distribution q as

$$q(c): \text{ place } \frac{c_i}{\sum_{i=1}^n c_i} \text{ on point } x_i \tag{224}$$

Proposition 69

If $\sum_{i \in S}(1-c_i) \le \sum_{i \notin S}(1-c_i)$, then $\mathrm{TV}(q(c), p^*) \le \frac{\varepsilon}{1-\varepsilon}$.

Proof. Define $\beta = \frac{1}{n}\sum_{i=1}^n (1-c_i)$, so $\sum_{i=1}^n c_i = (1-\beta)n$. Proceed by case analysis on $\beta \le \varepsilon$.

When $\beta \le \varepsilon$,

$$\mathrm{TV}(p, q) = \frac{1}{2}\int |p(x) - q(x)|dx = \int \max(p(x) - q(x), 0)dx \tag{225}$$

$$\mathrm{TV}(p^*, q(c)) = \sum_{i \in S} \max\left(\underbrace{\frac{c_i}{(1-\beta)n} - \frac{1}{(1-\varepsilon)n}}_{\frac{1}{1-\beta} \le \frac{1}{1-\varepsilon} \implies \cdot \le 0}, 0\right) + \sum_{i \notin S} \frac{\overbrace{c_i}^{\le 1}}{(1-\beta)n} \tag{226}$$

$$\le \frac{\varepsilon}{1-\varepsilon} \tag{227}$$

Now for $\beta \ge \varepsilon$

$$\mathrm{TV}(p^*, q(c)) = \sum_{i \in S} \max\left(\frac{1}{(1-\varepsilon)n} - \frac{c_i}{(1-\beta)n}, 0\right) + 0 \tag{228}$$

$$= \frac{1}{(1-\varepsilon)(1-\beta)n} \sum_{i \in S} \max\left((1-\beta)(1-c_i) + \underbrace{(\varepsilon - \beta)}_{\le 0} c_i, 0\right) \tag{229}$$

$$\le \frac{\cancel{(1-\beta)}}{(1-\varepsilon)\cancel{(1-\beta)}n} \sum_{i \in S}(1-c_i) \tag{230}$$

$$= \frac{\varepsilon}{1-\varepsilon} \tag{231}$$

□

8.2 Other norms

Definition 70

Given a norm $\|\cdot\|$, the ***dual norm*** is

$$\|u\|_* = \sup_{\|v\| \le 1} \langle u, v \rangle \tag{232}$$

Example 71

$\|\cdot\|_2$ is self-dual: $\|\cdot\|_* = \|\cdot\|_2$.
$\|\cdot\|_\infty$ has dual norm $\|\cdot\|_1$.
$\|\cdot\|_1$ has dual norm $\|\cdot\|_\infty$.

Theorem 72

$\|\cdot\|_{**} = \|\cdot\|$ *if finite dimensional.*
$\|v\|_{(k)}$ = *sum of* k *largest coordinates (in absolute value).*
$\|u\|_{(k)*} = \max(\|u\|_\infty, \|u\|1/k)$. *To explain this last one, notice that*

$$\|u\|_{(k)}^* \le 1 \iff \text{convex hull of } \{-1, 0, +1\} \text{ and } k \text{ non-zero} \tag{233}$$

$$\sup_{\|u\|_{(k)*}\le 1} \langle u, v\rangle = \|v\|_{(k)} \tag{234}$$

We can now generalize our definitions to other norms:

$$\mathcal{G}_{cov}(\sigma) = \{p : \sup_{\|v\|_2\le 1} v^\top \operatorname{Cov}_p[X] v \le \sigma^2\} \tag{235}$$

$$\mathcal{G}_{cov}(\sigma, \|\cdot\|) = \{p : \sup_{\|v\|_*\le 1} v^\top \operatorname{Cov}_p[X] v \le \sigma^2\} \tag{236}$$ {{eq:bdd-cov-other-norms}}

Since we were previously using resilience of $\mathcal{G}_{cov}(\sigma)$ to get modulus bounds, we would like to show resilience as well:

Proposition 73

If $p \in \mathcal{G}_{cov}(\sigma, \|\cdot\|)$, *then* p *is* $(\mathcal{O}(\sigma\sqrt{\varepsilon}), \varepsilon)$*-resilient in* $\|\cdot\|$.
If $r \le \frac{p}{1-\varepsilon}$, *then* $\|\mu(r) - \mu(p)\| \le \sigma\sqrt{\frac{2\varepsilon}{1-\varepsilon}}$

Proof.

$$\|\mu(r) - \mu(p)\| = \sup_{\|v\|_*\le 1} \langle \mu(r) - \mu(p), v\rangle \tag{237}$$

For any $\|v^*\|_* \le 1$, by $p \in \mathcal{G}_{cov}(\sigma, \|\cdot\|)$ we have the bound

$$\operatorname{Var}_p[\langle X, v^*\rangle] = (v^*)^\top \operatorname{Cov}_p[X](v^*) \le \sigma^2 \tag{238}$$

Applying the previous argument used to prove resilience of $\mathcal{G}_{cov}(\sigma)$ (Corollary 5) yields the reuslt. □

How can we generalize the algorithm?

- Use thes ame algo
- Replace max eigenvector with $\sup_{\|v\|_*\le 1} v^\top \Sigma v$. NP-hard generally, so we want to approximate.

Example 74 (*Distribution learning using the 1-norm*)

Let π be a distribution on $[m]$.
Goal: Recover $\hat\pi$ such that $\mathrm{TV}(\hat\pi, \pi) = \frac{1}{2}\|\hat\pi - \pi\|_1$.
Trusted batches: $p^* = \pi^k$ k-tuples of independent.
Goal: Given $\tilde p$ such that $\mathrm{TV}(\tilde p, p^*) \le \varepsilon$, recover π in TV

Proposition 75

Can recover $\hat\pi$ such that $\mathrm{TV}(\hat\pi, \pi) \le \sqrt{\frac{\varepsilon}{k}}$.

Remark 76. Compare to the trivial bound of $\le \varepsilon$, we see that this is better whenever $k \ge \frac{1}{\varepsilon}$.

Proof. Represent samples $X \sim p^*$ as normalized count vector/histograms Z where

$$Z_j = \frac{1}{k} \sum_{i=1}^{k} \delta_{X_i = j} \tag{239}$$

so in particular $\mathbb{E}_{p^*}[Z] = \pi$. From here forwards we will use X to denote the normalized histogram.

Lemma 77

$$\sup_{\|v\|_\infty \le 1} v^\top \operatorname{Cov}_{p^*}[X] v \le \frac{1}{k} \tag{240}$$

Proof.

$$v^\top \operatorname{Cov}_{p^*}[X] v = \frac{1}{k} v^\top \operatorname{Cov}_{\pi}[X] v = \frac{1}{k} \operatorname{Var}_{\pi}[\langle X, v \rangle] \tag{241}$$

$$\le \frac{1}{k} \mathbb{E}_{\pi}[\langle X, v \rangle^2] = \frac{1}{k} \sum_{j=1}^{m} \pi_j \underbrace{v_j^2}_{\le 1} \tag{242}$$

$$\le \frac{1}{k} \sum_j \pi_j \tag{243}$$

$$\le \frac{1}{k} \tag{244}$$

□

□

Definition 78 (*κ-approximate oracle*)

A κ-approximate oracle A is a matrix-valued function $M = A(\Sigma)$ such that

1. $\langle M, \Sigma \rangle \ge \sup_{\|v\|_* \le 1} v^\top \Sigma v$: it's correctly large on bad points (overall)
2. For any Σ', $\langle M, \Sigma' \rangle \le \kappa \sup_{\|v\|_* \le 1} v^\top \Sigma' v$: it doesn't accidentally think the good points are bad
3. $M \succeq 0$

Modify the filtering step of our algorithm:

$$q(c) = \frac{c_i}{\sum_i c_i} \tag{245}$$

a distribution on x_i.

- Initialize $c_i = 1$ for all i
- Compute

$$\hat{\mu}_c = \mathbb{E}_{q(c)} X \tag{246}$$

$$\hat{\Sigma}_c = \operatorname{Cov}_{q(c)} X \tag{247}$$

$$M = A(\Sigma) \tag{248}$$

- If $\langle M, \hat{\Sigma}_c \rangle \le 20 \kappa \sigma^2$, output $q(c)$

- Else

$$\tau_i = (x_i - \hat{\mu}_c)^\top M (x_i - \hat{\mu}_c) \tag{249}$$
$$c_i \leftarrow c_i(1 - \tau_i/\tau_{\max}) \tag{250}$$

Proposition 79

If $p^* \in \mathcal{G}_{cov}(\sigma, \|\cdot\|)$, then

$$\|\mu(p^*) - \mu(q(c))\| \le \mathcal{O}(\sigma\sqrt{\kappa\varepsilon}) \tag{251}$$

How do we get to a κ-approximate oracle? One way is to consider relaxation of eigenvalue problem:

$$\max v^\top \Sigma v \qquad \text{st } \|v\|_\infty \le 1 \tag{252}$$
$$\max \langle vv^\top, \Sigma \rangle \qquad \text{st } \|v\|_\infty \le 1 \tag{253}$$
$$\max \langle M, \Sigma \rangle \qquad \text{st } M_{jj} = 1 \ \forall j, M \succeq 0, \operatorname{rank}(M) = 1 \tag{254}$$

The rank constraint is the only problem, so we will just relax it to get the SDP (which is solvable in polynomial time)

$$\begin{aligned} &\max \langle M, \Sigma \rangle \\ &\text{st } M \succeq 0 \\ &\quad \operatorname{diag}(M) = 1 \end{aligned} \tag{255}$$

{{eq:kappa-approx-oracle-sdp}}

Theorem 80 (*Grothendieck*)

$$\textit{Optimal value of Eq. (255)} \le \frac{\pi}{2} \max_{\|v\|_\infty \le 1} v^\top \Sigma v \tag{256}$$

Hence, the SDP Eq. (255) is a $\frac{\pi}{2}$-approximate oracle (assuming $\Sigma \preceq 0$).

Proof. Define

$$\arcsin[X]_{ij} = \arcsin[X_{ij}] \tag{257}$$

We will show two identities:

1. $\sup_{\|v\|_\infty \le 1} v^\top \Sigma v = \frac{2}{\pi} \sup_{\substack{M \succeq 0 \\ \operatorname{diag}(M)=1}} \langle \arcsin[M], \Sigma \rangle$
2. $\arcsin[X] \succeq X$

For the first, $M \succeq 0$ means we can write $M = UU^\top$ where $M_{ij} = \langle u_i, u_j \rangle$. Since $1 = M_{ii} = \|u_i\|^2$, we have that u_i are unit vectors. We will do two things:

1. $M \implies$ distribution over $v \in \{\pm 1\}^d$ such that $\mathbb{E} vv^\top = \frac{2}{\pi} \arcsin[M]$ (think randomized rounding)
2. $\frac{2}{\pi} \arcsin(vv^\top) = vv^\top$ (just a calculation)

For the first, let $g \sim N(0, I)$ and consider

$$v_i = \operatorname{sgn}(\langle u_i, g \rangle) \tag{258}$$

Notice

$$\mathbb{E}_g[v_i v_j] = \frac{2}{\pi} \arcsin \langle u_i, u_j \rangle \tag{259}$$

Figure 9.26.1 □

9 10/1/2019

9.1 Semidefinite Programing and Sum of Squares

Theorem 81 (*Grothendieck's Inequality*)

$$\frac{\pi}{2} \max_{\|v\|_\infty \le 1} v^\top \Sigma v \ge \max_{\substack{M \succeq 0 \\ \operatorname{diag} M = 1}} \langle M, \Sigma \rangle \tag{260}$$

Proof. We first consider (1) and Will show:

1. $\max_{\|v\|_\infty \le 1} v^\top \Sigma v = \frac{2}{\pi} \max_{\substack{M \succeq 0 \\ \operatorname{diag} M = 1}} \langle \arcsin M, \Sigma \rangle$
2. $\arcsin X \succeq X$

after which composing the two gives our desired result.

Easy: LHS $\le$ RHS. LHS max attained for $v \in \{\pm 1\}^d$, set $M = vv^\top$ in RHS. Since $\arcsin(1) = \frac{\pi}{2} \cdot 1$ and $\arcsin(-1) = \frac{\pi}{2} \cdot (-1)$,

$$\frac{2}{\pi} \langle \underbrace{vv^\top}_{= \frac{\pi}{2} vv^\top}, \Sigma \rangle = \langle vv^\top, \Sigma \rangle = v^\top \Sigma v \tag{261}$$

Harder: RHS $\le$ LHS. Will do randomized rounding. Given M, construct $\rho(v)$ such that

$$\mathbb{E}_{v \sim \rho}[v^\top \Sigma v] = \frac{2}{\pi} \langle \arcsin(M), \Sigma \rangle \iff \mathbb{E}_{v \sim \rho}[vv^\top] = \frac{2}{\pi} \langle \arcsin(M), \Sigma \rangle \tag{262}$$

$M \succeq 0 \implies M = UU^\top \implies M_{ij} = \langle u_i, u_j \rangle$. $\operatorname{diag} M = 1 \implies u_i$ are unit vectors.

The distribution will be $g \sim N(0, I)$ and $v_i = \operatorname{sgn}(g \cdot u_i)$.

Figure 10.1.1:

$$\begin{aligned} \mathbb{E}[v_i v_j] &= \mathbb{E}[\operatorname{sgn}[\langle g, u_i \rangle] \operatorname{sgn}[\langle g, u_j \rangle]] && (263) \\ &= 2 \Pr[\operatorname{sgn}[\langle g, u_i \rangle]] \operatorname{sgn}[\langle g, u_j \rangle] - 1 && (264) \end{aligned}$$

This is really how likely for both to be on the same side of a hyperplane.

Figure 10.1.2: the projection of g onto u_i and u_j have opposite sign only when u_i and u_j are split by the hyperplane orthogonal to g.

$$\begin{aligned} \theta &= \arccos \langle u_i, u_j \rangle && (265) \\ 1 - \Pr[\text{opposite signs}] &= 1 - \frac{\theta}{\pi} && (266) \\ &= \frac{\pi - \arccos \langle u_i, u_j \rangle}{\pi} && (267) \\ &\ge \frac{\pi - 2 \arccos \langle u_i, u_j \rangle}{\pi} && (268) \\ &\vdots \text{ algebra} && (269) \\ &= \frac{2}{\pi} \arcsin \langle u_i, u_j \rangle && (270) \\ &= \frac{2}{\pi} \arcsin(M_{ij}) && (271) \end{aligned}$$

Now we consider (2).

$$\arcsin X \succeq X \tag{272}$$

$$\arcsin(Z) = \underbrace{Z + \frac{Z^3}{6} + \cdots}_{\text{positive coeffs}} \tag{273}$$

$$\arcsin(X) = X + \frac{X^{\odot 3}}{6} + \cdots \succeq X \tag{274}$$

where $X_i^{\odot k} j = (X_{ij})^k$ is elementwise power.

