Lecture 5: Finite-Sample Bounds for Minimum Distance Functionals

— Hw 1: extended to next Tuesday

<u>Recap</u>

Concentration inequalities + union bound

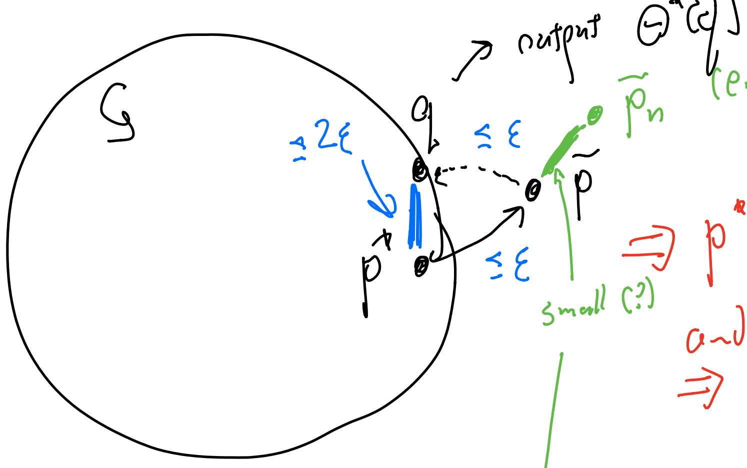
$\Rightarrow$ bound supremum

- Maximum eigenvalue of random matrix (ε-net)
- VC dimension bound (symmetrization)

$$\boxed{\begin{array}{l}\text{If } \mathcal{H} \text{ is a family of functions } f: \mathcal{X} \to \{0,1\}, \\ \text{and } vc(\mathcal{H}) = d, \text{ then w.p. } 1-\delta, \\ \displaystyle \sup_{f \in \mathcal{H}} |\underbrace{\nu_n(f)}_{\hat{\mathbb{P}}_n[f(x)=1]} - \underbrace{\nu(f)}_{\mathbb{P}[f(x)=1]}| \leq \mathcal{O}\left(\sqrt{\frac{d + \log(1/\delta)}{n}}\right) \\ \hat{\mathbb{P}}_n[f(x)=1] = \frac{1}{n}\sum_{i=1}^{n} f(x_i), \quad x_1, \dots, x_n \sim \mathbb{P}\end{array}}$$

<u>Today</u>: Finite-sample of MD functional

$D = TV$



(empirical distribution of n draws from $\tilde{p}$)

$\Rightarrow$ p^*, q close ($\leq 2\varepsilon$) and both lie in $\mathcal{G}$

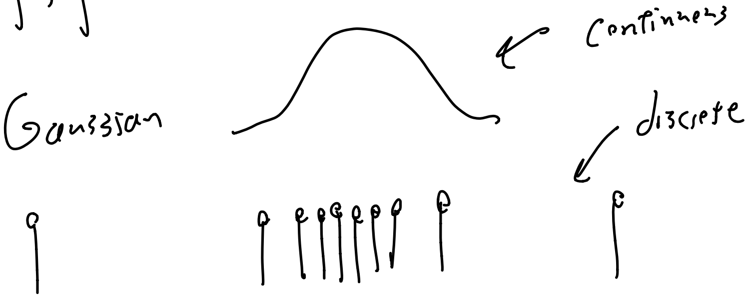
$\Rightarrow$ modulus of continuity $\mathfrak{m}(\mathcal{G}, 2\varepsilon)$ (resilient, e.g.)

