STAT260: Robust Statistics Course Notes

Feynman Liang*
Department of Statistics, UC Berkeley

Last updated: September 20, 2019

Contents

*feynman@berkeley.edu

1 9/3/2019

$$\begin{tikzcd}
\begin{array}{c}\text{train}\\ \tilde{p}\end{array} \arrow[r, leftrightarrow, "D(\tilde{p},p^*)\leq\varepsilon"] \arrow[d] \arrow[dr] & \begin{array}{c}\text{test}\\ p^*\in\mathcal{G}\end{array} \arrow[dr] & \\
\begin{array}{c}X_1,\ldots,X_n\\ \text{samples}\end{array} \arrow[r] & \begin{array}{c}\hat{\theta}(\tilde{p})\\ \text{estimator}\end{array} \arrow[r] & \begin{array}{c}L(p^*,\hat{\theta})\\ \text{loss}\end{array}
\end{tikzcd}$$

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.

1.1 Minimum distance functional

Introduced in 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*)

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*)

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.

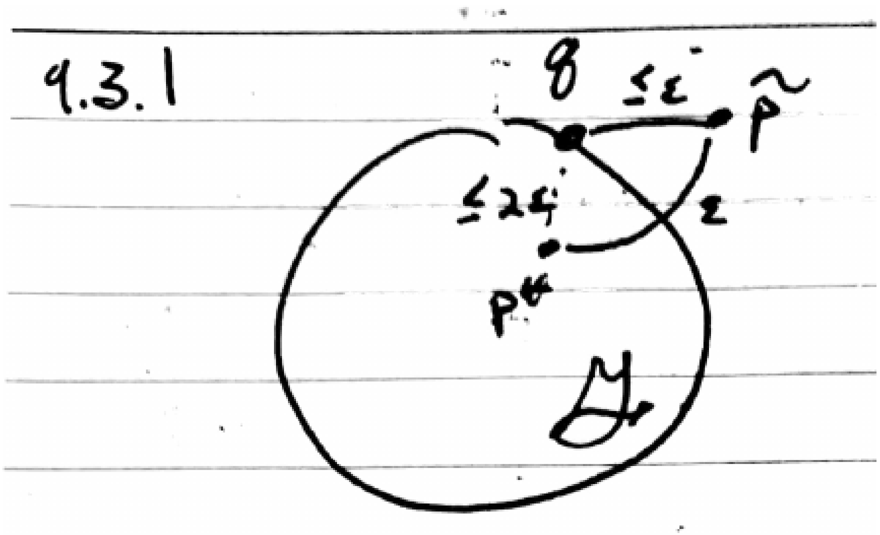


Figure 2: Given $D(p^*, \tilde{p}) \leq \varepsilon$, $p^* \in \mathcal{G}$, and q is the projection of $\tilde{p}$ onto $\mathcal{G}$ under D, we must have $D(\tilde{p}, q) \leq \varepsilon$ and by triangle inequality $D(p^*, q) \leq 2\varepsilon$

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 $D = \text{TV}$ 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

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

Therefore

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

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

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

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

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

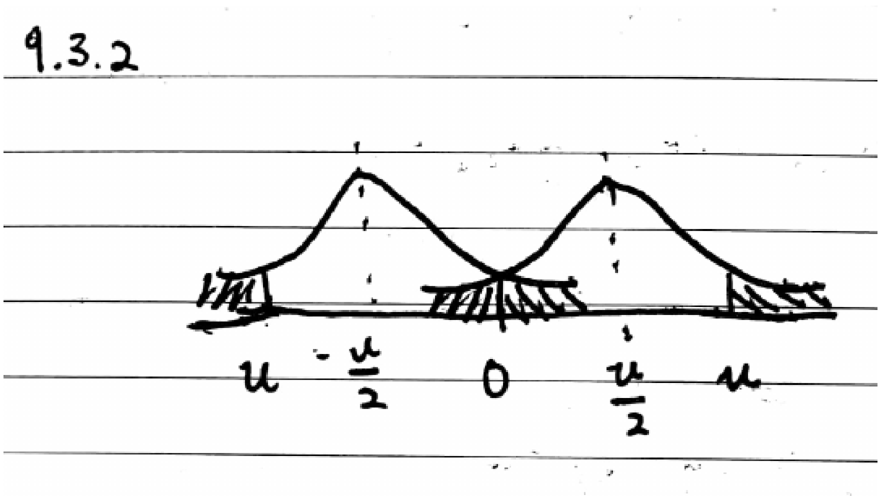


Figure 3: Both Gaussians exhibit identical $\pm\frac{u}{2}$ tails with opposite signs in the expression for TV, so the TV is equivalent to the area in $[-u/2, u/2]$ drawn out by the pointwise max between the two PDFs. By symmetry, this is just twice the area inside $[-u/2, u/2]$ which after cutting and 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. $\square$

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*)

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.

