Lecs 3-4

Wednesday, August 30, 2017 10:13 AM

ORIE 6745 Lec 3 8/30/17

Last class: Defined counterfactuals using potential outcome

$Y(z)$: the outcome (for a generic unit) we would have observed if we applied z

In general, we don't observe all of these $\{Y(z) : z \in \mathcal{Z}\}$

We just observe a single factual outcome $Y = Y(Z)$ corresponding to some actually applied treatment Z.

Association $\neq$ Causation

Prediction $\neq$ Decision

$$Y \mid Z=z \overset{d}{\neq} Y(z)$$

- not everything held constant
- "Alternative explanations" may exist

Later on: we'll talk about creating statistical homogeneity w/ randomization & controlled trials

Also: we'll talk about "unconfoundedness" holding everything constant conditional on observables

For now: The easy case
Learning to Decide With
Fully observed Counterfactuals

Observe $Y_i(z)\ \forall z \in \mathcal{Z}\ \forall i = 1, \dots, n$

Want to find $\pi : \mathcal{X} \to \mathcal{Z}$

Policy risk $R(\pi) = \mathbb{E}[Y(\pi(X))]$

Empirical policy risk $\hat{R}_n(\pi) = \frac{1}{n} \sum_{i=1}^n Y_i(\pi(X_i))$

Last time:

Thm (Hoeffding Inequality)

If $V_i \in [a_i, b_i]$ indep $i = 1, \dots, n$

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_i V_i - \frac{1}{n} \sum_i \mathbb{E} V_i \right| \geq \varepsilon \right) \leq 2e^{\frac{-2n\varepsilon^2}{\frac{1}{n}\sum_i (b_i - a_i)^2}}$$

So: if $|Y(z)| \leq c$ then for a fixed π

$\hat{R}_n(\pi)$ tightly concentrated near $R(\pi)$

Note: Can be generalized for $Y(z)$ subgaussian (light tails) instead bounded.

Want actually:

If we choose $\hat{\pi}_n \in \operatorname{argmin}_{\pi \in \Pi} \hat{R}_n(\pi)$

then we really want

$$D(\Pi) = \sup_{\pi \in \Pi} \left| \hat{R}_n(\pi) - R(\pi) \right|$$

small.

One reason is empirical risk min (ERM)

$$\begin{aligned} R(\hat{\pi}_n) &\leq \hat{R}_n(\hat{\pi}_n) + D(\Pi) && \text{(by def)} \\ &\leq \hat{R}_n(\pi^*) + D(\Pi) && \text{(by opt)} \\ &\leq R(\pi^*) + 2D(\Pi) && \text{(by def)} \end{aligned}$$

for any $\pi^* \in \Pi$

in particular for $\pi^* \in \operatorname{argmin}_{\pi \in \Pi} R(\pi)$

best-in-class policy

This WHOLE 2 lecs will be about bounding $D(\Pi)$ with high prob in terms of the complexity of the class Π.

Key result

Theorem (McDiarmid's Inequality)

Let $f : \mathcal{V}^n \to \mathbb{R}$ be s.t.

$\forall i, v_1, \dots, v_n, v_i'$

$$|f(v_1, \dots, v_i, \dots, v_n) - f(v_1, \dots, v_i', \dots, v_n)| \leq c_i$$

(satisfies the "bounded difference condition")

Let $V_1, \dots, V_n$ indep RVs in $\mathcal{V}$. then

$$\mathbb{P}\left(f(V_1, \dots, V_n) - \mathbb{E} f(V_1, \dots, V_n) \geq \varepsilon \right) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2} \right)$$

Pf: Construct the Doob martingale

$$W_i = \mathbb{E}[f(V_{1:n}) \mid V_{1:i}]$$

Note: ① $W_0 = \mathbb{E}f(V_{1:n})$

② $W_n = f(V_{1:n})$

③ $\mathbb{E}[W_i - W_{i-1} \mid V_{1:i-1}] = 0$

④ $\mathbb{E}[e^{t(W_i - W_{i-1})} \mid V_{1:i-1}] \le e^{t^2 c_i^2/8}$

Therefore: for $t>0$

$$\mathbb{P}(W_n - W_0 \ge \varepsilon) = \mathbb{P}(e^{t(W_n - W_0)} \ge e^{t\varepsilon})$$