Lemma 82

If $X \succeq 0$, then $X^{\odot k} \succeq 0$ for $k \in \mathbb{N}$.

Proof. Recall the Hadamard (i.e. tensor) product:

$$A, B \succeq 0, \qquad A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}n' \times n' \tag{275}$$

$$A \otimes B \in \mathbb{R}^{(n \cdot n') \times (n \cdot n')} \tag{276}$$

$$(A \otimes B)_{ii',jj'} = A_{ij} B_{i'j'} \tag{277}$$

$$u \in \mathbb{R}^n \qquad v \in \mathbb{R}^{n'} \tag{278}$$

$$u \otimes v \in \mathbb{R}^{n \cdot n'} \tag{279}$$

$$(u \otimes v)_{ii'} = u_i v_{i'} \tag{280}$$

Note in particular $(A \otimes B)(u \otimes v) = (Au) \otimes (Bv)$ so the eigenvalues of $A \otimes = \lambda_i \lambda'_j$ for $\text{eig}(A) = \{\lambda_i\}_i$ and $\text{eig}(B) = \{\lambda'_j\}_j$. Hence, if $A, B \succeq 0$ then $A \otimes B \succeq 0$. But since $A_{ij}^{\odot k} = (A^{\otimes k})_{ij,ij}$ is a principal submatrix of a PSD matrix, $A^{\odot k}$ is PSD. □

□

9.2 Semidefinite programing

$$\max \langle A, X \rangle \qquad \text{objective} \tag{281}$$

$$\text{s.t.`} X \succeq 0 \qquad \text{PSD constraint} \tag{282}$$

$$\tag{283}$$

where $\langle X, Y \rangle = \sum_{i,j} X_{ij} Y_{ij}$, $X, A, B \in \mathbb{R}^{n \times n}$, and $c_j \in \mathbb{R}$.

SDP preserving operations:

- min instead of max (i.e. $A \mapsto -A$)
- equality constraints
- $X \succeq 0 \implies \mathcal{L}(X) \succeq 0$ (e.g. X_1, $X_2 \succeq 0$, $X_1 + 2X_2 \succeq 0$)
- $\mathcal{L}_1(X) \succeq 0$, …, $\mathcal{L}_k(X) \succeq 0$, then $\text{diag}(L_i(X)) \succeq 0$.

10 10/3/2019

Goal: Bound $2k$th moments $\sup_{\|v\|_2 \leq 1} \mathbb{E}[\langle X - \mu, v \rangle^2 k] = \sup_{\|v\| \leq 1} \langle M_{2k}, v^{\otimes 2k} \rangle$ where $M_{2k} \in \mathbb{R}^{d^{2k}}$ is the $2k$th moment tensor.

Idea: Polynomial program is NP hard. Approximate via SoS program

$$\min \ \lambda \tag{284}$$

$$\text{st } \lambda\|v\|_2^{2k} - \langle M_{2k}, v^{\otimes 2k}\rangle \geq_{SoS} 0 \tag{285}$$

Today: Analyze the SoS program. Show that λ is small.

- Poincaré inequality
- SoS proofs

10.1 Sum-of-squares proofs

Definition 83

The inequality $p(v) \leq q(v)$ has a ***sum-of-squares proof*** (i.e. is ***sum-of-squares certifiable***) if $q(v) - p(v) \geq_{SoS} 0$. In this case, we write $p \leq_{SoS} q$.

Switching from SDPs to SoS is motivated by the following nice composition properties for sum-of-squares proofs.

Proposition 84

$\leq_{SoS}$ is similar to $\leq$:

1. $p_1 \leq_{SoS} p_2, p_2 \leq_{SoS} p_3 \implies p_1 \leq_{SoS} p_3$
2. $p_1 \leq_{SoS} q_1, p_2 \leq_{SoS} q_2 \implies p_1 + p_2 \leq_{SoS} q_1 + q_2$
3. $p_1, p_2 \geq_{SoS} 0 \implies p_1, p_2 \geq_{SoS} 0$
4. $p_1 \leq_{SoS} p_2, q_1 \leq_{SoS} q_2, p_1 \geq_{SoS} 0, q_2 \geq_{SoS} 0 \implies p_1 q_1 \leq_{SoS} p_2 q_2$

Proof of (3).

$$p_1(v)p_2(v) = \left(\sum_i p_{1i}(v)^2\right)\left(\sum_j p_{2j}(v)^2\right) = \sum_{ij}(p_{1i}(v)p_{2j}(v))^2 \tag{286}$$

□

Proof of (4).

$$p_2q_2 - p_1q_1 = p_2 \underbrace{(q_2 - q_1)}_{\geq_{SoS} 0} + q_1 \underbrace{(p_2 - p_1)}_{\geq_{SoS} 0} \geq_{SoS} 0 \tag{287}$$

□

Remark 85. Most "standard" inequalities have SoS proofs:

- Arithmetic mean geometric mean
- Cauchy-Schwarz
- Hölder's inequality

Our general strategy: construct SoS proof that $\langle M_{2k}, v^{\otimes 2k}\rangle \leq_{SoS} \lambda\|v\|_2^{2k}$

Lemma 86 (*PSD implies SoS*)

{lem:psd-implies-sos}

If $P \succeq 0$, then $v^\top P v \geq_{SoS} 0$.

Proof. $P = \sum_i \lambda_i u_i u_i^\top$ so $v^\top P v = \sum_i \lambda_i \langle u_i, v\rangle^2 \geq_{SoS} 0$. □

Proof for Gaussians.

$$\mathbb{E}_{X\sim N(\mu,\Sigma)}\left[\langle X-\mu, v\rangle^{2k}\right] = \underbrace{(2k-1)(2k-3)\cdots 3\cdot 1}_{\text{product of odd numbers}} \cdot (v^\top \Sigma v)^k \tag{288}$$

(see Issirlis's Theorem).

Applying Lemma 86 shows $v^\top \Sigma v \leq_{SoS} \|\Sigma\|_{op}\|v\|_2^2$, and

$$(v^\top \Sigma v)^k \leq_{SoS} (\|\Sigma\|_{op}\|v\|_2^2)^k = \|\Sigma\|_{op}^k \|v\|_2^{2k} \tag{289}$$

So $\lambda = \|\Sigma\|_{op}^2((2k-1)\cdots 3\cdot 1)$ provides a $2k$th moment bound for Gaussians. □

10.2 Poincaré inequality

Definition 87

A distribution p on $\mathbb{R}^d$ satisfies ***Poincaré inequality with parameter*** σ if

$$\mathrm{Var}_p[f(x)] \leq \sigma^2 \mathbb{E}_p[\|\nabla f(x)\|_2^2] \tag{290}$$

for all differentiable $f : \mathbb{R}^d \to \mathbb{R}$.

Remark 88.

- Interpret as global $\leftrightarrow$ local property. If f has a lot of variation under p, then for a typical $x \sim p$ it is also changing a lot (i.e. large derivative)
- No "holes": the support of p must be connected and have trivial fundamental group. This excludes discrete distributions. Also see Figure 10.3.1. (Fig 10.3.1: Non-example. Take two disjoint sets both of probability 1/2, f the Urysohn's Lemma separating function, $\mathrm{Var}[f] = 1/4$ and $\mathbb{E}[\|\nabla f\|_2^2] = 0$ so this p with a "hole" between A and B fails to satisfy any Poincaré inequality.)

Example 89

- $\mathcal{N}(\mu, \sigma^2)$ is σ-Poincaré
- p, p' σ-Poincaré, then $p \times p'$ is σ-Poincaré. In particular, $\mathcal{N}(\mu, \sigma^2 I)$ is σ-Poincaré.
- $X \sim p$ σ-Poincaré, A a linear mp, then AX is $(\|A\|_{op}\sigma)$-Poincaré. In particular, $\mathcal{N}(\mu, \Sigma)$ is $\|\Sigma\|_{op}^{1/2}$-Poincaré.
- $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ is L-Lipschitz, then $f(X)$ is $(L\sigma)$-Poincaré.

Theorem 90 (*Bakry + Émery, 1985*)

If p is log-*concave, $\nabla^2 \log p(x) \leq -\frac{1}{\sigma^2} I$, then p is σ-Poincaré.*

Remark 91. This theorem covers Gaussians, where $p(x) = \exp(-\psi(x))$ with $\psi(x) = (X-\mu)^\top \Sigma^{-1}(X-\mu)$.

Theorem 92 (*Bounded distributions are Poincaré after convolving with Gaussians*)

$X \sim p$, $\mathrm{supp}(p)$ has radius $\leq R$, $Z \sim N(0, \tau^2 I)$ with $\tau \geq 2R$, then $X + Z$ is $(\tau\sqrt{e})$-Poincaré

Theorem 93 (*Lipschitz functions of Poincaré RVs are SE*)

If p is σ-Poincaré, f is L-Lipschitz, then

$$\Pr[|f(x) - \mathbb{E}f(x)| \geq t] \leq 6\exp\left(-\frac{t}{\sigma L}\right) \tag{291}$$

Example 94

Let $(X, Y) \in \mathbb{R}^{2d}$, $X \sim N(0, I)$, $Y = \varepsilon \cdot X$, $\varepsilon \sim \text{Rad}$.
Consider $f(X, Y) = \sum_i X_i Y_i$. Then

$$\nabla f(X, Y) = (Y_1, \ldots, Y_d, X_1, \ldots, X_d) \tag{292}$$
$$\|\nabla f(X, Y)\|_2^2 \approx 2d \tag{293}$$
$$\text{Var}[f] \approx d^2 \tag{294}$$

where the first $\approx$ is because each of the $2d$ coordinates of ∇f is Gaussian and $\|X\|_2^2 \approx 1$, and the second $\approx$ is because

$$Y_i = \varepsilon X_i \tag{295}$$
$$f(X, Y) = \varepsilon \cdot (X_1^2 + \cdot + X_d^2) \tag{296}$$

So f is approximately $+d$ or $-d$ with equal probability.

Theorem 95 (*Adamczak and Wolff, 2015*)

If $f : \mathbb{R}^d \to \mathbb{R}$ such that

$$\mathbb{E}[\nabla f(x)] = 0 \tag{297}$$
$$\mathbb{E}[\nabla^2 f(x)] = 0 \tag{298}$$
$$\vdots \tag{299}$$
$$\mathbb{E}[\nabla^{k-1} f(x)] = 0 \tag{300}$$

Then $\text{Var}_p[f] \leq c_k \sigma^{2k} \mathbb{E}[\| \underbrace{\nabla^k f(x)}_{\mathbb{R}^{d^k}} \|_F^2]$

10.3 SoS proofs for 2k moments

Recall our goal of showing

$$\mathbb{E}_p[\langle X - \mu, v\rangle^{2k}] \leq_{SoS} \lambda \|v\|_2^2 \tag{301}$$

p is σ-Poincaré $\implies$ $\lambda = c_k \sigma^{2k}$. We will consider $k = 1, 2, 3$ here.

Define the moment tensor

$$M_k = \mathbb{E}_p[(X - \mu)^{\otimes k}] \tag{302}$$
$$M_k(v) = \langle M_k, v^{\otimes k}\rangle = \mathbb{E}_p[\langle X - \mu, v\rangle^k] \tag{303}$$

Proof for k=1. Poincaré

$$f_v(X) = \langle X - \mu, v\rangle \tag{304}$$
$$\nabla f_v(X) = v \tag{305}$$
$$M_2(v) = \text{Var}[f_v] \leq \sigma^2 \mathbb{E}[\|v\|_2^2] = \sigma^2 \|v\|^2 \tag{306}$$

But this means $\sigma^2 I - M_2 \succeq 0$ so by Lemma 86 (PSD implies SoS) $M_2(v) \leq_{SoS} \sigma^2 \|v\_2^2$. □

Proof for k=2.

$$f_v(X) = \langle X - \mu, v\rangle^2 \tag{307}$$
$$\nabla f_v(X) = 2\langle X - \mu, v\rangle \cdot v \tag{308}$$
$$\implies \mathbb{E}[\nabla f_v(x)] = 0 \tag{309}$$

$$\mathrm{Var}[f_v] \le c_2\sigma^2\mathbb{E}[\|\nabla^2 f_v\|_F^2] \tag{310}$$
$$= c_2\sigma^2\mathbb{E}[\|4vv^\top\|_F^2] \tag{311}$$
$$= 4c_2\sigma^2\|v\|_2^4 \tag{312}$$

But we know

$$\mathrm{Var}[f_v] = M_4(v) - M_2(v)^2 \tag{313}$$
$$M_4(V) = \mathrm{Var}[f_v] + M_2(v)^2 \tag{314}$$
$$\le 4c_2\sigma^4\|v\|_2^4 + \sigma^4\|v\|_2^4 \tag{315}$$
$$\le (4c_2+1)\sigma^4\|v\|_2^4 \tag{316}$$

{{eq:var-upper-bd-frob}}

How do we turn this to a SoS proof? One part is easy: by the multiplicative composition property for SoS ordering

$$M_2(v) \le_{SoS} \sigma^2\|v\|_2^2 \tag{317}$$
$$M_2(v)^2 \le_{SoS} \sigma^4\|v\|_2^4 \tag{318}$$

The Poincaré term is harder. Passing from vector v to matrix A

$$f_v(X) = \langle X - \mu, v\rangle^2 \tag{319}$$
$$f_A(X) = (X-\mu)^\top A(X-\mu) \tag{320}$$
$$\mathrm{Var}[f_A] \le c_2\sigma^4 4\|A\|_F^2 \tag{321}$$

where we used Eq. (315). But also

$$\mathrm{Var}[f_A] = \mathbb{E}[((X-\mu)^\top A(X-\mu))^2] - \mathbb{E}[(X-\mu)^\top A(X-\mu)]^2 \tag{322}$$
$$= \langle M_4, A\otimes A\rangle - \langle M_2\otimes M_2, A\otimes A\rangle \tag{323}$$

Parsing the tensor notation, this again says that M_4 satisfies a PSD constraint where we can apply Lemma 86 (PSD implies SoS). □

k=3. Consider $f_v(X) = \langle X-\mu, v\rangle^3$. This doesn't work because

$$\nabla f_v(x) = 3\langle X-\mu, jv\rangle^2 v \tag{324}$$
$$\mathbb{E}\nabla f_v(x) = 3\Sigma(v)\cdot v \neq 0 \tag{325}$$

So instead we pick

$$f_v(x) = \langle X-\mu, v\rangle^3 - 3(v^\top\Sigma v)\langle X-\mu, v\rangle \tag{326}$$

and verify

$$\mathbb{E}[\nabla f_v] = 0 \tag{327}$$
$$\mathbb{E}[\nabla^2 f_v] = 0 \tag{328}$$
$$\nabla^3 f_v = 6(v\otimes v\otimes v) \tag{329}$$

Apply the theorem to the third derivative, use our previous bounds to handle the even moments. There is an additional $M_3(v)^2$ term, which we can handle with Hölder's inequality to show

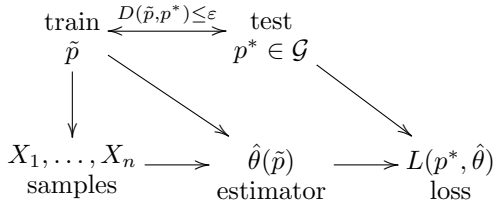
$$M_3(v)^2 \le_{SoS} M_2(v)M_4(v) \tag{330}$$

□

11 10/8/2019

11.1 Resilience Beyond Mean Estimation

Recall our general framework for robust statistics (Fig. 1, reproduced below):



So far we have only considered mean estimation where the discrepancy $D = \mathrm{TV}$ and the loss $L(p, \theta) = \|\theta - \mu(p)\|_2$. In this section, we will continue to take $D = \mathrm{TV}$ but will now consider more general losses. Developing this theory will require suitable generalizations of:

- Modulus of continuity
- Resilience
- Analogue of $\mathcal{G}_{\mathrm{Cov}}(\sigma)$
  - Moment estimation
  - Linear regression

For now, assume $n = \infty$ so we neglect finite sample issues and our estimator $\hat{\theta}(\tilde{p})$ directly uses the corrupted population distribution $\tilde{p}$.