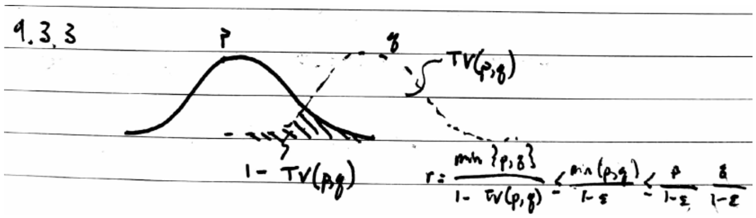


Figure 4: The midpoint distribution $r = \frac{\min(p,q)}{1-\mathrm{TV}(p,q)}$ can be reached from both p and q by deleting ϵ-mass and renormalizing.

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. $\square$

Corollary 5

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

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

$$\begin{aligned} \|\mathbb{E}_p[X - \mu(p) \mid E^c]\|_2 &= \sup_{\|v\|_2 \le 1} \langle \mathbb{E}_p[X - \mu(p) \mid E^c], v\rangle && (12)\\ &= \sup_{\|v\|_2 \le 1} \mathbb{E}_p[\langle X - \mu(p), v\rangle \mid E^c] && (13)\\ &\le \sup_{\|v\|_2 \le 1} \sqrt{\mathbb{E}_p[\langle X - \mu(p), v\rangle^2 \mid E^c]} && (14) \end{aligned}$$

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

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

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 $\|\mu(p) - \mu(q)\|_2 \le 4\sigma\sqrt{\varepsilon}$. $\square$

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:

Definition 7 (*Resilient distribution*)

A distribution is (ρ, ε)-resilient if

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

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.

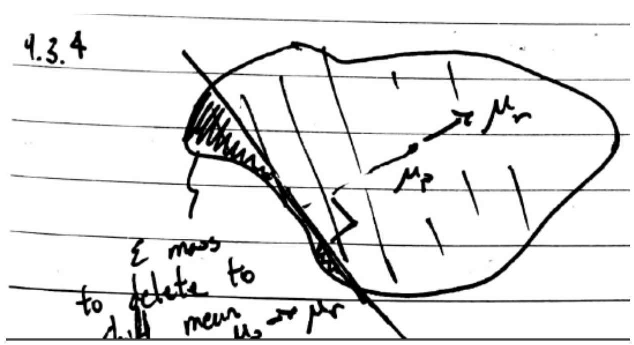


Figure 5: Deleting ε mass from a resilient distribution p shifts the mean by a controlled amount $\|\mu_p - \mu_r\|_2 \leq \rho$.

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

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

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

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

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*)

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*)

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

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*)

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*)

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

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

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

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

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

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

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*)

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

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

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

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

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*)

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*)

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*)

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*)

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*)

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

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

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

$$\begin{aligned} \Pr[X-\mu \geq t] &\leq \inf_{\lambda\geq 0} m_X(\lambda)e^{-\lambda t} && (54)\\ &\leq \inf_{\lambda\geq 0} \exp\left(\frac{1}{2}\lambda^2(\sigma')^2 - \lambda t\right) && (55)\\ &= \exp\left(-\frac{t^2}{2(\sigma')^2}\right) && (56)\end{aligned}$$

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

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

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*)

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

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)

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

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

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

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*)

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

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

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

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

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

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

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

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:

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

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

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 $\|\mathbb{E} X_i X_i^\top\| = O\left(\frac{\sigma^2}{n} \log \frac{1}{\delta}\right)$ 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

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

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 $v \in \mathcal{S}^{d-1}$

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

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

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

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

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

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

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

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

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 $\mathrm{Ball}_{9/8}(0)$ Therefore, by the (converse of the) pidgeonhole principle, $|N| \leq \frac{\mathrm{Vol}(\mathrm{Ball}_{9/8}(0))}{\mathrm{Vol}(\mathrm{Ball}_{1/8}(0))} = 9^d$.