(by Markov) $\le e^{-t\varepsilon}\, \mathbb{E}[e^{t(W_n - W_0)}]$

$= e^{-t\varepsilon}\, \mathbb{E} \prod_{i=1}^n e^{t(W_i - W_{i-1})}$

(iter expec) $= e^{-t\varepsilon}\, \mathbb{E}[\mathbb{E}[\prod_{i=1}^n e^{t(W_i - W_{i-1})} \mid V_{1:n-1}]]$

$= e^{-t\varepsilon}\, \mathbb{E}[\prod_{i=1}^{n-1} e^{t(W_i - W_{i-1})}\, \mathbb{E}[e^{t(W_n - W_{n-1})} \mid V_{1:n-1}]]$

$\le e^{-t\varepsilon}\, \mathbb{E}[\prod_{i=1}^{n-1} e^{t(W_i - W_{i-1})}\, e^{t^2 c_n^2/8}]$

$\le \;\ldots$

$\le e^{-t\varepsilon} \prod_{i=1}^n e^{t^2 c_i^2/8}$

$= \exp(-t\varepsilon + t^2 \sum_i c_i^2/8)$

Let $t = 4\varepsilon / \sum_{i=1}^n c_i^2$ (opt by 1st order opt condns)

$$\Rightarrow -t\varepsilon + t^2 \sum_i c_i^2/8 = -2\varepsilon^2 / \sum_{i=1}^n c_i^2 . \quad \square$$

Suppose $|Y_i(z)| \le c$. Then

$$\Big| \sup_{\pi\in\Pi} \Big| \overbrace{\tfrac{1}{n}\textstyle\sum_{j\ne i} Y_j(\pi(X_j)) + \tfrac{1}{n} Y_i(\pi(X_i)) - \mathbb{E}Y(\pi(X))}^{Ⓐ} \Big|$$

$$- \sup_{\pi\in\Pi} \Big| \underbrace{\tfrac{1}{n}\textstyle\sum_{j\ne i} Y_j(\pi(X_j)) + \tfrac{1}{n} Y_i'(\pi(X_i')) - \mathbb{E}Y(\pi(X))}_{Ⓑ} \Big| \Big|$$

$\le \sup_{\pi\in\Pi} \big| |Ⓐ| - |Ⓑ| \big|$ $\quad (\sup_x f(x) - \sup_x g(x) \le \sup_x (f(x) - g(x)))$

$\le \sup_{\pi\in\Pi} |Ⓐ - Ⓑ|$ $\quad$ (triangle ineq)

$= \sup_{\pi\in\Pi} \frac{1}{n} |Y_i(\pi(X_i)) - Y_i'(\pi(X_i'))| \le 2c/n$

Therefore, letting $\varepsilon = c\sqrt{2\log(1/\delta)/n}$,

McDiarmid's will give:

$V_i = (X_i, \{Y_i(z) : z \in Z\})$

w.p. $\ge 1-\delta$,

$$D(\Pi) \le \mathbb{E}D(\Pi) + c\sqrt{2\log(1/\delta)/n}$$

So what's $\mathbb{E}D(\Pi)$?

$$\mathbb{E}D(\Pi) = \mathbb{E} \sup_{\pi\in\Pi} \Big| \tfrac{1}{n}\textstyle\sum_i Y_i(\pi(X_i)) - \mathbb{E}Y(\pi(X)) \Big|$$

$= \mathbb{E} \sup_{\pi\in\Pi} \Big| \frac{1}{n}\sum_i Y_i(\pi(X_i)) - \frac{1}{n}\sum_i \mathbb{E}[Y_i'(\pi(X_i') \mid X_{1:n}, Y_{1:n}] \Big]$

$\le \mathbb{E}\Big[\sup_{\pi\in\Pi} \Big| \frac{1}{n}\sum_i Y_i(\pi(X_i)) - \frac{1}{n}\sum_i Y_i'(\pi(X_i')) \Big|\Big]$

$= \mathbb{E} \sup_{\pi\in\Pi} \Big| \frac{1}{n}\sum_i (Y_i(\pi(X_i) - Y_i'(\pi(X_i'))) \Big|$

Let $\sigma_i = \pm 1$ equip indep (Rademacher random variables)

$= \mathbb{E} \sup_{\pi\in\Pi} \Big| \frac{1}{n}\sum_i \sigma_i (Y_i(\pi(X_i)) - Y_i'(\pi(X_i'))) \Big|$