Example 96

$D = \mathrm{TV}$ and $L(p, \theta) = \|\theta - \mu(p)\|_2$ is the previous mean-estimation framework considered in previous sections.

For second-moment estimation, we can consider the loss function

$$L(p, \underbrace{S}_{\in \mathbb{R}^{d \times d}}) = \|S - \mathbb{E}_p[XX^\top]\| \tag{331}$$

However, the operator norm is only sensitive to the top eigenspace so other times a more natural loss is

$$\|I - \Sigma^{-1} \operatorname{Cov}_p[X]\|_F \tag{332}$$

Here, the Frobenius norm now weights all eigenvalues equally.

In linear regression, we will consider the ***excess squared loss***

$$L(p, \theta) = \mathbb{E}_{(x,y)\sim p}[(y - \langle \theta, x \rangle)^2 - (y - \langle \theta^*(p), x \rangle)^2] \tag{333}$$

11.1.1 Generalizing the modulus of continuity bound

Proposition 2 (Modulus of continuity bound) generalizes naturally: define the modulus of continuity

$$\mathfrak{m}(\mathcal{G}, 2\varepsilon, L) = \sup_{\substack{p,q \in \mathcal{G} \\ \mathrm{TV}(p,q) \leq 2\varepsilon}} L(p, \theta^*(q)) \tag{334}$$

This modulus bounds the minimax loss (i.e. worst loss for minimum distance functional), or more precisely:

Proposition 97

Let the minimum distance functional (MDF) be

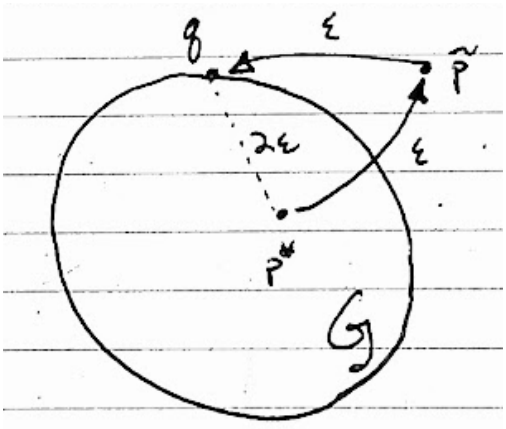
$$\hat{\theta}(\tilde{p}) = \theta^*(q) \text{ where } q = \operatorname{argmin}_{q \in \mathcal{G}} \mathrm{TV}(\tilde{p}, q) \tag{335}$$

Then

$$\mathrm{TV}(p^*, q) \leq 2\varepsilon \tag{336}$$

$$L(p^*, \theta^*(q)) \leq \mathfrak{m}(\mathcal{G}, 2\varepsilon) \tag{337}$$

Proof. By assumption, $p^* \in \mathcal{G}$ and $\mathrm{TV}(\tilde{p}, p^*) \leq \varepsilon$. Since q is the minimum-TV-distance projection of $\tilde{p}$ onto $\mathcal{G}$, $\mathrm{TV}(q, p^*) \leq \varepsilon$ and by the triangle inequality



We have that $\mathrm{TV}(p^*, q) \leq 2\varepsilon$). □

11.1.2 Resilience

Resilience is less trivial. Recall from Definition 7 (Resilient distribution) that p is (ρ, ε)-resilient if

$$\|\mu(p) - \mu(r)\|_2 \leq \rho \text{ whenever } r \leq \frac{p}{1-\varepsilon} \tag{338}$$

Our argument for robust mean estimation for $p^* \in \mathcal{G}_{\mathrm{TV}}$ relied on (1) the existance of a midpoint distribution, and (2) the triangle inequality. A sketch of the argument is below:

Lemma 98

If $\mathrm{TV}(p, q) \leq \varepsilon$*, there is a* **midpoint** *r such that* $r \leq \frac{p}{1-\varepsilon}$, $r \leq \frac{q}{1-\varepsilon}$.

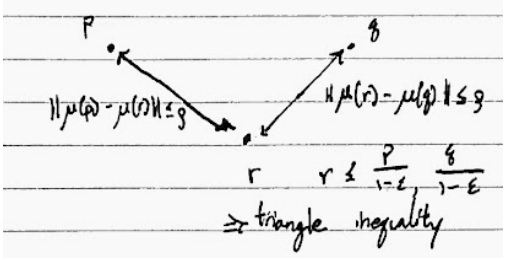


Figure 15: For resilient distributions, $\mathfrak{m}(\mathcal{G}_{\mathrm{TV}}, \varepsilon) \leq 2\rho$.

{fig:res-dist
-mdf-triangle
}

While so long as $D = \mathrm{TV}$ we still have Lemma 4, more general losses $L(p, \theta)$ may not satisfy the triangle inequality. Handling this requires the following generalized definition for resilience:

{def:resilience-general}

Definition 99 (*Resilience for general losses*)

p is $(\rho_1, \rho_2, \varepsilon)$-***resilient (for general L)*** if all of the following hold:

Downwards condition $\mathcal{G}_\downarrow$ $L(r, \theta^*(p)) \leq \rho_1$ for all $r \leq \frac{p}{1-\varepsilon}$. So the parameter $\theta^*(p)$ should do well for all ε-deletions r.

Upwards condition $\mathcal{G}_\uparrow$ If $L(r, \theta) \leq \rho_1$ for any $r \leq \frac{p}{1-\varepsilon}$ and θ, then $L(p, \theta) \leq \rho_2$. So any parameter θ which does well on some ε-deletion also does well on p.

Example 100 (*Compatibility of generalized resilience for mean estimation*)

To see what these conditions imply for the familiar setting of mean estimation, take $L(p, \theta) = \|\mu(p) - \theta\|_2$ and $\theta^*(p) = \mu(p)$.

The downward condition requires

$$L(r, \theta^*(p)) = \|\mu(r) - \mu(p)\|_2 \tag{339}$$

$$\|\mu(r) - \mu(p)\|_2 \leq \rho_1 \qquad \forall r \leq \frac{p}{1-\varepsilon} \tag{340}$$

In other words, the mean of any ε-deletion r is within ρ_1 of the original mean. This is precisely Definition 7 (Resilient distribution) for $\mathcal{G}_{\mathrm{TV}}$.

The upward condition says

$$\|\mu(r) - \theta\|_2 \leq \rho_1 \implies \|\mu(p) - \theta\|_2 \leq \rho_2 \tag{341}$$

But notice from the downward condition

$$\|\mu(p) - \theta\|_2 \leq \|\mu(p) - \mu(r)\| + \|\mu(r) - \theta\| \leq \rho_1 + \rho_1 \tag{342}$$

So choosing $\rho_2 = 2\rho_1$, we see that the downwards condition includes the upwards condition. Together, we see that generalized resilience is compatible with our previous definition for $\mathcal{G}_{\mathrm{TV}}$. In other words

$$(\rho, \varepsilon)\text{-resilience} \iff (\rho, 2\rho, \varepsilon)\text{-resilience}$$

Definition 101

Let $\mathcal{G}_\downarrow(\rho_1, \varepsilon) = \{$ all p satisfying $\downarrow\}$, $\mathcal{G}_\uparrow(\rho_1, \rho_2, \varepsilon) = \{$ all p satisfying $\uparrow\}$, and $\mathcal{G}_{\mathrm{TV}}(\rho_1, \rho_2, \varepsilon) = \mathcal{G}_\downarrow(\rho_1, \varepsilon) \cap \mathcal{G}_\uparrow(\rho_1, \rho_2, \varepsilon)$.

The key property of generalized resilience is that this family has a nice modulus bound:

Proposition 102

$\mathfrak{m}(\mathcal{G}_{\mathrm{TV}}(\rho_1, \rho_2, \varepsilon), \varepsilon) \leq \rho_2$.

Proof. Need to show for any $p, q \in \mathcal{G}_{\mathrm{TV}}$ such that $\mathrm{TV}(p, q) \leq \varepsilon$, we have $L(p, \theta^*(q)) \leq \rho_2$. Since $D = \mathrm{TV}$, Lemma 4 (Midpoint lemma) is still applicable so there exists some midpoint distribution r. Consider the following figure:

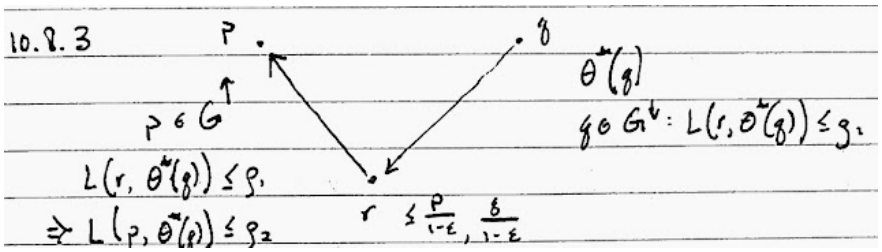


Figure 16: We use $\mathcal{G}_\downarrow$ to move $q \to r$ and $\mathcal{G}_\uparrow$ to move $r \to p$. Compare to Fig. 15, where the triangle inequality is used to combine resilience bounds between $q \to r$ and $r \to p$ to control $\|\mu(q) - \mu(p)\|_2$.

Note that $L(r, \theta^*(q)) \le \rho_1$ since $q \in \mathcal{G}_\downarrow$, and therefore by $\mathcal{G}_\uparrow$ $L(p, \theta^*(q)) \le \rho_2$. □

Example 103 (*Second moment estimation*)

Let $L(p, S) = \|S - \mathbb{E}_p[XX^\top]\|$. The resilience conditions now become:

- $\mathcal{G}_\downarrow(\rho, \varepsilon) \implies \|\mathbb{E}_r[XX^\top] - \mathbb{E}_p[XX^\top]\| \le \rho_1$ whenever $r \le \frac{p}{1-\varepsilon}$. This is just saying that $XX^\top$ is (old) (ρ_1, ε)-resilient under operator norm.
- $\mathcal{G}_\uparrow(\rho, \varepsilon) \implies \|\mathbb{E}_r[XX^\top] - S\| \le \rho_1$, $\rho_2 = 2\rho_1$, $\implies \|\mathbb{E}_p[xx^\top] - S\| \le \rho_2$.

We will verify these resilience properties when p has bounded moments.

For $\mathcal{G}_\downarrow$, $XX^\top$ is $(2\sigma\sqrt{\varepsilon}, \varepsilon)$-resilient in operator norm provided $\mathrm{Var}[\langle XX^\top, Z\rangle] \le \sigma^2$ for all $\|Z\|_* \le 1$ (Example 9 and Eq. (236)). For $\|\cdot\|$ the operator norm, the dual norm is the nuclear norm

$$\|Z\|_* = \sum_i \sigma_i(Z) \tag{343}$$

where $\sigma_i(Z)$ are the singular values of Z. For $\|Z\|_* \le 1$, the extreme points are $\pm vv^\top$ with $\|v\|_2 = 1$, so we want to show

$$\mathrm{Var}[\langle XX^\top, vv^\top\rangle] = \mathrm{Var}[\langle X, v\rangle^2] \le \mathbb{E}[|\langle X, v\rangle^2|^2] = \mathbb{E}[\langle x, v\rangle^4] \underbrace{\le}_{\text{WTS}} \sigma^2 \tag{344}$$

If we have bounded 4th moments $\mathbb{E}[\langle x, v\rangle^4]^{1/4} \le \tau$, then $\sigma = \tau^2$ and we get $(2\tau^2\sqrt{\varepsilon}, \varepsilon)$-resilience.

$\mathcal{G}_\uparrow$ is implied by $\mathcal{G}_\downarrow$ after taking $\rho_2 = \rho_1$ and using the same argument as the mean estimation example.

More generally, for any k and distributional assumptions $\mathcal{G}_{2k}$ we have $(2\sigma\varepsilon^{1-1/k}, \varepsilon)$-resilience for mean estimation and $(2\sigma^2\varepsilon^{1-2/k}, \varepsilon)$-resilience for second moment estimation.

The previous example relied on the symmetry of the loss as well as triangle inequality of the operator norm, which is not satisfied in the next setting.