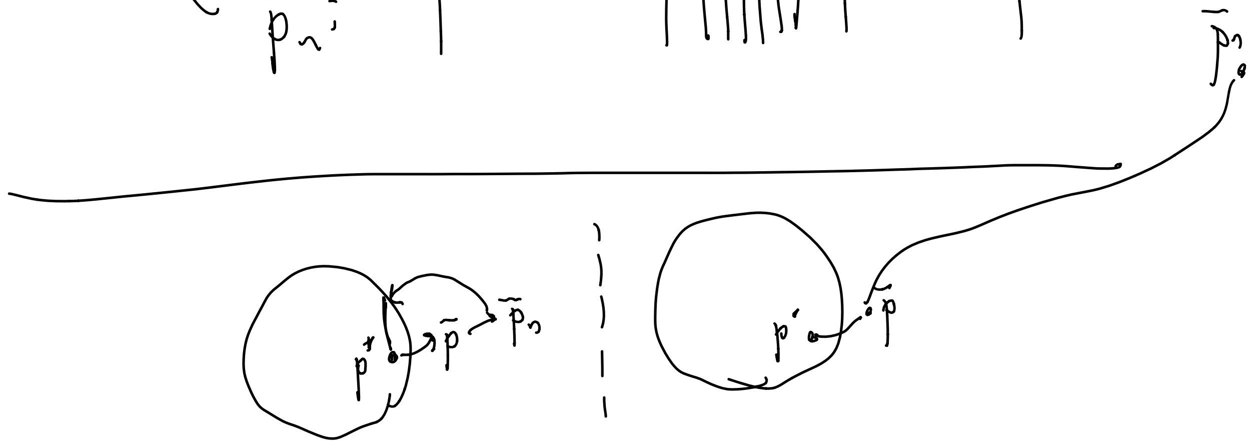
hopefully? going to 0 as $n \to \infty$

$TV(\tilde{p}, \tilde{p}_n) = 1$ even as $n \to \infty$.

$TV = 1$ ($\tilde{p}$ = Gaussian

$\hat{p}_n$:





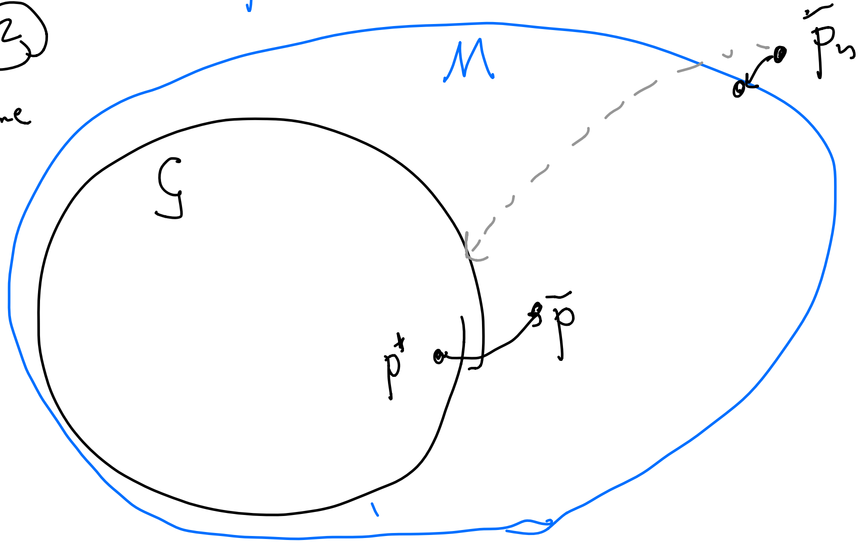
Issue. $\mathrm{TV}(\tilde{p}, \tilde{p}_n)$ large $\Rightarrow$ can't directly apply Δ-ineq.