$\le 2\,\mathbb{E} \sup_{\pi\in\Pi} \Big| \frac{1}{n}\sum_i \sigma_i Y_i(\pi(X_i)) \Big|$

Def Given a set of point $\mathcal{A} \subseteq \mathbb{R}^n$ the Rademacher complexity of $\mathcal{A}$ is $\mathfrak{R}_n(\mathcal{A}) = \underbrace{\frac{1}{2^n} \sum_{\sigma \in \{-1,1\}^n}}_{\mathbb{E}_\sigma} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^n \sigma_i a_i$

So: $\mathbb{E} D(\Pi) \leq 2 \mathbb{E}[\mathfrak{R}_n(\mathcal{Z}_n(\Pi))]$

When $\mathcal{Z}_n(\Pi) = \{(Y_1(\pi(X_1)), \ldots, Y_n(\pi(X_n)) : \pi \in \Pi\}$

Again, if $|Y_i(z)| \leq c$ then

replace $X_i, Y_i(\cdot)$ by $X_i', Y_i'(\cdot)$

$$\left| \mathfrak{R}_n(\mathcal{Z}_n(\Pi)) - \mathfrak{R}_n(\mathcal{Z}_n'(\Pi)) \right|$$

$$= \Bigg| \mathbb{E}_\sigma \Bigg(\sup_{\pi \in \Pi} \Big(\overbrace{\frac{1}{n} \sum_{j \neq i} \sigma_j Y_j(\pi(X_j)) + \frac{1}{n} \sigma_i Y_i(\pi(X_i))}^{Ⓐ} \Big)$$

$$- \sup_{\pi \in \Pi} \Big(\underbrace{\frac{1}{n} \sum_{j \neq i} \sigma_j Y_j(\pi(X_j)) + \frac{1}{n} \sigma_i Y_i'(\pi(X_i'))}_{Ⓑ} \Big) \Bigg) \Bigg|$$

$$\leq \mathbb{E}_\sigma \left| \sup_{\pi \in \Pi} Ⓐ - \sup_{\pi \in \Pi} Ⓑ \right| \leq \mathbb{E}_\sigma \sup_{\pi \in \Pi} |Ⓐ - Ⓑ|$$

$$= \mathbb{E}_\sigma \sup_{\pi \in \Pi} |\sigma_i| \cdot \left| \frac{1}{n} \left(Y_i(\pi(X_i)) - Y_i'(\pi(X_i')) \right) \right|$$

$$\leq \frac{2c}{n}$$

So by McDiarmid's:

wp $\geq 1 - \delta$

$$\mathbb{E} \mathfrak{R}_n(\mathcal{Z}_n(\Pi)) \leq \mathfrak{R}_n(\mathcal{Z}_n(\Pi)) + c \sqrt{\frac{2 \log(4/\delta)}{n}}$$

So far:

$$D(\Pi) \leq \mathfrak{R}_n(\mathcal{Z}_n(\Pi)) + 2c \sqrt{\frac{2 \log(4/\delta)}{n}}$$

Next: interpret $\mathfrak{R}_n(\mathcal{Z}_n(\Pi))$ as the complexity of Π

(An extension of Ledoux-Talagrand Lipschitz comparison lemma)

Theorem Suppose $\mathcal{Z} \subseteq \mathbb{R}^d$

Fix $X_i, Y_i(\cdot)$ $i = 1, \ldots, n$

Suppose each $Y_i(\cdot)$ is L-Lipschitz

$Y_i(z) - Y_i(z') \leq L \|z - z'\|_\infty$

for all i

Let $\mathcal{A}_n^{(k)}(\Pi) = \{((\pi(X_1))_k, \ldots, (\pi(X_n))_k) : \pi \in \Pi\}$

Then
$$\mathfrak{R}_n(\mathcal{Z}_n(\Pi)) \leq L \sum_{k=1}^d \mathfrak{R}_n(\mathcal{A}_n^{(k)}(\Pi))$$

Pf: WLOG $L = 1$ (divide LHS by L)

Will show (stronger than thm)

$$\mathbb{E}_\sigma \sup_{\pi \in \Pi} \frac{1}{n} \sum_i \sigma_i Y_i \pi(X_i) \leq \mathbb{E}_{\sigma \in \{-1,1\}^{n \times d}} \sup_{\pi \in \Pi} \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^d \sigma_{ik} (\pi(X_i))_k$$