Proposition 104 (*Linear regression*)

Suppose $(X, Y) \sim p$, L is excess squared loss

$$L(p, \theta) = \mathbb{E}_{(X,Y)\sim p}[(Y - \langle\theta, X\rangle)^2 - (Y - \langle\theta^*(p), X\rangle)^2] \tag{345}$$

Define the "noise" $Z = Y - \langle\theta^*(p), X\rangle$. If

Bounded noise $\mathbb{E}[XZ^2X^\top] \preceq \sigma^2\mathbb{E}[XX^\top]$

Hypercontractivity $\mathbb{E}[\langle X, v\rangle^4] \le \kappa\mathbb{E}[\langle X, v\rangle^2]^2$ *for all v*

and $\varepsilon \le 1/2$, $\varepsilon(\kappa - 1) \le 1/16$, then p is $(\rho, 8\rho, \varepsilon)$-resilient with $\rho = 3\sigma\sqrt{\varepsilon}$.

Proof. Some general observations:

- The loss is a quadratic form in the second moment matrix for p, that is:

$$L(p,\theta) = \mathbb{E}_p[(y - \langle\theta, x\rangle)^2 - (y - \langle\theta^*(p), x\rangle)^2] \tag{346}$$
$$= (\theta - \theta^*(p))^\top S_p(\theta - \theta^*(p)) \tag{347}$$
$$= \|\theta - \theta^*(p)\|_{S_p}^2 \tag{348}$$

  where $S_p = \mathbb{E}_p[XX^\top]$.

- An ε-deletion should not change too much in second moment matrix: $S_r \approx S_p$
- Nor should an ε-deletion change much in mean: $\theta^*(r) \approx \theta^*(p)$

We first use hypercontractivity to make precise $S_r \approx S_p$:

Lemma 105

{lem:linreg-second-moment-close}

If $\varepsilon(\kappa - 1) \le \frac{1}{16}$, then

$$\frac{1}{2}S_p \preceq S_r \preceq \frac{3}{2}S_p \tag{349}$$

if $r \le \frac{p}{1-\varepsilon}$

Proof. As r is an ε-deletion of p, let $r = p \mid E$ for some event with $p(E) > 1 - \varepsilon$. Then by Cauchy-Schwarz and $\varepsilon \le 1/2$, we have (note similarity between this proof and that for Corollary 5)

$$v^\top S_p v - v^\top S_r v = |\mathbb{E}_p[\langle X, v\rangle^2] - \mathbb{E}_r[\langle X, v\rangle^2]| \tag{350}$$
$$= |\mathbb{E}_p[\langle X, v\rangle^2] - \mathbb{E}_p[\langle X, v\rangle^2 \mid E]| \tag{351}$$
$$= \frac{\left|\mathbb{E}_{X\sim p}\left[\mathbb{1}_E\left(\langle X, v\rangle^2 - \mathbb{E}_{X\sim p}[\langle X, v\rangle^2\right)\right]\right|}{p(E)} \tag{352}$$
$$\le \frac{1}{1-\varepsilon}\sqrt{\mathbb{E}[\mathbb{1}_E^2]\operatorname{Var}_p[\langle x, v\rangle^2]} \tag{353}$$
$$\le 2\sqrt{\varepsilon \operatorname{Var}_p[\langle x, v\rangle^2]} \tag{354}$$

Furthermore, by hyercontractivity

$$\operatorname{Var}_p[\langle x, v\rangle^2] = \underbrace{\mathbb{E}_p[\langle x, v\rangle^4]}_{\le \kappa \mathbb{E}_p[\langle x, v\rangle^2]^2} - \mathbb{E}_p[\langle x, v\rangle^2]^2 \tag{355}$$

Hence

$$|\mathbb{E}_p[\langle x, v\rangle^2] - \mathbb{E}_r[\langle x, v\rangle^2]| \le 2\sqrt{\varepsilon(\kappa-1)}\mathbb{E}_p[\langle x, v\rangle^2] \le \frac{1}{2}\mathbb{E}_p[\langle x, v\rangle^2] \tag{356}$$

Hence $\mathbb{E}_r[\langle x, v\rangle^2] \in (1/2\mathbb{E}_p[\langle x, v\rangle^2], 3/2\mathbb{E}_p[\langle x, v\rangle^2])$ for any v. □

Next we analyze the upwards and downwards conditions for resilience. First note $\theta^*(p) = S_p^{-1}\mathbb{E}_p[XY]$ by pseudoinverse formula for least squares and

$$\theta^*(r) - \theta^*(p) = S_r^{-1}\mathbb{E}_r[XY] - S_p^{-1}\mathbb{E}_p[XY] \tag{357}$$
$$= S_r^{-1}\mathbb{E}_r[XY - XX^\top S_p^{-1}\mathbb{E}_p[XY]] \tag{358}$$
$$= S_r^{-1}\mathbb{E}_r[X(Y - \langle\theta^*(p), X\rangle)] \tag{359}$$
$$= S_r^{-1}\mathbb{E}_r[XZ] \tag{360}$$

{{eq:linreg-second-moment-diff-theta-star}}

Take for granted $\|S_r^{-1/2}\mathbb{E}_r[XZ]\|_2 \le 2\sigma\sqrt{\varepsilon}$ (will prove next time). Starting with $\mathcal{G}_\downarrow$

$$L(r,\theta^*(p)) = \|\theta^*(p)-\theta^*(r)\|_{S_r}^2 \tag{361}$$
$$= (S_r^{-1}\mathbb{E}_r[XZ])^\top S_r (S_r^{-1}\mathbb{E}[XZ]) \tag{362}$$
$$= \mathbb{E}_r[XZ]^\top S_r^{-1}\mathbb{E}_r[XZ] \tag{363}$$
$$= \|S_r^{-1/2}\mathbb{E}_r[XZ]\|_2^2 \tag{364}$$
$$\le 4\sigma^2\varepsilon = \rho \tag{365}$$

For $\mathcal{G}_\uparrow$, we want to show

$$\|\theta-\theta^*(r)\|_{S_r}^2 \le \rho \implies \|\theta-\theta^*(p)\|_{S_p}^2 \le 8\rho \tag{366}$$

Using triangle inequality on matrix norms, applying Lemma 105 to replace S_p with S_r, and using our assumption $\|\theta-\theta^*(r)\|_{S_r}^2 \le \rho$ as well as $\mathcal{G}_\downarrow$ we have

$$\|\theta-\theta^*(p)\|_{S_p} \le \|\theta-\theta^*(r)\|_{S_p} + \|\theta^*(r)-\theta^*(p)\|_{S_p} \tag{367}$$
$$\le \sqrt{2}\|\theta-\theta^*(r)\|_{S_r} + \sqrt{2}\|\theta^*(r)-\theta^*(p)\|_{S_r} \tag{368}$$
$$\le \sqrt{2\rho}+\sqrt{2\rho} \le \sqrt{8\rho} \tag{369}$$

□

12 10/15/2019

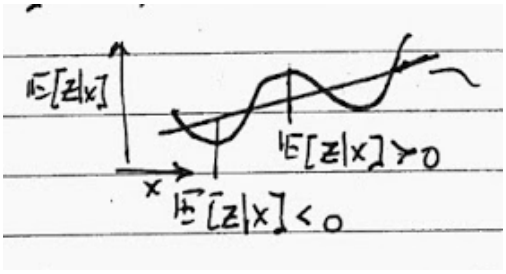
12.1 Finishing up linear regression

We first collect some useful facts about linear regression. First, by optimality of $\theta^*(p)$

$$0 = \frac{1}{2}\left.\frac{\partial L(p,\theta)}{\partial\theta}\right|_{\theta=\theta^*(p)} = \mathbb{E}[X(Y-\langle\theta^*(p),X\rangle)] = \mathbb{E}[XZ] \tag{370}$$

Furthermore, if the regression has an intercept term (e.g. if one coordinate X_i were deterministically a constant) then $\mathbb{E}[Z]=0$.

Caution: We can have $\mathbb{E}[Z \mid X] \ne 0$, which is the situation when residuals are correlated with X. For example, consider the following scenario:



The fact that $\mathbb{E}[XZ] = 0$ is what gives us the quadratic form representation for excess loss:

$$L(q,\theta) = \mathbb{E}_q[(Y - \langle\theta, X\rangle)^2] - \mathbb{E}_q[(Y - \langle\theta^*(q), X\rangle)^2] \tag{371}$$
$$= \mathbb{E}_q\left[\langle\theta, X\rangle^2 + \langle\theta^*(q), X\rangle^2 - 2Y\langle\theta - \theta^*(q), X\rangle\right] \tag{372}$$
$$= \mathbb{E}_q\left[\langle\theta - \theta^*(q), X\rangle^2 - 2\langle\theta^*(q), X\rangle\langle X, \theta^*(q) - \theta\rangle + 2Y\langle X, \theta^*(q) - \theta\rangle\right] \tag{373}$$
$$= \mathbb{E}_q\left[\langle\theta - \theta^*(q), X\rangle^2 + 2(Y - \langle\theta^*(q), X\rangle)\langle X, \theta^*(q) - \theta\rangle\right] \tag{374}$$
$$= \mathbb{E}_q\left[\langle\theta - \theta^*(q), X\rangle^2 + 2Z\langle X, \theta^*(q) - \theta\rangle\right] \to 0 \tag{375}$$
$$= \|\theta - \theta^*(q)\|_{S_q}^2 \tag{376}$$

In this setting, the resilience conditions we need to show are

- $\mathcal{G}_\downarrow$: $\|\theta^*(p) - \theta^*(r)\|_{S_r}^2 \le \rho$
- $\mathcal{G}_\uparrow$: $\|\theta^*(r) - \theta)\|_{S_r}^2 \le \rho \implies \|\theta^*(p) - \theta\|_{S_p}^2 \le 5\rho$

Last time we showed that $\mathcal{G}_\uparrow$ is implied by $\mathcal{G}_\downarrow$ and we took for granted $\|S_r^{-1/2}\mathbb{E}_r[XZ]\|_2 = \|\mathbb{E}_r[XZ]\|_{S_r} \le 2\sigma\sqrt{\varepsilon}$. This was required for verifying $\mathcal{G}_\downarrow$ because by Eq. (360) and Lemma 105

$$\|\theta^*(p) - \theta^*(r)\|_{S_r}^2 = \|\mathbb{E}_r[XZ]\|_{S_r}^2 \le 2\|\mathbb{E}_r[XZ]\|_{S_p}^2 = 2\|S_p^{-1/2}\mathbb{E}_r[XZ]\|_2^2 \tag{377}$$

The RHS is now just ordinary mean resilience in the ℓ_2 norm (note that when $r = p$ the mean is zero), so we can control it by controlling the covariance and applying Example 9. By bounding covariance with second moment and using the bounded noise hypothesis:

$$\mathrm{Cov}_p[S_p^{-1/2}XZ] \le \mathbb{E}_p[S_p^{-1/2}XZ^2X^\top S_p^{-1/2}] = S_p^{-1/2}\mathbb{E}[XZ^2X^\top]S_p^{-1/2} \tag{378}$$
$$\le \sigma^2 S_p^{-1/2}\mathbb{E}[XX^\top]S_p^{-1/2} = \sigma^2 I \tag{379}$$

12.2 Linear Classification

Let $(X, Y) \in \mathbb{R}^d \times \{\pm 1\}$. We can consider the ***classification loss***

$$L(p,\theta) = \Pr_p[Y \ne \operatorname{sgn}\langle\theta, X\rangle] \tag{380}$$

or ***hinge loss***

$$B(p,\theta) = \mathbb{E}_{X,y}\max(1 - \langle\theta, x\rangle y, 0) \tag{381}$$

Proposition 106

Suppose p satisfies

1. $\mathbb{E}_{(X,Y)\sim p}\max(1 - y\langle\theta^*, x\rangle, 0) \le (1-\varepsilon)\rho_1$
2. *For all θ satisfying*

$$\Pr[y\langle\theta, X\rangle \le \frac{1}{2}] \le \varepsilon + 2(1-\varepsilon)\rho_1 \tag{382}$$

we have

$$\Pr[y\langle\theta, X\rangle \le 0] \le \rho_2 \tag{383}$$

Then p is $(\rho_1, \rho - 2, \varepsilon)$ resilient.

which loss

?

Proof. HW □

add when done

Remark 107. The second condition is a statement about the rate of tail decay. It says that once the tail starts to decay, then it falls off very fast.

Figure 10.15.XX

Compare this to bounded variance, which was stronger and required the probability mass to be tightly concentrated (i.e. tails both fall off very fast and start to fall off close to the mean).

13 10/17/2019

13.1 Efficient Algorithms for Robust Linear Regression

- Last time
  - Resilience for linear regression (hypercontractivity and bounded noise) and classification (tails decay quickly once they decay at all)
- This time
  - Efficient algorithm for linear regression, generalizing previous down-weighting algorithm to filter wrt hypercontractivity and bounded noise
  - **Problem**: hypercontractivity is not easy to check.
    - To work around, replace with stronger condition of "SoS-certifiably hypercontractive"
    - Will need a new SoS tool called "pseudoexpectations"

13.2 Pseudoexpectations

Recall that a sum-of-squares (SoS) program looks like:

$$\max_y \ c^\top y \tag{384}$$

$$\text{st } p_y(v) \geq_{SoS} 0 \text{ (in } v) \tag{385}$$

where $p_y(v)$ is a polynomial with coefficients linear in y.

Definition 108 (*Pseudoexpectation*)

A ***degree-2k pseudoexpectation*** is a linear map $E : \{\text{degree-}2k \text{ polynomials}\} \to \mathbb{R}$ such that

1. $E[1] = 1$
2. $E[q] \geq 0$ if $q \geq_{SoS} 0$

Let $\mathcal{E}_{2k}$ be the set of all degree-$2k$ pseudoexpectations.

Remark 109. By linearity and the second property, $E[p] \leq E[q]$ if $p \leq_{SoS} q$

Example 110

$p \mapsto p(v_0)$
$p \mapsto \mathbb{E}_{v_0 \sim N(\mu,\Sigma)}[p(v_0)]$

13.2.1 Efficiency

Question: Can we efficiently impose a $\mathcal{E}_{2k}$ constraint?

Yes! Can check membership in this set efficiently (need to build a ***separation oracle***).

To check $E \in \mathcal{E}_{2k}$

$$\min_q \ E[q] \tag{386}$$

$$\text{st } q \geq_{SoS} 0 \tag{387}$$

$$\deg q \leq 2k \tag{388}$$

Can parameterize q using d^{2k} coefficients.

13.2.2 Algorithm

Setup: $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i \in [n]$, "good" set $S \subset [n]$ with $|S| = (1-\varepsilon)n$ of the points. Think of p^* as uniform on S and $\tilde{p}$ as uniform on $[n]$.

We will be generalizing the bounded covariance algorithm from before.