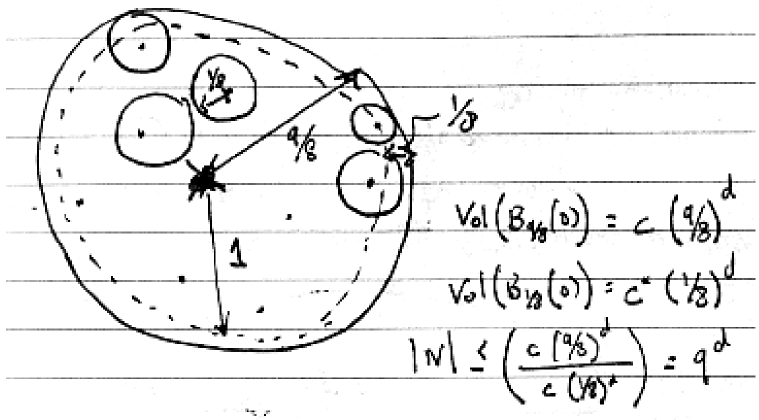


Figure 6: 1/8-radius balls centered at all packing points are disjoint, the union of all these balls is contained in $B_{9/8}(0)$, so the cardinality $|N| \leq \left(\frac{9/8}{1/8}\right)^d = 9^d$.

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.

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

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

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

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

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

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

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***

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

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*)

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.

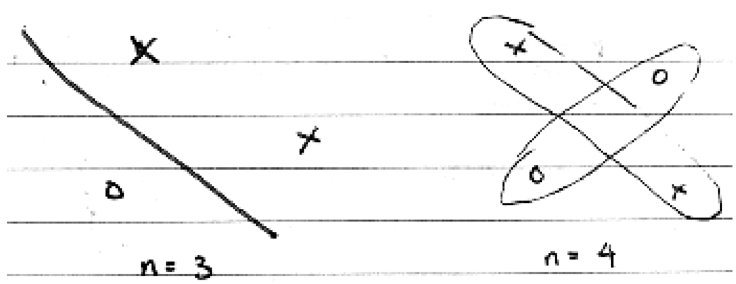


Figure 7: $n = 3$ can always be shattered by a line, but the crossings possible when $n = 4$ prevent this.

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*)

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 $O(n^{\mathbf{c}(\mathcal{H})})$ many terms of the form $\{f(X_i)\}_{i=1}^d$.

Theorem 42 (*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 $\varepsilon_i \sim \text{Rad}$

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

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*)

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 $S_+, S_- \subset \{x_1, \dots, x_{d+2}\}$ such that $S_+ \cap S_- = \emptyset$ but $\operatorname{conv}(S_+) \cap \operatorname{conv}(S_-) \neq \emptyset$. 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

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

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

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

or equivalently in matrix form

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

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)

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

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

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 $A = \{(p, q) \in \mathcal{G} : \widetilde{\mathrm{TV}}(p, q) \leq \varepsilon\}$ 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

Suppose $\tilde{\mathrm{TV}}$ *is a pseudometric such that* $\widetilde{\mathrm{TV}} \leq \mathrm{TV}$. *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*

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

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

Proof. By Proposition 2 (Modulus of continuity bound)

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

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

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

□

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*)

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

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

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*)

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

$\widetilde{\mathrm{TV}}(\tilde{p}, \tilde{p}_n)$ 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*)

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

$$\begin{aligned}
\underbrace{\mathbb{E}_p \langle v, x\rangle}_{=\langle v,\mu_p\rangle} &\le \mathbb{E}_{r_p}[\langle v, x\rangle] + \rho && \text{resilience of } p && (129)\\
&\le \mathbb{E}_{r_q}[\langle v, x\rangle] + \rho && \text{mean crossing} && (130)\\
&\le \underbrace{\mathbb{E}_q[\langle v, x\rangle]}_{=\langle v,\mu_q\rangle} + 2\rho && \text{resilience of } q && (131)
\end{aligned}$$

Hence

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

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

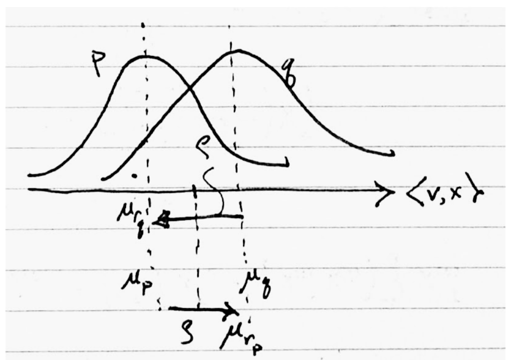


Figure 8: Resilience allows us to perform an ε-deletion to move from $\mu_p \to \mu_{r_p}$ and $\mu_q \to \mu_{r_q}$ and pick up a factor of $+2\rho$. Mean crossing allows us to relate μ_{r_p} and μ_{r_q}.