Solutions. "relaxing the distance"

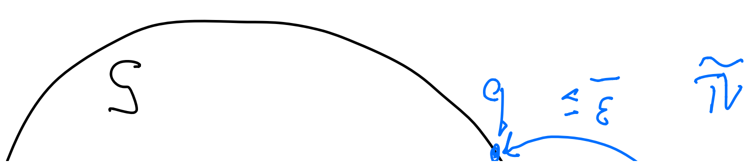
today → ① Replace TV w/ relaxed distance $\widetilde{\mathrm{TV}}$ where $\widetilde{\mathrm{TV}}(\tilde{p}, \tilde{p}_n)$ actually is small

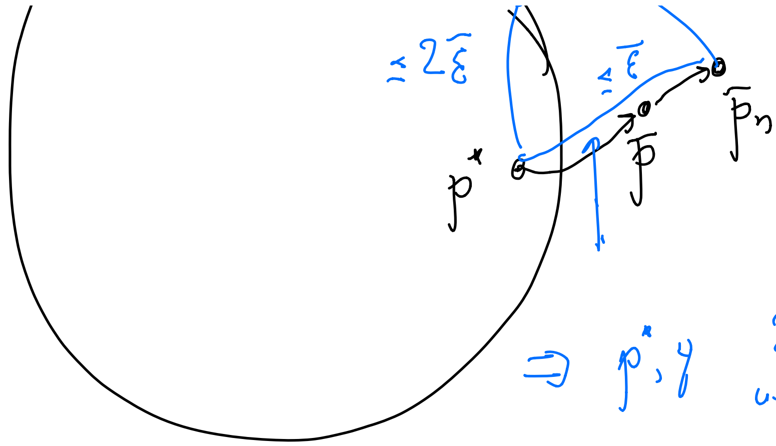
"expanding the set"

next time → ②



Lemma. Suppose $\widetilde{\mathrm{TV}}$ is a distance s.t. $\mathrm{TV} \geq \widetilde{\mathrm{TV}}$ and let $\tilde{\varepsilon} = \varepsilon + \mathrm{TV}(\tilde{p}, \tilde{p}_n)$. Then, the error of MD functional for $\widetilde{\mathrm{TV}}$ is at most $m(S, \widetilde{\mathrm{TV}}, 2\tilde{\varepsilon})$.





$\Rightarrow$ p^*, y $2\tilde{\varepsilon}$-close under $\widetilde{TV}$ distance

$\Rightarrow$ error bound by $\mathrm{Im}(\mathcal{G}, \widetilde{TV}, 2\varepsilon)$. $\blacksquare$

Conclusion need: (0) $TV \ge \widetilde{TV}$

(1) $TV(\bar{p}, \bar{p}_n) \to 0$

(2) need modulus to still be small

$\mathcal{H}$: family of fnc's $f: \mathcal{X} \to \mathbb{R}$

"variational representations"

$$\widetilde{TV}_{\mathcal{H}}(p,q) = \sup_{f \in \mathcal{H},\, \tau \in \mathbb{R}} \left| \underbrace{\mathbb{P}_{x \sim p}[f(x) \ge \tau]}_{\text{bounded}} - \mathbb{P}_{x \sim q}[f(x) \ge \tau] \right|$$

$\sup |v(f) - v_n(f)|$

(0) $TV \ge \widetilde{TV}_{\mathcal{H}}$

$$TV(p,q) = \sup_E |\mathbb{P}_p[E] - \mathbb{P}_q[E]|$$

(1) $\widetilde{TV}_{\mathcal{H}}(\bar{p}, \bar{p}_n) \to 0$ as $n \to \infty$ (at least intuitively)

↓ formalize via VC-dim argument

③ modulus → next

$$TV_1(p,q) = \sup_{1\text{-Lipschitz } f} \left|\mathbb{E}_p[f(x)] - \mathbb{E}_q[f(x)]\right|$$

Claim: For (ρ,ϵ)-resilient distributions, modulus still bounded by ρ.

$$\begin{tikzcd} p \arrow[rr, dash, "TV \le \epsilon"] \arrow[dr] & & q \arrow[dl] \\ & r & \end{tikzcd}$$

If $TV(p,q) \le \epsilon$, then

$\exists\, r \le \frac{p}{1-\epsilon}$, $r \le \frac{q}{1-\epsilon}$

$\Rightarrow \mu(r) \approx \mu(p)$

$\mu(r) \approx \mu(q)$

Δ-ineq. $\Rightarrow \mu(p) \approx \mu(q)$.