Need two conditions for our algorithm to work. The first is ***certifiable hypercontractivity***

$$\frac{1}{n}\sum_{i\in S} \langle X_i, v\rangle^4 \leq_{SoS} \kappa \left(\frac{1}{n}\sum_{i\in S}\langle X_i, v\rangle^2\right)^2 \tag{389}$$

and the second is a generalization of ***bounded noise*** from before

$$\frac{1}{n}\sum_{i\in S}\underbrace{(Y_i - \langle \theta^*, X_i\rangle)}_{Z_i}{}^2 X_i X_i^\top \preceq \sigma^2 \frac{1}{n}\sum_{i\in S} X_i X_i^\top \tag{390}$$

- **Input**: $(x_i, y_i)_{i\in[n]}$
- **Initialize**: $c_i = 1$ for all $i \in [n]$
- Let $q(\boldsymbol{c}) = \sum_{i\in[n]} \frac{c_i}{\sum_i c_i}\delta_{(x_i,y_i)}$ be the empirical distribution weighted by $\boldsymbol{c}$.
- Repeat: $(\hat{\theta}_c = \text{argmin}_\theta \frac{1}{n}\sum_{i=1}^n c_i (y_i - \langle\theta, x_i\rangle)^2$
  - (check hypercontractivity) Find (if possible) $E \in \mathcal{E}_4$ such that

$$E\left[\frac{1}{n}\sum_{i=1}^n c_i \langle x_i, v\rangle^4\right] \geq 3\kappa E\left[\left(\frac{1}{n}\sum_{i=1}^n c_i \langle x_i, v\rangle^2\right)\right] \tag{391}$$

  - If E exists,

$$\tau_i \leftarrow E[\langle x_i, v\rangle^4] \tag{392}$$

$$c_i \leftarrow c_i\left(1 - \frac{\tau_i}{\tau_{\max}}\right) \tag{393}$$

    and go to next loop iteration
  - (check bounded noise) Find (if possible) $v \in \mathbb{R}^d$ such that

$$\frac{1}{n}\sum_{i=1}^n c_i (y_i - \langle\hat{\theta}_c, x_i\rangle)^2 \langle x_i, v\rangle^2 \geq 24\sigma^2 \frac{1}{n}\sum_{i=1}^n c_i \langle x_i, v\rangle^2 \tag{394}$$

– If v exists,

$$\tau_i \leftarrow (y_i - \langle \hat{\theta}_c, x_i \rangle)^2 \langle x_i, v \rangle^2 \tag{395}$$

$$c_i \leftarrow c_i \left(1 - \frac{\tau_i}{\tau_{\max}}\right) \tag{396}$$

and go to next loop iteration

- Output $\hat{\theta}_c$

Each iteration repeatedly finds candidates violating hypercontractivity or bounded noise and updates c to reduce their influence.

The algorithm is guaranteed to terminate, because some c_i gets set to zero each iteration so there are at most n iterations.

The following lemma says that $q(c)$ continues to satisfies certifiably hypercontractive and bounded noise after the updates:

Lemma 111

Suppose S satisfies hypercontractivity and bounded noise conditions, and $c_i \in [0,1]$ satisfy $\frac{1}{n}\sum_{i\in S}(1 - c_i) \le \varepsilon$. Then

$$\frac{1}{n}\sum_{i\in S} c_i \langle x_i, v \rangle^4 \le_{SoS} \frac{\kappa}{1-\kappa\varepsilon}\left(\frac{1}{n}\sum_{i\in S} c_i \langle x_i, v \rangle^2\right)^2 \tag{397}$$

$$\frac{1}{n}\sum_{i\in S} c_i \langle x_i, v \rangle^2 \ge (1-\kappa\varepsilon)\frac{1}{n}\sum_{i\in S} \langle x_i, v \rangle^2 \tag{398}$$

The second result implies bounded noise with parameter $\frac{\sigma^2}{1-\kappa\varepsilon}$.

Proof sketch. For the hypercontractivity, if we define

$$A = \frac{1}{n}\sum_{i\in S} \langle x_i, v \rangle^4 \tag{399}$$

to be the initial distribution for the 4th moment over S and

$$B = \frac{1}{n}\sum_{i\in S} \langle x_i, v \rangle^2 \tag{400}$$

and if we define C and D to be the amount we're taking away

$$C = \frac{1}{n}\sum_{i\in S} (1 - c_i) \langle x_i, v \rangle^4 \tag{401}$$

$$D = \frac{1}{n}\sum_{i\in S} (1 - c_i) \langle x_i, v \rangle^2 \tag{402}$$

Then we know $\kappa B^2 - A \ge_{SoS} 0$ and $\frac{1}{\varepsilon}c - \left(\frac{1}{\varepsilon}D\right)^2 \ge_{SoS} 0$, and we want to show

$$\frac{\kappa}{1-\kappa\varepsilon}(B-D)^2 - (A-C) \ge_{SoS} 0 \tag{403}$$

We can check this is true by factoring. □

Proposition 112

If $\varepsilon \le \frac{1}{100}$ and $\kappa\varepsilon \le \frac{1}{50}$ and $q(c)$ satisfies certifiable hypercontractivity and bounded noise conditions,

then the output of the algorithm has excess squared loss $\leq 250\sigma^2\varepsilon$.

Proof structure. Same as covariance case:

1. Remove more bad points than good points (so close in TV distance)

$$\sum_{i\in S} c_i\tau_i \leq \frac{1}{2}\sum_{i=1}^n c_i\tau_i \tag{404}$$

- Hypercontractive
- Bounded noise

2. $\hat{\theta}_c$ okay if we terminate (use resilience and small $\mathrm{TV}(q(c), p^*)$

For hypercontractivity, "check hypercontractivity" filtering step terminates when does not exist pseudo-expectation $E \in \mathcal{E}_4$ refuting hyercontractivity

$$\frac{1}{n}\sum_{i\in S} c_i E[\langle x_i, v\rangle^4] \leq \frac{1}{2}\frac{1}{n}\sum_{i=1}^n c_i E[\langle x_i, v\rangle^4] \tag{405}$$

$$\frac{1}{n}\sum_{i\in S} c_i E[\langle x_i, v\rangle^4] \leq \frac{\kappa}{1-\kappa\varepsilon} E\left[\left(\frac{1}{n}\sum_{i\in S} c_i \langle x_i, v\rangle^2\right)^2\right] \tag{406}$$

$$\leq \frac{\kappa}{1-\kappa\varepsilon} E\left[\left(\frac{1}{n}\sum_{i=1}^n c_i \langle x_i, v\rangle^2\right)^2\right] \tag{407}$$

$$\frac{1}{n}\sum_{i=1}^n c_i E[\langle x_i, v\rangle^4] \geq 3\kappa E\left[\left(\frac{1}{n}\sum_{i=1}^n c_i E[\langle x_i, v\rangle^2]\right)^2\right] \tag{408}$$

so we need $\frac{\kappa}{1-\kappa\varepsilon} \leq \frac{3\kappa}{2}$, which is true since $\kappa\varepsilon \leq \frac{1}{50}$.

For bounded noise, need to show

$$\frac{1}{n}\sum_{i\in S} c_i z_i^2 \langle x_i, v\rangle^2 \leq \frac{1}{2}\frac{1}{n}\sum_{i=1}^n c_i z_i^2 \langle x_i, v\rangle^2 \tag{409}$$

We already have

$$\langle x_i, v\rangle^2 \leq 12\sigma^2 \sum_{i=1}^n c_i \langle x_i, v\rangle^2 \tag{410}$$

and the RHS is lower bounded

$$\frac{1}{2}\frac{1}{n}\sum_{i=1}^n c_i z_i^2 \langle x_i, v\rangle^2 \geq 24\sigma^2 \sum_{i=1}^n c_i \langle x_i, v\rangle^2 \tag{411}$$

Expanding z_i^2, we have

$$z_i^2 = (y_i - \langle\hat{\theta}_c, x_i\rangle)^2 \leq 2((y_i - \langle\theta^*, x_i\rangle)^2 + \langle\hat{\theta}_c - \theta^*, x_i\rangle^2) \tag{412}$$

the first term has been done and the second is $\leq \frac{\sigma^2}{1-\kappa-\varepsilon}\frac{1}{n}\sum_{i\in S} c_i \langle x_i, v\rangle^2$, so our goal is to bound

$$\underbrace{\frac{1}{n}\sum_{i\in S}\langle\hat{\theta}_c - \theta^*, x_i\rangle^2}_{(1-\varepsilon)(\text{excess predictive loss}:=R)} = \frac{1-\varepsilon}{|S|}(\theta^* - \hat{\theta}_c) \underbrace{\sum_{i\in S} x_i x_i^\top}_{\text{second moment matrix}} (\theta^* - \hat{\theta}_c) \tag{413}$$

By Cauchy-Schwarz and hypercontractivity (applied twice to move back down to 2nd powers)

$$\frac{1}{n}\sum_{i\in S} c_i(y_i - \langle\hat{\theta}_c, x_i\rangle)^2 \le \frac{\sigma^2}{1-\kappa\varepsilon}\left(\frac{1}{n}\sum_{i\in S} c_i \langle x_i, v\rangle^2\right) + \frac{1}{n}\sum_{i\in S} c_i \langle\hat{\theta}_c - \theta^*, x_i\rangle^2 \langle x_i, v\rangle^2 \tag{414}$$

Bounding the RHS terms individually

$$\frac{1}{n}\sum_{i\in S} c_i \langle\hat{\theta}_c - \theta^*, x_i\rangle^2 \langle x_i, v\rangle^2 \le \sqrt{\left(\frac{1}{n}\sum_{i\in S} c_i \langle\hat{\theta}_c - \theta^*, x_i\rangle^4\right)\left(\frac{1}{n}\sum_{i\in S} c_i \langle x_i, v\rangle^4\right)} \tag{415}$$

$$\le \frac{\kappa}{1-\kappa\varepsilon}\underbrace{\left(\frac{1}{n}\sum_{i\in S} c_i \langle\hat{\theta}_c - \theta^*, x_i\rangle^2\right)}_{(1-\varepsilon)R}\left(\frac{1}{n}\sum_{i\in S} c_i \langle x_i, v\rangle^2\right) \tag{416}$$

$$\frac{1}{n}\sum_{i\in S} c_i(y_i - \langle\hat{\theta}_c, x_i\rangle)^2 \langle x_i, v\rangle^2 \le \underbrace{\left(\frac{\sigma^2}{1-\kappa\varepsilon} + \frac{(1-\varepsilon)\kappa R}{1-\kappa\varepsilon}\right)}_{\le\sigma^2/2}\left(\frac{1}{n}\sum_{i\in S} c_i \langle x_i, v\rangle^2\right) \tag{417}$$

We can show

$$R \le \frac{10\varepsilon\tilde{\sigma}_c^2}{1-\varepsilon} \tag{418}$$

if $\varepsilon \le \frac{1}{8}$ and $\frac{\varepsilon(\tilde{\kappa}-1)}{3\kappa} \le \frac{1}{6}$. So we are happy if

$$\frac{\sigma^2}{1-\kappa\varepsilon} + \frac{10\kappa\varepsilon\tilde{\sigma}^2}{1-\kappa\varepsilon} \le \frac{\tilde{\sigma}^2}{2} \tag{419}$$

Doing the algebra, we will find that this holds if $\tilde{\sigma}^2 \ge 256\sigma^2$. □

14 10/23/2019 and 10/25/2019

Was travelling, see course notes for generalizing from $D = \text{TV}$ to Wasserstein distances (need ε-deletion $\to$ ε-"friendly" and a generalized midpoint lemma).

15 10/29/2019

15.1 Setting for test-time robustness (classification)

- Train: $(x_1, y_1), \ldots, (x_n, y_n) \sim p$, $y \in Y$ ($Y = \{\pm 1\}$ binary, $Y = [K]$ multiclass)
- Test: $(x, y) \sim p$. Observe $\bar{x}$ such that $d(x, \bar{x}) \le \varepsilon$ (before $D(\tilde{p}, p^*) \le \varepsilon$)
- Goal: Predict y from $\bar{x}$

For various $\ell(\theta; x, y)$, e.g.:

- 0/1-loss, measure accuracy of classifier
- log-loss / hinge-loss: measure "margin" of classification

We were previously using a (non-robust) loss

$$L(p, \theta) = \mathbb{E}_{(x,y)\sim p}[\ell(\theta; x, y)] \tag{420}$$

Consider instead a ***robust loss***

$$L(p, \theta) = \mathbb{E}_{(x,y)\sim p}\left[\sup_{\bar{x}: d(x,\bar{x})\le\varepsilon} \ell(\theta; \bar{x}, y)\right] \tag{421}$$

We care about this kind of loss because most models are very non-robust (Figure 10.29.1: a 1/128 perturbation in ℓ_∞ causes a DNN to become very confused)

15.1.1 Relation to train-time robustness

Before we thought of the training distribution $\tilde{p}$ as corrupted, and want to ensure performance on the test-time distribution p^*. Here, we think of the train distribution p as nice and the test distribution $\bar{p}$ as corrupted.

For classification, we can formalize test-time corruption in terms of discrepancy as follows: $D(\tilde{p}, q) \le \varepsilon$ if there exists a coupling π from $\tilde{p}$ to q such that for $(x, y), (x', y') \sim \pi$ if $y = y'$ then $d(x, x') \le \varepsilon$ almost surely.

This generalizes Wasserstein distance

$$w_c(p, q) = \inf_{\pi \in \Pi} \mathbb{E}_\pi[c(x, x')^k]^{1/k} \tag{422}$$

Notice here that $\mathcal{G}$ is not needed.

15.1.2 Natural algorithm

A natural algorithm is to just minimize empirical loss. For $(x_i, y_i) \overset{\text{iid}}{\sim} p$, fit our model by

$$\min_\theta \rho(\theta) + \frac{1}{n} \sum_{i=1}^n \sup_{\bar{x}_i : d(x_i, \bar{x}_i) \le \varepsilon} \ell(\theta; \bar{x}_i, y_i) \tag{423}$$

A couple of issues arise:

1. sup over $\bar{x}$
2. Generalizataion
3. What if d wasn't what you cared about?

Some examples of different d we can care about.
Figure 10.29.2

- ℓ_∞ : each pixel $\le \varepsilon$
- ℓ_1: total change $\le \varepsilon$
- ℓ_2
- JPEG: $\|JPEG(x) - JPEG(\bar{x})\|_2 \le \varepsilon$
- $\bar{x}$ obtained from x via elastic warping of size ε
- Fog
- Snow

15.2 Sup over $\bar{x}$

The robust loss requires us to compute

$$\sup_{\bar{x} : d(x, \bar{x}) \le \varepsilon} \ell(\theta; \bar{x}, y) \tag{424}$$

Unfortunately, ℓ is usually not convex in $\bar{x}$ and is hard to optimize.