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 $\Pr_q[\langle v, x\rangle \ge \tau] - \Pr_p[\langle v, x\rangle \ge \tau] \le \widetilde{\mathrm{TV}}(p, q) \le \varepsilon$, 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):

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

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:

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

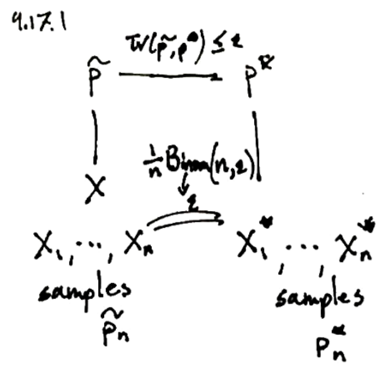
Under this analysis, we can achieve a tighter $\mathcal{O}\left(\varepsilon^{1-1/k} + \sqrt{d/n}\right)$ error.

Outline for today:

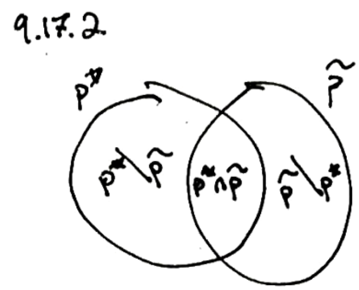
- True "empirical distribution"
- Expand the set idea
- Analyze concentration for bounded kth moments
  - symmetrization
  - truncated moments
  - ledoux-talagrand

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

Lemma 53 (*Tail bound for binomials*)

With probability $\ge 1 - \delta$

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

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

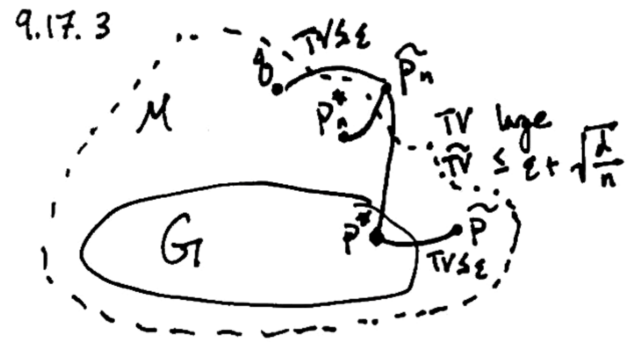


Figure 9: If we can expand $\mathcal{G} \subset \mathcal{M}$ so that $p_n^* \in \mathcal{M}$, then we can form $q = \min_{q\in\mathcal{M}} \mathrm{TV}(\tilde{p}_n, q)$ by projecting $\tilde{p}_n$ onto $\mathcal{M}$ and use the "true empirical distribution" p_n^* to connect q with p^*. This is made precise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

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

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

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

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

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

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

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 $\mathcal{M} = \mathcal{G}_{\text{TV}}(\rho, \varepsilon)$ with $\rho = O(\varepsilon^{1-1/k})$. We already have a modulus bound $\mathfrak{m}(\mathcal{M}, \varepsilon) \leq 2\rho = O(\varepsilon^{1-1/k})$ 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***:

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

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

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

This is equal to ψ_k for $x \leq x_0$ and linearly interpolates beyond $x \geq x_0$.

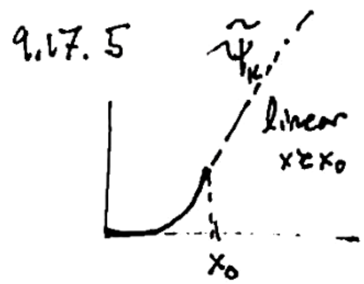


Figure 11: The Orlicz function $\tilde{\psi}_k$ used for truncating moments

$\tilde{\psi}_k$ 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

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

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

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

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

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

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

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

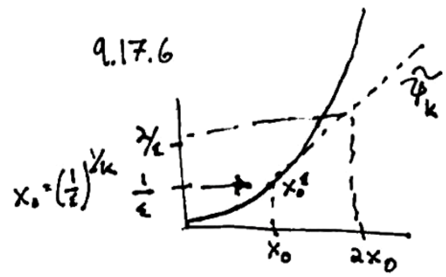


Figure 12: A proof by picture why $\psi^{-1}(2/\varepsilon) \leq 2x_0 = 2\varepsilon^{-1/k}$

5.4.2 Ledoux-Talagrand contraction

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?

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

□

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 $\mathcal{M} = \mathcal{G}$)
- $\mathcal{M} = \mathcal{G}_{\mathrm{TV}}(\rho, \varepsilon)$ = (ρ, ε)-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

$$\tilde{\psi}_k = \begin{cases} x^k & \text{if } x \le x_0 \\ kx_0^{k-1}(x - x_0) + x_0^k & \text{if } x > x_0 \end{cases} \tag{160}$$

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

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.