Suppose we have that for any $S \subseteq \mathbb{R} \times \mathcal{Z}$ and a 1-Lipschitz $\varphi : \mathcal{Z} \to \mathbb{R}$

$$⊛ \quad \mathbb{E}_\sigma \sup_{t, z \in S} (t + \sigma \varphi(z)) \leq \mathbb{E}_\sigma \sup_{t, z \in S} (t + \sum_k \sigma_k z_k)$$

Then:

$$\mathbb{E}_\sigma \sup_{\pi\in\Pi} \frac{1}{n}\sum_i \sigma_i \psi_i(\pi(x_i))$$

$$= \mathbb{E}_{\sigma_1,\dots,\sigma_{n-1}}\left[\mathbb{E}_{\sigma_n}\left[\sup_{\substack{\pi\in\Pi\\ t=\frac{1}{n}\sum_{i=1}^{n-1}\sigma_i\psi_i(\pi(x_i))\\ z=\pi(x_n)}} (t+\sigma_n\psi_n(z)) \middle| \sigma_{1:n-1}\right]\right]$$

$$\le \mathbb{E}_{\sigma_1,\dots,\sigma_{n-1}} \mathbb{E}_{\sigma_n\in\{-1,1\}^d} \sup_{\substack{\pi\in\Pi\\ t=\dots\\ z=\dots}} t+\sum_k \sigma_{nk} z_k$$

$$\le \dots \le \mathbb{E}_\sigma \sup_{\pi\in\Pi} \frac{1}{n}\sum_i \sum_k \sigma_{ik} (\pi(x_i))_k$$

Let's show $\otimes$

Fix any $(t^{(+1)}, z^{(+1)}), (t^{(-1)}, z^{(-1)}) \in S$

Let $s^* = \pm 1$ $k^* = 1,\dots,d$ be

$$\|z^{(+1)} - z^{(-1)}\|_\infty = s^*(z^{(+1)}_{k^*} \; z^{(-1)}_{k^*})$$

$$\text{RHS } \otimes \ge \mathbb{E}_{\sigma_{k^*}}\left[\sup_{t,z\in S} \mathbb{E}_{\sigma_k : k\ne k^*}\left[t+\sum_k \sigma_k z_k \mid \sigma_{k^*}\right]\right)$$

$$\ge \frac{1}{2}\mathbb{E}_{\sigma_k : k\ne k^*}\left[t^{(+s^*)} + z^{(+s^*)}_{k^*} + \sum_{k\ne k^*} \sigma_k z^{(+s^*)}_k\right]$$

$$+ \frac{1}{2}\mathbb{E}_{\sigma_k : k\ne k^*}\left[t^{(-s^*)} - z^{(-s^*)}_{k^*} + \sum_{k\ne k^*} \sigma_k z^{(-s^*)}_k\right]$$

$$= \frac{1}{2}\left(t^{(+1)} + t^{(-1)} + \|z^{(+1)} - z^{(-1)}\|_\infty\right)$$

$$\ge \frac{1}{2}\left(t^{(+1)} + \varphi(z^{(+1)})\right) + \frac{1}{2}\left(t^{(-1)} - \varphi(z^{(-1)})\right)$$

Taking suprema over $t^{(\pm 1)}, z^{(\pm 1)}$ we get the LHS of $\otimes$. $\square$

All together:

wp $\ge 1-\delta$

$$D(\Pi) \le L \sum_{k=1}^d \mathfrak{R}_n(A_n^{(k)}(\Pi)) + 2C\sqrt{\frac{2\log(2/\delta)}{n}}$$

Ex: Supervised regression w/ bounded outcomes

$\psi_i(z) = (L_i - z)^2$ is $2C$-Lipschitz

$z\in\mathbb{R}$ univariate so

$$D(\Pi) \le \underbrace{\mathfrak{R}_n(A_n(\Pi))}_{\text{what's this?}} + 2C\sqrt{\frac{2\log(2/\delta)}{n}}$$

One kind bound:

If $\|X\|_2 \le M$ $\quad \Pi = \{x \mapsto \beta^T x : \|\beta\|_2 \le R\}$

then

$$\mathfrak{R}_n(A_n(\Pi)) = \mathbb{E}_\sigma \sup_{\|\beta\|_2 \le R} \frac{1}{n} \sum_{i=1}^n \sigma_i \beta^T X_i$$

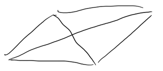
$$= \mathbb{E}_\sigma \sup_{\|\beta\|_2 \le R} \beta^T \left(\frac{1}{n} \sum_i \sigma_i X_i \right)$$