One strategy is to heuristically maximize $\bar{x}$ by just running a bunch of gradient steps.

$$\ell_{\text{robust}} = \sup_{\bar{x}} \ell(\theta; \bar{x}, y) \tag{425}$$

$$\ell_{\text{proxy}} = \text{heuristic max} \tag{426}$$

$$\ell_{\text{proxy}} \le \ell_{\text{robust}} \tag{427}$$

This is not a good idea: Figure 10.29.3: minimizing θ over the proxy will choose points where the proxy is not a good approximation to ℓ_{robust}

16 10/31/2019

16.1 Certified adversarial training

Recap last time

- Adversarial robustness / fragility
  - Robustness is hard to evaluate: most papers are wrong, follow best practices
  - Robustness seems to require more data
  - Random vs adversarial noise are very different in high dimensions
  - Robustness under different perturbation types partially (but doesn't completely) transfer

Today we will resolve this first issue that robustness is hard to evaluate, using a method called ***certified adversarial training***.

Recall that our goal is to minimize the robust loss

$$\mathbb{E}_{x,y}\left[\underbrace{\sup_{\bar{x}:d(x,\bar{x})\leq\varepsilon}\ell(\theta;\bar{x},y)}_{=\ell_{\text{robust}}(\theta;\bar{x},y)}\right] \tag{428}$$

We will (somehow) obtain $\ell_{\text{cert}} \geq \ell_{\text{robust}}$, after which we can minimize $\mathbb{E}\ell_{\text{cert}}$.

To start: why is an upper bound on ℓ_{robust} better than a lower bound (e.g. ℓ_{proxy} from before)?

Figure 10.31.1: minimization in θ of an upper bound prefers spots where the upper bound is a good approximation, whereas it favors regions where the lower bound is bad!

What should we do if we know ℓ_{robust}? If d and ℓ are sufficiently nice, then by the ***envelope theorem***

$$\nabla_\theta \sup_{\bar{x}:d(x,\bar{x})\leq\varepsilon}\ell(\theta;\bar{x},y) = \nabla_\theta\ell(\theta;x^*,y)|_{x^*=\text{argmax}_{\bar{x}:d(x,\bar{x})\leq\varepsilon}\ell(\theta;\bar{x},y)} \tag{429}$$

Sufficient conditions include $B = \{d(x,\bar{x}) \leq \varepsilon\}$ compact and ℓ continuously differentiable in $(\theta,\bar{x})$. More generally, see Danskin's theorem (subgradient = convex hull of gradients of argmaxes).

16.1.1 Adversarial training

The previous envelope theorem says that the gradient of the worst case loss is equal to the gradient of the loss at the worst case example x^*. This motivates adversarial training:

- Approximately optimize over $\bar{x}$
- Take step in direction $\nabla_\theta\ell(\theta;\bar{x},y)$

This is a continuation of last time's ℓ_{proxy}, which is an under-approximation.

16.1.2 Certified adversarial training

Consider a feedforward neural network

$$z^{(i)} = A^{(i-1)}x^{(i-1)} \tag{430}$$

$$x^{(i)} = \sigma(z^{(i)}) \tag{431}$$

where $\sigma(z) = \max(z,0)$ is ReLU activation and $x^{(0)} = x$ is the input. Consider cross-entropy loss

$$\ell(\theta;x^{(0)},y) = \log\left(\sum_{y'\in[k]}e^{z_{y'}^{(L)}}\right) - z_y^{(L)} \tag{432}$$

with $y \in [k]$.

Goal: Fix y and y'. Upper bound (at some fixed $\theta = (A^{(i)})_{i\in[L-1]}$)

$$\max_{x^{(i)},z^{(i)},i\in[L]} z_{y'}^{(L)} - z_y^{(L)} \tag{433}$$

$$\text{st } d(x, x^{(0)}) \le \varepsilon \tag{434}$$

$$z^{(i)} = A^{(i-1)}x^{(i-1)} \quad \forall i \tag{435}$$

$$x^{(i)} = \max(z^{(i)}, 0) \quad \forall i \tag{436}$$

In the case $d = \ell_\infty$, we can rewrite $d(x, x^{(0)}) \le \varepsilon$ as linear inequalities

$$x_j^{(0)} \le x_j + \varepsilon \tag{437}$$

$$x_j^{(0)} \ge x_j - \varepsilon \tag{438}$$

for all j. Similarly, we have linear equalities

$$z^{(i)} = A^{(i-1)}x^{(i-1)} \tag{439}$$

The problematic constraint is the equality constraint $x^{(i)} = \max(z^{(i)}, 0)$, which is not a convex region. How do we write this as something convex?

Figure 10.31.2

One way is to relax to inequality $x \ge \max(z, 0)$, but this is not great because we could have $x \to \infty$.

Approach 1: LP relaxation. Suppose we knew $z \in [l, u]$. Then, we can consider $x \ge z$, $x \ge 0$, $x \le \frac{z-l}{u-l}u$

Figure 10.31.3

This gives an LP we can solve to get a bound for each layer $i \in [L]$.

While not very efficient, there are speed-up tricks:

- Take the dual
- Guess good variables for intermediate layers
- Take gradient descent tsteps in dual instead of solving all the way
- Train "verifier network" that guesses dual variables

Approach 2: SDP. Again focus on $x = \max(z, 0)$. Rewrite as system of polynomial constraints

$$x \ge z \tag{440}$$

$$x \ge 0 \tag{441}$$

$$x(x - z) = 0 \tag{442}$$

Trick for Grothendieck

$$M = \begin{bmatrix} 1 & x & z \\ x & x^2 & xz \\ z & xz & z^2 \end{bmatrix} = \begin{bmatrix} 1 & m_x & m_z \\ m_x & m_{xx} & m_{xz} \\ m_z & m_{xz} & m_{zz} \end{bmatrix} \tag{443}$$

Our constraints are equivalent to

$$m_x \ge m_z \tag{444}$$

$$m_x \ge 0 \tag{445}$$

$$m_{xx} = m_{xz} \tag{446}$$

$$M \succeq 0 \tag{447}$$

$$\text{rank}(M) = 1 \tag{448}$$

We also know $M \succeq 0$, because it's an outer product of a vector with itself.

Recall Grothendieck, which rewrote quadratic forms

$$\max y^\top \Sigma y \tag{449}$$
$$\text{st } y \in \{\pm 1\}^m \tag{450}$$
$$M = yy^\top \tag{451}$$

as big SDPs

$$\max \langle M, \Sigma \rangle \tag{452}$$
$$\text{st } M \succeq 0 \tag{453}$$
$$\text{diag}(M) = 1 \tag{454}$$

Decision variables $m^{(i)}$, $m^{(i,i)}$, $m^{(i,i-1)}$

Constraints

$$m^{(i)} \geq 0 \tag{455}$$
$$m^{(i)} \geq A^{(i-1)} m^{(i-1)} \tag{456}$$
$$\text{diag}(m^{(i,i)} - m^{(i,i-1)}(A^{(i-1)})^\top) = 0 \tag{457}$$
$$\begin{bmatrix} 1 & (m^{(i-1)})^\top & m^{(i)\top} \\ m^{(i-1)} & m^{(i-1,i-1)} & (m^{(i,i-1)})^\top \\ m^{(i)} & m^{(i,i-1)} & m^{(i,i)} \end{bmatrix} \succeq 0 \tag{458}$$

17 11/5/2019

17.1 Randomized smoothing

$$f_\theta : \mathcal{X} \to \Delta_k = [0,1]^k \tag{459}$$

Noise π, define ***smoothed classifier***

$$\bar{f}_\theta(x) = \mathbb{E}_{\delta \sim \pi} f_\theta(x + \delta) \tag{460}$$

Approximate with error ε in ℓ_∞ by drawing $O(1/\varepsilon^2)$ samples from , $O(\log k/\varepsilon^2)$ for ℓ_2?

Proposition 113

Let π_x be distribution of $x + \delta$. For any inputs x, x'

$$\|\bar{f}_\theta(x) - \bar{f}_\theta(x')\|_\leq \text{TV}(\pi_x, \pi_{x'}) \tag{461}$$

In particular, if $\sup_{\bar{x}:d(x,\bar{x})\leq\varepsilon} \text{TV}(\pi_x, \pi_{\bar{x}}) \leq \tau$ and correct class $\bar{f}_\theta(x) \geq 2\tau$ larger than any incorrect class, i.e. $y \in [k]$ and actual class y then for all $y' \neq y$

$$\bar{f}_\theta(x)_y \geq 2\tau + \bar{f}_\theta(x)_{y'} \tag{462}$$

then $\bar{f}_\theta$ is **adversarially robust** *at x.*

Proof.

$$|\bar{f}_\theta(x)_j - \bar{f}_\theta(x')_j| = |\mathbb{E}_{z\sim\pi_x}[f_\theta(z)_j] - \mathbb{E}_{z\sim\pi_{x'}}[f_\theta(z)_j]| \leq \text{TV}(\pi_x, \pi_{x'}) \tag{463}$$
$$\text{TV}(p,q) = \sup_{f:\mathcal{X}\to[0,1]} |\mathbb{E}_p[f] - \mathbb{E}_q[f]| \tag{464}$$

Remaining questions:

- Pick π.
- How to train $\bar{f}_\theta$.

Let $d(x, \bar{x}) = \|x - \bar{x}\|_2$, $\pi = N(0, \sigma^2 I)$. Then

$$\sup_{\|x-x'\|_2 \le \varepsilon} \mathrm{TV}(\pi_x, \pi_{x'}) = \sup_{\|x-x'\|_2 \le \varepsilon} \mathrm{TV}(N(x, \sigma^2 I), N(x', \sigma^2 I)) \tag{465}$$

$$= \mathrm{TV}(N(-\varepsilon/2, \sigma^2), N(\varepsilon/2, \sigma^2)) \tag{466}$$

$$= \Phi(\varepsilon/2\sigma) - \Phi(-\varepsilon/2\sigma) \tag{467}$$

where Φ is the normal CDf.

Figure 11.5.1

$$\int_{-\varepsilon/2}^{\infty} p_\sigma(x)dx - \int_{\varepsilon/2}^{\infty} p_\sigma(x)dx = \Phi(\varepsilon/2\sigma) - \Phi(-\varepsilon/2\sigma) \tag{468}$$

We previously showed the above quantity $\Theta(\varepsilon/\sigma)$ if $\varepsilon/sigma$ is small. where?

Similar to calculation in linear case: need lots of noise ($\sigma\sqrt{d}$ vs σ)

How to do training? Traditionally to get f_θ we

$$\min_\theta \mathbb{E}_{(x,y)\sim p}[-\log f_\theta(x)_y] \tag{469}$$

For this randomized smoothing classifier, could consider

$$\min_\theta \mathbb{E}_{(x,y)\sim p}[-\log \bar{f}_\theta(x)_y] \tag{470}$$

But this is not the most convenient object. To see why, Fisher scoring requires computing

$$\nabla_\theta[\log \bar{f}_\theta(x)_y] = \frac{1}{\bar{f}_\theta(x)_y} \nabla_\theta \bar{f}_\theta(x)_y \tag{471}$$

$$= \frac{1}{\bar{f}_\theta(x)_y} \nabla_\theta \mathbb{E}_{\delta\sim\pi}[f_\theta(x+\delta)_y] \tag{472}$$

$$= \frac{1}{\bar{f}_\theta(x)_y} \mathbb{E}_{\delta\sim\pi}[\nabla_\theta f_\theta(x+\delta)_y] \tag{473}$$

$$= \frac{1}{\bar{f}_\theta(x)_y} \mathbb{E}_{\delta\sim\pi}[f_\theta(x+\delta)_y \nabla_\theta \log f_\theta(x+\delta)_y] \tag{474}$$

$$= \mathbb{E}_{\delta\sim\pi}[\frac{f_\theta(x+\delta)_y}{\bar{f}_\theta(x)_y} \nabla_\theta \log f_\theta(x+\delta)_y] \tag{475}$$

Notice $\nabla_\theta \log f_\theta(x+\delta)_y$, the typical gradient term in maximum likelihood. Think of $\frac{f_\theta(x+\delta)_y}{\bar{f}_\theta(x)_y}$ as an importance weight.

Another way we could have generalized the loss is

$$-\log \bar{f}_\theta(x)_y = -\log \mathbb{E}_\delta[f_\theta(x+\delta)_y] \tag{476}$$

resulting in

$$\mathbb{E}_{(x,y)\sim p}\mathbb{E}_\delta[-\log f_\theta(x+\delta)_y] \tag{477}$$

which is saying that ***all*** of the δ have high probability of the true label. The derivative is much nicer:

$$\nabla_\theta \mathbb{E}_{(x,y)\sim p}\mathbb{E}_\delta[-\log f_\theta(x+\delta)_y] = \mathbb{E}_{(x,y)\sim p}\mathbb{E}_\delta[\nabla_\theta[-\log f_\theta(x+\delta)_y]] \tag{478}$$

So to train, we (1) sample a training point (x, y) then (2) sample a pertubation δ and use the gradient at $-\log f_\theta(x+\delta)_y$. By Jension's, this is also an upper bound on the previous. □

17.2 Covariate shifts

Training $\tilde{p}$ and test p^*, ***covariate shift*** is an assumption

$$\tilde{p}(y \mid x) = p^*(y \mid x) \tag{479}$$

for all x, y.

It says that "with infinite data I'm fine" because

$$\tilde{p}(x, y) \to \tilde{p}(y \mid x) \tag{480}$$
$$p^*(x) \quad p^*(y \mid x) \tag{481}$$

and $p^*(y \mid x)$ is what you need to predict well.

Two issues:

- Extrapolating to low probaility regions under p^*: Figure 11.5.2. Will address with importance weighting.
- Handling model mis-specification.

18 11/7/2019

18.1 Propensity weighting

Here we consider covariate shifts and will use propensity score importance weighting to handle extrapolation to low probability regions under p^*.

For now asusme p^* and $\tilde{p}$ known, $\ell(\theta; x, y)$ squared loss $(y - \langle x, \theta \rangle)^2$ or $\log(1 + \exp(-y \langle \theta, x \rangle))$.