r_p $\qquad$ r_q

"mean of p crosses the mean of q"

For any $f \in \mathcal{H}$, $\mathbb{E}_p[f(x)]$

Lemma (Mean cross lemma) If $\widetilde{TV}_{\mathcal{H}}(p,q) \le \varepsilon$, and $f \in \mathcal{H}$, then $\exists\ r_p \le \frac{p}{1-\varepsilon}$, $r_q \le \frac{q}{1-\varepsilon}$, such that

$$\mathbb{E}_{r_q}[f(x)] \le \mathbb{E}_{r_p}[f(x)].$$

$\ge$

$r_p \le \frac{p}{1-\varepsilon} \Rightarrow$ For all events E, $r_p(E) \le \frac{p(E)}{1-\varepsilon}$.

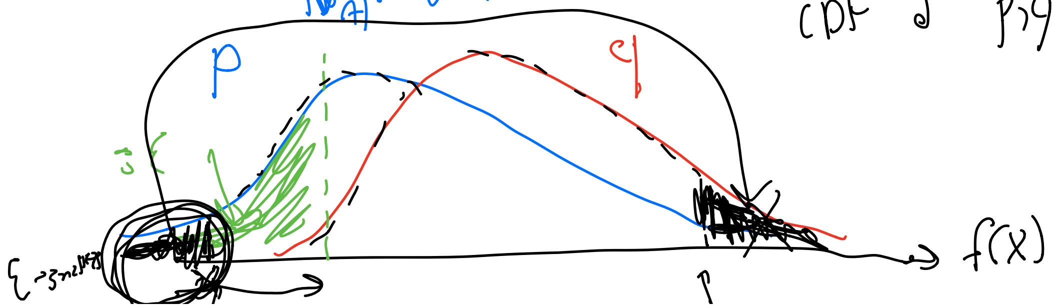
$\not\Rightarrow r_p(E) = \frac{p(E)}{1-\varepsilon} \Rightarrow r_p(\mathbb{R}^n) = \frac{1}{1-\varepsilon} \neq 1$

Pf of lemma, $\widetilde{TV}_{\mathcal{H}}(p,q) \le \varepsilon$

$$\Rightarrow \sup_{\tau \in \mathbb{R}} \left| \mathbb{P}_p[f(x) \ge \tau] - \mathbb{P}_q[f(x) \ge \tau] \right| \le \varepsilon,$$

$\widetilde{TV}_{\mathcal{H}}$: $\forall f \in \mathcal{H}$

CDF of p, q ε-close



r_p $\Leftarrow r_q$ r_q

$r_p \Rightarrow$

For all τ, $\mathbb{P}_{r_p}[f(x) \geq \tau] \geq \mathbb{P}_{r_q}[f(x) \geq \tau]$.

$\longrightarrow$ r_p stochastically dominates r_q

$\Downarrow$

$\longrightarrow$ $\mathbb{E}_{r_q}[f(x)] \leq \mathbb{E}_{r_p}[f(x)]$

$$\mathbb{E}[Z] = \int_0^\infty \mathbb{P}[Z \geq \tau]\, d\tau$$

$\Leftarrow$ stoch. dom.

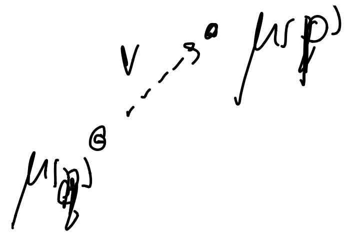
Back to bounding modulus.