$\sup_{\|v\|_2 \le 1} v^T u = \|u\|_2$

$$= \mathbb{E}_\sigma \left[R \cdot \left\| \frac{1}{n} \sum_i \sigma_i X_i \right\|_2 \right]$$

Paralledogram Law

$\frac{1}{2}\|x_1 + x_2\|_2^2$
$+ \frac{1}{2}\|x_1 - x_2\|_2^2 = \|x_1\|_2^2 + \|x_2\|_2^2 + \cancel{\langle x_1, x_2 \rangle}$
$= \|x_1\|_2^2 + \|x_2\|_2^2 \qquad - \cancel{\langle x_1, x_2}$



$$\to \le R \sqrt{\mathbb{E}_\sigma \left\| \frac{1}{n} \sum_i \sigma_i X_i \right\|_2^2}$$

$$\to = \frac{R}{n} \sqrt{\sum_{i=1}^n \|X_i\|_2^2} \le \frac{R\sqrt{nM^2}}{n}$$

$$= \frac{RM}{\sqrt{n}}$$

whp: $\forall \pi \in \Pi$

$$R(\pi) \le \hat{R}_n(\pi) + 2C\, RM/\sqrt{n} + 2C\sqrt{2\log(2/\delta)/n}$$

→ Structural risk minimization

Choose $\hat{\pi}_n$ to minimize

$$\hat{R}_n(\pi_\beta) + \lambda \|\beta\|_2$$

Another kind of bound: VC-dimension

Suppose $\mathcal{Z} = \{\pm 1\}$

e.g. - predict ±ive (classification)
- give or don't give treatment (actual decision)

Then $\psi(\cdot)$ is 2C-Lipschitz

What about $\mathfrak{R}_n(A_n(\Pi))$?

Lemma: (Massart's lemma)

Let $\mathcal{A} \subseteq [-M, M]^n$

Then: $\mathcal{R}_n(\mathcal{A}) \le M \cdot \sqrt{\frac{2}{n} \log|\mathcal{A}|}$

Pf: Let $\lambda > 0$ Then

(jensen's) $$e^{\lambda \mathcal{R}_n(\mathcal{A})} \le \mathbb{E}_\sigma\left[\exp\left(\lambda \sup_{a\in\mathcal{A}} \frac{1}{n} \textstyle\sum_i \sigma_i a_i\right)\right]$$

$$= \mathbb{E}_\sigma\left[\sup_{a\in\mathcal{A}} \exp\left(\lambda \frac{1}{n} \textstyle\sum_i \sigma_i a_i\right)\right]$$

$$\le \mathbb{E}_\sigma\left[\sum_{a\in\mathcal{A}} \exp\left(\frac{\lambda}{n} \textstyle\sum_i \sigma_i a_i\right)\right]$$

$$= \textstyle\sum_{a\in\mathcal{A}} \prod_{i=1}^n \mathbb{E}_{\sigma_i} \exp\left(\frac{\lambda}{n} \sigma_i a_i\right)$$

$\cosh(x) \le e^{x^2/2}$ $$= \textstyle\sum_{a\in\mathcal{A}} \prod_{i=1}^n \left(\frac{1}{2}\exp\left(\frac{\lambda}{n} a_i\right) + \frac{1}{2}\exp\left(\frac{-\lambda}{n} a_i\right)\right)$$

$$\le \textstyle\sum_{a\in\mathcal{A}} \prod_{i=1}^n e^{\lambda^2 a_i^2 / 2n^2}$$

$$\le \textstyle\sum_{a\in\mathcal{A}} \prod_{i=1}^n e^{\lambda^2 M^2 / 2n^2}$$

$$= \textstyle\sum_{a\in\mathcal{A}} e^{\lambda^2 M^2 / 2n} =$$

$$= |\mathcal{A}|\, e^{\lambda^2 M^2 / 2n}$$

$$\Rightarrow \mathcal{R}_n(\mathcal{A}) \le \frac{\log|\mathcal{A}|}{\lambda} + \frac{\lambda M^2}{2n}$$