By the covariate shift assumption

$$\mathbb{E}_{(x,y)\sim p^*}[\ell(\theta; x, y)] = \mathbb{E}_{(x,y)\sim \tilde{p}}\left[\frac{p^*(x,y)}{\tilde{p}(x,y)}\ell(\theta; x, y)\right] = \mathbb{E}_{(x,y)\sim \tilde{p}}\left[\frac{p^*(x)\cancel{p^*(y \mid x)}}{\tilde{p}(x)\cancel{\tilde{p}(y \mid x)}}\ell(\theta; x, y)\right] \tag{482}$$

From which we see that the loss under p^* is equal to the ***propensity weighted loss***

$$\mathbb{E}_{(x,y)\sim \tilde{p}}\left[\frac{p^*(x)}{\tilde{p}(x)}\ell(\theta; x, y)\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{p^*(x_i)}{\tilde{p}(x_i)}\ell(\theta; x_i, y_i) \tag{483}$$

where the examples $(x, y) \sim \tilde{p}$ are from the training distribution.

Few concerns:

1. Variance
2. How to get $p^*(x)$ and $\tilde{p}(x)$?
3. Finite-sample corrections for learning θ (relates to first point)

Looking at the variance

$$\mathrm{Var}_{\tilde{p}}\left[\frac{p^*(x)}{\tilde{p}(x)}\ell(\theta; x, y)\right] = \mathbb{E}_{\tilde{p}}\left[\frac{p^*(x)^2}{\tilde{p}(x)^2}\ell(\theta; x, y)^2\right] - \underbrace{\mathbb{E}_{\tilde{p}}\left[\frac{p^*(x)}{\tilde{p}(x)}\ell(\theta; x, y)\right]^2}_{\mathbb{E}_{p^*}[\ell(\theta;x,y)]^2 \le 0} \tag{484}$$

Assume $\ell(\theta; x, y) \le B$. Then

$$\mathbb{E}_{\tilde{p}}\left[\frac{p^*(x)^2}{\tilde{p}(x)^2}\ell(\theta; x, y)^2\right] = \mathbb{E}_{\tilde{p}}\left[\frac{p^*(x)^2}{\tilde{p}(x)^2}\right]B^2 = \left(D_{\chi^2}(\tilde{p}||p^*) + 1\right)B^2 \tag{485}$$

So we see that the variance is controlled by the chi-square divergence.

Definition 114 (*Chi-square divergence*)

The ***chi-square divergence***

$$D_{\chi^2}(\tilde{p}||p^*) = \int \frac{p^*(x)^2}{\tilde{p}(x)} dx - 1 \tag{486}$$

18.1.1 Properties of chi-square

Proposition 115 (*KL lower bounds chi-square*)

$$D_{\chi^2}(p||q) \geq KL(q||p) \tag{487}$$

Proof.

$$\int q(x) \log \frac{q(x)}{p(x)} dx = \int q(x) \left(\log q(x) - \log p(x)\right) dx \leq \int \frac{q(x)}{p(x)} \left(q(x) - p(x)\right) dx \tag{488}$$

□

Similar to how KL-divergence satisfies a distributivity across product measures:

Proposition 116 (*KL of products*)

$$KL\left(\prod_{i=1}^n p_i || \prod_{i=1}^n q_i\right) = \sum_i KL(p_i||q_i) \tag{489}$$

Proof.

$$\int \prod^n p_i(x_i) \log \frac{\prod^n p_i(x_i)}{\prod^n q_i(x_i)} \prod^n dx_i = \int \prod^n p_i(x_i) \left(\sum^n \log \frac{p_i}{q_i}\right) \prod^n dx_i \tag{490}$$

$$= \sum_i \int p_i(x_i) \log \frac{p_i(x_i)}{q_i(x_i)} \tag{491}$$

$$= \sum_i D_{KL}(p_i||q_i) \tag{492}$$

□

A similar property holds for D_{χ^2}, though we recover a product rather than a sum:

Proposition 117 (*Chi-square divergence of products*)

$$D_{\chi^2}\left(\prod_{i=1}^n p_i || \prod_{i=1}^n q_i\right) = \prod_i (1 + D_{\chi^2}(p_i||q_i)) - 1 \tag{493}$$

Proof.

$$D_{\chi^2}(\prod^n p_i || \prod^N q_i) = \int \frac{\prod_i p_i(x_i)^2}{\prod_i q_i(x_i)} \prod^n dx_i - 1 \tag{494}$$

$$= \left(\prod_{i=1}^n \int \frac{p_i(x_i)^2}{q_i(x_i)} dx \right) - 1 \tag{495}$$

$$= \left(\prod_{i=1}^n D_{\chi^2}(q_i || p_i) \right) + 1 \tag{496}$$

□

Proposition 118 (*Chi-square divergence for Gaussians*)

$$D_{\chi^2}(N(\mu, I), N(\mu', I)) = \exp(\|\mu - \mu'\|_2^2) \tag{497}$$

Proof. Let Z be the normalization constant.

$$\frac{1}{Z} \int \frac{\exp(-\|x-\mu\|_2^2)}{\exp(-\frac{1}{2}\|x-\mu'\|_2^2)} dx = \frac{1}{Z} \int \exp\left(-\frac{1}{2}\|x\|_2^2 + \langle 2\mu - \mu', x \rangle - \|\mu\|_2^2 + \frac{1}{2}\|\mu'\|_2^2 \right) \tag{498}$$

$$= \frac{1}{Z} \int \exp\left(-\frac{1}{2}\|x - (2\mu - \mu')\|_2^2 + \frac{1}{2}\|2\mu - \mu'\|_2^2 - \|\mu\|_2^2 + \frac{1}{2}\|\mu'\|_2^2 \right) \tag{499}$$

$$= \exp\left(\frac{1}{2}\|2\mu - \mu'\|_2^2 - \|\mu\|_2^2 + \frac{1}{2}\|\mu'\|_2^2 \right) \tag{500}$$

$$= \exp\left(\|\mu\|_2^2 + \|\mu'\|_2^6 2 - 2\langle \mu, \mu' \rangle \right) \tag{501}$$

$$= \exp(\|\mu - \mu'\|_2^2) \tag{502}$$

□

18.2 Causal inference

Suppose X were patients, $t \in \{0, 1$ an indicator for whether treatment was given, and Y the outcome.

We want to estimate the ***treatment effect***

$$\mathbb{E}[Y \mid t = 1] - \mathbb{E}[y \mid t = 0] \tag{503}$$

This would be easy if we could perform randomized trials, but in historical data treatment may only be given to certain X.

To handle this, consider potential outcomes $Y(0)$ and $Y(1)$ for a patient X which represent the outcome that would have happened if the patient did / didn't receive treatment.

We make an ***unconfoundedness assumption***

$$Y(0), Y(1) \perp t \mid X \tag{504}$$

Reduction to covariate shifts: let $\tilde{p}(x, t, y(t))$ be the observational distribution and

$$p_1^*(x, t, y(t)) = p^*(x) \underbrace{p(t \mid x)}_{\mathbb{1}\{t=1\}} p^*(y(t) \mid x, t) \tag{505}$$

$$p_0^*(x, t, y(t)) = p^*(x) \underbrace{p(t \mid x)}_{1 - \mathbb{1}\{t=1\}} p^*(y(t) \mid x, t) \tag{506}$$

two distributions where everyone did / didn't get the treatment.

Proposition 119

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_{p_1^*}[Y(t)] - \mathbb{E}_{p_0^*}[Y(t)] \tag{507}$$

Proposition 120

$\tilde{p}$ and p_1^* satisfy covariate (x, t) shift, that is

$$\tilde{p}(y(0), y(1) \mid x, t) = p^*(y(0), y(1) \mid x, t) \tag{508}$$

$$\mathbb{E}_{p_1^*}[Y(t)] = \mathbb{E}_{\tilde{p}}\left[\frac{p_1^*(x,t)}{\tilde{p}(x,t)} Y(t)\right] \tag{509}$$

$$= \mathbb{E}_{\tilde{p}}\left[\frac{\cancel{p_1^*(x)}}{\cancel{\tilde{p}(x)}} \frac{p_1^*(t \mid x)}{\tilde{p}(t \mid x)} Y(t)\right] \tag{510}$$

$$= \mathbb{E}_{\tilde{p}}\left[\frac{\mathbb{1}\{t=1\}}{\tilde{p}(t=1 \mid x)} Y(1)\right] \tag{511}$$

$$\mathbb{E}_{p_0^*}[Y(0)] = \mathbb{E}_{\tilde{p}}\left[\frac{\mathbb{1}\{t=0\}}{p(t=0 \mid x)} Y(0)\right] \tag{512}$$

$$\mathbb{E}_{p_1^*}[Y(t)] - \mathbb{E}_{p_0^*}[Y(t)] = \mathbb{E}_{\tilde{p}}\left[\left(\frac{\mathbb{1}\{t=1\}}{\tilde{p}(t=1 \mid x)} - \frac{\mathbb{1}\{t=0\}}{\tilde{p}(t=0 \mid x)}\right) Y(t)\right] \tag{513}$$

$$\approx \frac{1}{n} \sum_{i=1}^{n} y_i \left(\frac{\mathbb{1}\{t=1\}}{\tilde{p}(t=1 \mid x)} - \frac{\mathbb{1}\{t=0\}}{\tilde{p}(t=0 \mid x)}\right) \tag{514}$$

We can interpret the effect of this as doing a reweighting which "undoes" the propensity to treat.

This is called ***inverse propensity weighting***, so we see that domain adaptation under covariate shifts and inverse propensity weighting are the same thing.

19 11/12/2019

- Last time
  - Domain adaptation and covariate shifts
  - Propensity weighting: works if D_{χ^2} small
  - Causal inference: inverse propensity weighting (IPV)
- This time
  - Review IPW
  - Better estimator: doubly-robust estimator
    - Start w/ model
  - Semi-parametric estimator

Recall the causal inference setup: observe $(X, T, Y(T))$. Randomized trials have $\Pr(T = 1 \mid X) = 1/2$, but more available observational settings often have $P(T = 1 \mid X)$ depends on X.

To deal with this, we make an ***unconfoundedness assumption***: $T \perp Y \mid X$. As a result, the ***average treatment effect***

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X[\mathbb{E}[Y(1) \mid X] - \mathbb{E}[Y(0) \mid X]] \tag{515}$$

$$= \mathbb{E}_X[\mathbb{E}[Y(1) \mid T = 1, X] - \mathbb{E}[Y(0) \mid T = 0, X]] \tag{516}$$

This is nice because we only ever observe one of the two potential outcomes: $(Y(1), T = 1)$ or $(Y(0), T = 0)$. Further expanding these terms reveals:

$$\mathbb{E}_{X\sim\tilde{p}}\mathbb{E}_{\tilde{p}}[Y(1) \mid T = 1, X] = \mathbb{E}_{p_1^*}[\mathbb{E}[Y(1) \mid X]] = \mathbb{E}_{p_1^*}[Y(1)] \tag{517}$$

$$p_1^*(x, t, y) = \tilde{p}(x)\mathbb{1}\{t = 1\}\tilde{p}(y \mid x, t = 1) \tag{518}$$

where we have extracted the propensity score reweighted distributions p_1^* and p_0^*.

Importance sampling allows us to go from $\tilde{p}$ to p_1^*, giving inverse propensity weighting (IPW):

$$\mathbb{E}_{p_1^*}[Y(1)] = \mathbb{E}_{\tilde{p}}\left[\frac{p_2^*(x,t)}{\tilde{p}(x,t)}Y(1)\right] = \mathbb{E}_{\tilde{p}}\left[\frac{\mathbb{1}\{t=1\}}{\tilde{p}(t=1 \mid X)}Y(1)\right] \tag{519}$$

$$\mathbb{E}_{\tilde{p}}\left[Y(1) - Y(0)\right] = \mathbb{E}_{\tilde{p}}\left[\left(\frac{\mathbb{1}\{T=1\}}{\tilde{p}(t=1 \mid X)} - \frac{\mathbb{1}\{T=0\}}{\tilde{p}(t=0 \mid X}\right)Y(T)\right] \tag{520}$$

All of the terms on the RHS are observable quantities, making causal inference tractable.

Definition 121

The ***IPW estimator*** of ATE is

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}\left[\left(\frac{\mathbb{1}\{T=1\}}{\tilde{p}(T=1 \mid X)} - \frac{\mathbb{1}\{T=0\}}{p(T=0 \mid X)}\right)Y(T)\right] \tag{521}$$

Some issues for this estimator are

- High variance: denominator could be small
- Requires knowing $\tilde{p}$ in advance; not robust to mis-specification

Idea: consider a predictor for Y and $\tilde{p}$

$$\bar{Y}(0, X) \approx \mathbb{E}[Y(0) \mid X] \tag{522}$$

$$\bar{Y}(1, X) \approx \mathbb{E}[Y(1) \mid X] \tag{523}$$

$$f(X) = \hat{p}(t = 1 \mid X) \approx \tilde{p}(t = 1 \mid X) \tag{524}$$

We can decompose the ATE as

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\bar{Y}(1, X) - \bar{Y}(0, X)] + \mathbb{E}\left[\underbrace{Y(1) - \bar{Y}(1, X)}_{\hat{Y}(1)} - \underbrace{Y(0) - \bar{Y}(0, X)}_{\hat{Y}(0)}\right] \tag{525}$$

We can try to estimate that $\hat{Y}$ using IPW with $\hat{p}$:

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\bar{Y}(1, X) - \bar{Y}(0, X)] + \mathbb{E}\left[\left(\frac{\mathbb{1}\{T=1\}}{\hat{p}(t=1 \mid X)} - \frac{\mathbb{1}\{T=0\}}{\hat{p}(t=0 \mid X)}\right)\underbrace{(Y(T) - \bar{Y}(T, X))}_{\ll Y(T)}\right] \tag{526}$$

Bias of the estimator

$$\mathbb{E}[Y(1) - Y(0)] - \mathbb{E}[\bar{Y}(1, X) - \bar{Y}(0, X)] - \mathbb{E}\left[\left(\frac{\mathbb{1}\{T=1\}}{\hat{p}(t=1 \mid X)} - \frac{\mathbb{1}\{T=0\}}{\hat{p}(t=0 \mid X)}\right)(Y(T) - \bar{Y}(T, X))\right] \tag{527}$$

$$= \mathbb{E}\left[(Y(1) - \bar{Y}(1, X))\left(1 - \frac{\mathbb{1}\{T=1\}}{\hat{p}(t=1 \mid X)}\right) - (Y(0) - \bar{Y}(0, X))\left(1 - \frac{\mathbb{1}\{T=0\}}{\hat{p}(t=0 \mid X)}\right)\right] \tag{528}$$

Focusing on just the $Y(1)$ term

$$\mathbb{E}_X\left[\mathbb{E}\left[(Y(1) - \bar{Y}(1, X) \mid X\right]\left(1 - \frac{\mathbb{1}\{T=1\}}{\hat{p}(t=1 \mid X)}\right)\right] \tag{529}$$

This bias term is zero (and the other $Y(0)$ term as well) if either $\bar{Y}$ or $\hat{p}$ is correct, giving the name ***doubly-robust estimator***.