Suppose p, q are (ρ, ε)-resilient.

$$\mathcal{H} = \{ x \mapsto \langle v, x \rangle \mid v \in \mathbb{R}^d \}$$

If $TV_{\mathcal{H}}(p, q) \leq \varepsilon$.

$$v = \frac{\mu(p) - \mu(q)}{\|\mu(p) - \mu(q)\|_2}$$



$$\begin{tikzcd}
p \arrow[r, dash, "TV \leq \varepsilon"] \arrow[d, dash, "\varepsilon\text{-deletions}"'] & q \arrow[d, dash, "\varepsilon\text{-deletions}"] \\
r_p \arrow[r, dash] & r_q
\end{tikzcd}$$

$$\|\mu(p) - \mu(q)\|_2$$

$$= \langle v, \mu(p) - \mu(q) \rangle$$

$$= \underline{\langle v, \mu(p) - \mu(r_p) \rangle} \longrightarrow \leq \|\mu(p) - \mu(r_p)\|_2 \leq \rho \text{ (by resilience)}$$

$$+ \langle v, \mu(r_p) - \mu(r_q) \rangle \leq \sigma \quad = \mathbb{E}_{r_p}[\langle v, x \rangle] - \mathbb{E}_{r_q}[\langle v, x \rangle] \leq \sigma$$

$$+ \langle v, \mu(r_q) - \mu(q) \rangle \longrightarrow \leq \rho$$

$\leq 2\rho. \quad \Rightarrow \quad \|\mu(p) - \mu(q)\|_2 \leq 2\rho$ even for $\widetilde{TV}$

$\Rightarrow$ modulus of continuity bound. ∎

Thm. If p^* is $(\rho, 2\tilde{\varepsilon})$-resilient

for $\tilde{\varepsilon} = \varepsilon + \widetilde{TV}(\tilde{p}, \tilde{p}_n)$, $\lesssim \sqrt{\frac{d}{n}}$ then

$\widetilde{TV}$ MD-functional recovers mean

w/ error $\leq 2\rho$, even in

finite samples.

$$\widetilde{TV}_{\mathcal{H}}(\tilde{p}, \tilde{p}_n) = \sup_{f \in \mathcal{H}, \tau \in \mathbb{R}} \left| \mathbb{P}_{\tilde{p}}[f(x) \geq \tau] - \mathbb{P}_{\tilde{p}_n}[f(x) \geq \tau] \right|$$

$$\mathcal{H}' = \{ \mathbb{I}[f(x) \geq \tau] \mid f \in \mathcal{H}, \tau \in \mathbb{R} \}$$

$$= \sup_{h \in \mathcal{H}'} \left| \mathbb{P}_{\tilde{p}}[h(x) = 1] - \mathbb{P}_{\tilde{p}_n}[h(x) = 1] \right|$$

$$= \sup_{h \in \mathcal{H}'} \underbrace{|\nu(h) - \nu_n(h)|}_{\text{VC-dim notation}} \leq \mathcal{O}\left(\sqrt{\frac{vc(\mathcal{H}') + \log(1/\delta)}{n}} \right)$$

w.p. $1-\delta$.

$$\widetilde{TV}(\tilde{p}, \tilde{p}_n) \lesssim \sqrt{\frac{d}{n}}, \text{ where } d = vc(\mathcal{H}').$$

Interpretation.

Resilience bound:

Bounded covariance: error $\mathcal{O}(\sigma\sqrt{\varepsilon})$ $\quad \varepsilon \mapsto \tilde{\varepsilon}$

Subgaussian: error $\mathcal{O}(\sigma \varepsilon \sqrt{\log(1/\varepsilon)})$.

$$\sigma\sqrt{\varepsilon+\sqrt{\frac{d}{n}}}$$

loose

$$\sqrt{\varepsilon}+\sqrt{\frac{d}{n}}$$

$\varepsilon \to 0$

$$\sigma\sqrt[4]{\frac{d}{n}}$$

(opt: $\sqrt{\frac{d}{n}}$)

gives us alternate way of estimating mean for heavy-tailed dist^ns

---

Connections b/w $\widetilde{TV}$ and Tukey median