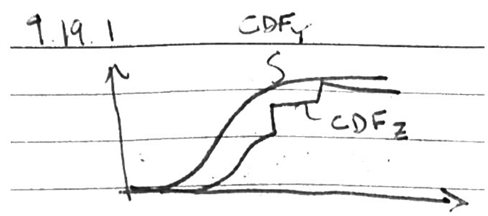


Figure 13: Intuition for $Z \succeq_1 Y$: going from Y to Z shifts cumulative distribution function (CDF) to the right

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

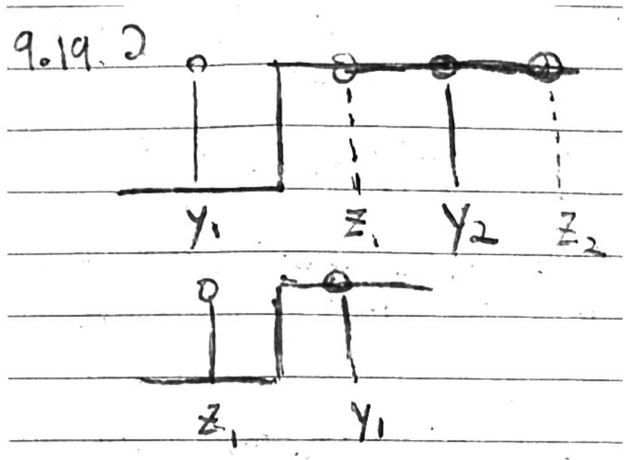
For $y_1 \le y_2$ and $z_1 \le z_2$, let

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

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

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

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

For the sufficiency,

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

□

Definition 64 (*Second order stochastic dominance*)

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

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

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*)

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

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

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

If

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

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

then $Z \succeq_2 Y$.

Proof. Necessity follows from considering ramp functions such as:

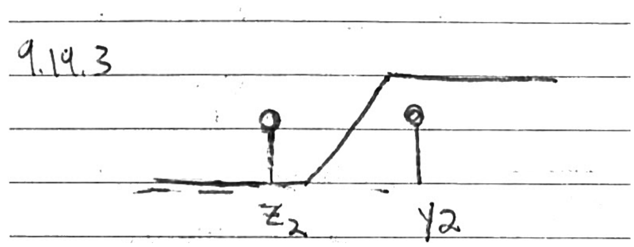


Figure 14: Example of a ramp function to show necessity that $y_2 \le z_2$. The ramp function should be drawn to be convex (i.e. continue increasing)

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

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 $(t_1, \dots, t_n)$ (think $t_i = \langle X_i - \mu, v\rangle$). Then

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

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 $|\phi(s) - \phi(t)| \le L|s-t|$.

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

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

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

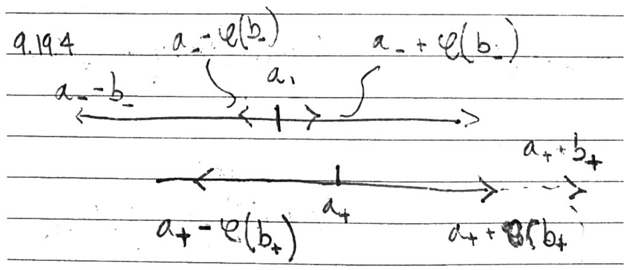
We will prove the theorem for $n = 2$ and then generalize by induction.

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

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

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:

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

By Lemma 65, we are done. □

Extending to $n > 2$.

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

$$\le \mathbb{E}_{\varepsilon_{1:n-1}}\left[\mathbb{E}_{\varepsilon_n}\left[g\left(\sup_{t\in T}\sum_{i=1}^{n-1}\varepsilon_i\phi(t_i) + \varepsilon_n t_n\right)\middle|\varepsilon_1,\dots,\varepsilon_{n-1}\right]\right] \tag{185}$$

$$= \mathbb{E}_{\varepsilon_{1:n}}\left[g\left(\sup_{t\in T}\sum_{i=1}^{n-1}\varepsilon_i\phi(t_i) + \varepsilon_n t_n\right)\right] \tag{186}$$

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

□

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*)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6.4 Zooming out

Goal: we want the following to concentrate

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

By symmetrization

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

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

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

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

Bibliography

Donoho, D. L., R. C. Liu, et al.
1988. The" automatic" robustness of minimum distance functionals. *The Annals of Statistics*, 16(2):552–586.