Denoting $Y^*(1,X) = \mathbb{E}[Y(1) \mid X]$, we have by Cauchy-Schwarz

$$\mathbb{E}_X\left[\mathbb{E}\left[(Y(1) - \bar{Y}(1,X) \mid X\right]\left(1 - \frac{\mathbb{1}\{T=1\}}{\hat{p}(t=1 \mid X)}\right)\right] \tag{530}$$

$$= Y^*(1,X) - \bar{Y}(1,X) \tag{531}$$

$$\leq \underbrace{\mathbb{E}_X[(Y^*(1,X) - \bar{Y}(1,X))^2]^{1/2}}_{\text{average squared error } \bar{Y}} \underbrace{\mathbb{E}_X\left[\left(1 - \frac{\bar{p}(t=1 \mid X)}{\hat{p}(t=1 \mid X)}\right)^2\right]^{1/2}}_{\text{average squared error } \hat{p}} \tag{532}$$

For the variance, we will focus only on the dominant second term for $T = 1$

$$\mathrm{Var}\left[\left(\frac{\mathbb{1}\{T=1\}}{\hat{p}(t=1 \mid X)}\right)(Y(T) - \bar{Y}(T,X))\right] \leq \mathbb{E}\left[\frac{\mathbb{1}\{T=1\}}{\hat{p}(t=1 \mid X)^2}(Y(1) - \bar{Y}(1,X))^2\right] \tag{533}$$

$$= \mathbb{E}_X\left[\frac{\tilde{p}(t=1 \mid X)}{\hat{p}(t=1 \mid X)^2}\mathbb{E}[(Y(1) - \bar{Y}(1,X))^2]\right] \tag{534}$$

Some common sources of noise here are:

- Intrinsic variance
- Rare treatement error (real $\tilde{p}$ or predicted $\hat{p}$)

19.1 Non-parametric regression

19.1.1 Rates of convergence for non-parametric regression

Consider the RMSE

$$\mathbb{E}[(y - \hat{f}(x))^2]^{1/2} \tag{535}$$

where $\hat{f}(x)$ is the function we are trying to estimate.

For linear regression (parametric): $\hat{f}(x) = \langle w, x\rangle$ and the RMSE decays with $n^{-1/2}$ (assuming asymptotic normality).

In non-parametric regression: $d = 1$ and $\hat{f}$ Lipschitz, we get $n^{-1/3}$. In general, we get $n^{-\alpha}$ with α depending on dimension and smoothness (derivatives)

$$\alpha = \frac{\beta}{2\beta + d} \tag{536}$$

In the regime where $\alpha > 1/4$, then we can get parametric rates even if we use a non-parametric estimator!

20 11/14/2019

- Recap last time
  - Doubly-robust estimators: predict either response $\bar{y}$ ***or*** propensity $\bar{p}$
  - Semi-parametric estimation

$$\underbrace{\text{Bias}}_{n^{-2\alpha}, \alpha > 1/4} + \underbrace{\text{Variance}}_{1/\sqrt{n}} \tag{537}$$

- Today
  - Uncertainty estimates
    - Assuming model coffect
    - Partial specification

20.1 Linear regression

Bias-variance decomposition

$$\mathbb{E}[(\mathbb{E}[Y] - \hat{Y})^2] = \underbrace{\mathbb{E}[\mathbb{E}[Y] - \hat{Y}]^2}_{\text{Bias}} + \underbrace{\text{Var}[\hat{Y}]}_{\text{variance}} \tag{538}$$

Let $(X_i, Y_i) \sim \tilde{p}$ and $\bar{X}_i \sim p^*$. Assume covariate shift, i.e.

$$\tilde{p}(y \mid x) = p^*(y \mid x) \tag{539}$$

and a linear model with homoscedastic noise

$$y = \langle \beta^*, x \rangle + z, \qquad z \sim N(0, \sigma^2) \tag{540}$$

$$\bar{y} = \langle \beta^*, \bar{x} \rangle + z \tag{541}$$

Question: Obtain $\hat{\beta}$ from $(X_i, Y_i)_{i=1}^n$. How well should I expect to do on p^*?
One way to measure it is just the squared error

$$\frac{1}{m} \sum_{i=1}^{m} (\bar{y}_i - \langle \hat{\beta}, \bar{x}_i \rangle)^2 \tag{542}$$

But this includes error introduced by z, so another way to measure it is excess risk

$$\frac{1}{m} \sum_{i=1}^{m} \langle \beta^* - \hat{\beta}, \bar{x}_i \rangle^2 \tag{543}$$

But we may not have labels $\bar{y}_i$ for p^*, so instead we measure using the expectations (over both p^* and $\hat{\beta}$) of the above.

We will see that it's often useful to condition the above expectations on the x_i and $\bar{x}_i$

$$\mathbb{E}\left[\frac{1}{m} \sum_{i=1}^{m} \langle \beta^* - \hat{\beta}, \bar{x}_i \rangle^2 \,\middle|\, x_1, \ldots, x_n, \bar{x}_1, \ldots, \bar{x}_m \right] \tag{544}$$

20.1.1 OLS estimator

Consider $\hat{\beta}$ the OLS estimator.

$$\hat{\beta} = \text{argmin}_\beta \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \beta, x_i \rangle)^2 \tag{545}$$

$$= \left(\underbrace{\frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top}_{\tilde{\Omega}} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i y_i \right) \tag{546}$$

$$= \tilde{\Omega}^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i (x_i^\top \beta^* + z_i) \right) \tag{547}$$

$$= \tilde{\Omega}^{-1} \left(\tilde{\Omega} \beta^* + \frac{1}{n} \sum_{i=1}^{n} x_i z_i \right) \tag{548}$$

$$= \beta^* + \tilde{\Omega}^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i z_i \right) \tag{549}$$

So we see immediately that this is ***unbiased***

$$\hat{\beta} - \beta^* = \tilde{\Omega}^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i z_i \right) \to \mathbb{E} = 0 \tag{550}$$

The variance of $\tilde{\Omega}^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i z_i \right)$ is

$$\text{Cov}(\tilde{\Omega}^{-1}(\frac{1}{n} \sum_{i=1}^{n} x_i z_i)) = \mathbb{E} \left[\tilde{\Omega}^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i z_i \right) \left(\frac{1}{n} \sum_{j=1}^{n} z_j x_j^\top \right) \tilde{\Omega}^{-1} \right] \tag{551}$$

$$= \tilde{\Omega}^{-1} \mathbb{E} \left[\frac{1}{n^2} \sum_{i=1}^{n} z_i^2 x_i x_i^\top \right] \tilde{\Omega}^{-1} \tag{552}$$

$$= \frac{\sigma^2}{n} \tilde{\Omega}^{-1} \tag{553}$$

If the errors are Gaussian, then squared error is excess loss plus σ, so consider excess loss.

$$\mathbb{E} \left[\frac{1}{m} \sum_{j=1}^{n} \langle \hat{\beta} - \beta^*, \bar{x}_j \rangle^2 \middle| x_{1:m}, \bar{x}_{1:m} \right] = \frac{1}{m} \sum_{j=1}^{m} (\hat{\beta} - \beta^*) \bar{x}_j \bar{x}_j^\top (\hat{\beta} - \beta^*) \tag{554}$$

$$= \mathbb{E} \left[(\hat{\beta} - \beta^*)^\top \left(\underbrace{\frac{1}{m} \sum_{j=1}^{m} \bar{x}_j \bar{x}_j^\top}_{\Omega^*} \right) (\hat{\beta} - \beta^*) \right] \tag{555}$$

$$= \mathbb{E} \left[\langle \Omega^*, \underbrace{(\hat{\beta} - \beta^*)(\hat{\beta} - \beta^*)^\top}_{\frac{\sigma^2}{n} \tilde{\Omega}^{-1}} \rangle \right] \tag{556}$$

$$= \frac{\sigma^2}{n} \langle \Omega^*, \tilde{\Omega}^{-1} \rangle \tag{557}$$

Interpretation: when $\tilde{\Omega} = \Omega^*$, we have

$$\langle \Omega^*, \tilde{\Omega}^{-1} \rangle = \text{Tr}(\Omega^* \tilde{\Omega}^{-1}) = \text{Tr}(I) = d \tag{558}$$

$$\therefore \text{Excess loss} = \sigma^2 \frac{d}{n} \tag{559}$$

- If $\tilde{\Omega}$ not invertible, then this result is inconclusive. Can show finite if $\|\beta^*\|_2 < \infty$
- Size of Ω^* vs $\tilde{\Omega}$: can double $\tilde{\Omega}$ and reduce error, this is because we are increasing scale of X but keeping σ constant
- $\tilde{\Omega}$ invertible, Ω^* not invertible. This is fine, because it just says that we got more dimensions at training $\tilde{p}$ than at test time p^*
- Easy to undo conditioning on $\bar{x}$ (just replace Ω^* with $\mathbb{E}_{p^*}[xx^\top]$), need concentration to undo $\tilde{\Omega}^{-1}$

We can undo the Gaussian assumption on z_i. We still have

$$\hat{\beta} = \beta^* + \tilde{\Omega}^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i z_i \right) \tag{560}$$

From which we see that we are ***unbiased*** if $\mathbb{E}[z_i \mid x_i] = 0$ for all i, and asymptotically unbiased if $\mathbb{E}[XZ] = 0$ and $\mathbb{E}[\bar{X}Z] = 0$.

Replace pseudo-inverse with Fisher information matrix in logistic regression

See Montanari and Bartlett for results characterizing bias and variance

This is avoided by determinantal design!

Considering the variance term (and no longer assuming unbiased), we find

$$\mathrm{Cov}(\hat{\beta}) = \mathrm{Cov}(\hat{\beta} - \beta^*) \tag{561}$$

$$= \mathrm{Cov}(\tilde{\Omega}^{-1} \sum_{i=1}^{n} x_i z_i) \tag{562}$$

$$= \frac{1}{n^2} \tilde{\Omega}^{-1} \sum_{i,j=1}^{n} \mathrm{Cov}(x_i z_i, x_j z_j) \tilde{\Omega}^{-1} \tag{563}$$

Since (x_i, z_j) is independent of (x_j, z_j), all off-diagonal terms are zero

This rules out experimental designs

$$= \frac{1}{n^2} \tilde{\Omega}^{-1} \sum_{i=1}^{n} \mathrm{Cov}(x_i z_j) \tilde{\Omega}^{-1} \tag{564}$$

$$= \frac{1}{n} \tilde{\Omega}^{-1} \left(\underbrace{\frac{1}{n} \sum_{i=1}^{n} x_i \, \mathrm{Cov}(z_i) x_i^\top}_{\tilde{M}} \right) \tilde{\Omega}^{-1} \tag{565}$$

Befroe we had $\frac{\sigma^2}{n} \tilde{\Omega}^{-1}$, whereas now we have the guarantee that $\mathrm{Var}(z_i \mid x_i) \le \sigma^2$ implies $\tilde{M} \le \sigma^2 \tilde{\Omega}$ and therefore

$$\mathrm{Var} \le \frac{\sigma^2}{n} \tilde{\Omega}^{-1} \tag{566}$$

If data is actually linear ($\mathbb{E}[Z_i \mid X_i] = 0$), then

$$\frac{1}{n} \langle \Omega^*, \tilde{\Omega}^{-1} \tilde{M} \tilde{\Omega}^{-1} \rangle \tag{567}$$

But if data is not linear (drop assumption $\mathbb{E}[z_i \mid x_i] = 0$, then we need to be careful about β^*. Define $\tilde{b}$ by

$$\hat{\beta} = \beta^* + \tilde{\Omega}^{-1} \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} x_i z_i \right)}_{\tilde{b}} \tag{568}$$

The predictive loss is no longer equal to the excess loss. Consider just the predictive loss

$$\mathbb{E}\left[\frac{1}{m} \sum_{j=1}^{m} \left(\bar{y}_j - \langle \hat{\beta}, \bar{x}_j \rangle \right)^2 \right] = \mathbb{E}\left[\frac{1}{m} \sum_{j=1}^{m} \left(\bar{z}_j + \langle \beta^* - \hat{\beta}, \bar{x}_j \rangle \right)^2 \right] \tag{569}$$

$$= \mathbb{E}\left[\frac{1}{m} \sum_{j=1}^{m} \bar{Z}_j^2 \right] + 2\mathbb{E}\left[\langle \underbrace{\frac{1}{m} \sum_{j=1}^{m} \bar{x}_j \bar{z}_j}_{= -b^* = \tilde{\Omega}^{-1} \tilde{b}}, \hat{\beta} - \beta^* \rangle \right] + \mathbb{E}\left[\frac{1}{m} \sum_{j=1}^{m} \langle \hat{\beta} - \beta^*, \bar{x}_j \rangle^2 \right] \tag{570}$$

where b^* is $\tilde{b}$ except defined using $\bar{x}_i$ and $\bar{z}_i$. The first term is easily computable. The last term has a bias and error term because $\hat{\beta}$ is biased. After some algebra, the error conditional on $\bar{x}_{1:m}$ and $x_{1:n}$ is

$$\langle \Omega^*, \tilde{\Omega}^{-1} \left(\frac{1}{n} \tilde{M} + \tilde{b} \tilde{b}^\top \right) \tilde{\Omega}^{-1} \rangle - 2 \langle b^*, \tilde{\Omega}^{-1} \tilde{b} \rangle + \frac{1}{m} \sum_{j=1}^{m} \mathbb{E}[\bar{z}_j^2 \mid \bar{x}_j] \tag{571}$$

If we define β^* to be the optimizer on $\tilde{p}$ the first two terms go away but we need to evaluate the third term. Alternatively, if we let β be the optimizer on p^*, then ??.

Bibliography

{donoho1988au
tomatic}

Donoho, D. L., R. C. Liu, et al.
1988. The" automatic" robustness of minimum distance functionals. *The Annals of Statistics*, 16(2):552–586.