Set $\lambda = \sqrt{2n \log|\mathcal{A}|}/M$

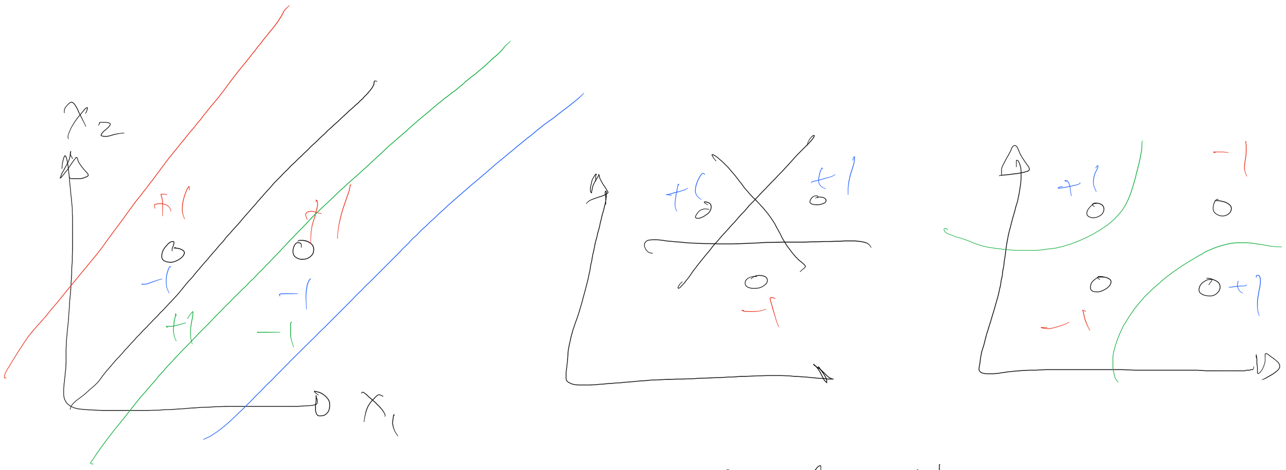
$$\Rightarrow \mathcal{R}_n(\mathcal{A}) \le \sqrt{2\log|\mathcal{A}|/n} \cdot M \,. \quad \square$$

Def:

$|\mathcal{A}_n(\Pi)|$ is called the growth function

Def: The VC-dimension of $\Pi \subseteq [\mathcal{X} \to \{-1,+1\}]$, $VC(\Pi)$, is the largest number n s.t. $\forall x_1, \dots, x_n \in \mathcal{X}$ we have

$$\underbrace{|A_n(\Pi)| = 2^n}_{\text{called "}\Pi\text{ shatters } x_1, \dots, x_n\text{"}}$$



Prop $\Pi = \{x \longmapsto \overset{\text{sign}(0)=+1}{\text{sign}}(\beta^T x) : \beta \in \mathbb{R}^p\}$, $\mathcal{X} \in \mathbb{R}^p$ has VC dimension p.

Pf: To show $VC(\Pi) \geq p$: find p point we can shatter

Consider $x_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \dots, x_p = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$

Let $z \in \{-1, 1\}^p$ be given. Set $\beta_i = z_i$ we get

$$(\Pi(x_1), \dots, \Pi(x_p)) = (z_1, \dots, z_p)$$

To show $VC(\Pi) \leq p$: show we cannot shatter $p+1$ points

Let $x_1, \dots, x_{p+1}$ be given

They cannot be linearly indep so

$\exists \lambda \neq 0 \ \sum_i \lambda_i x_i = 0$

$\Rightarrow \exists \lambda, j : x_j = \sum_{i \neq j} \lambda_i x_i$

Let $\beta \in \mathbb{R}^p$ and suppose

$$\text{sign}(\beta^T x_i) = \text{sign}(a_i) \quad \forall i \neq j$$

$$\Rightarrow \text{sign}(\beta^T x_j) = \text{sign}\left(\sum_{i \neq j} a_i \beta^T x_i\right) = +1 \quad . \quad \square$$

Why VC dimension?

Lemma (Sauer) If $VC(\Pi) \leq v$

then $|A_n(\Pi)| \leq \sum_{i=0}^{v} \binom{n}{i} \leq \left(\frac{en}{v}\right)^v$

Rademacher + Lipschitz Comparison
+ 2 × McDiarmid
+ Massart 3
+ Sauer's ;

For binary decision policies
& bounded outcomes

$$D(\Pi) = \sup_{\pi \in \Pi} \left| \hat{R}_n(\pi) - R(\pi) \right|$$

$$\leq 2C \sqrt{2 \log(2/\delta)/n} + 4C \sqrt{\frac{2 \log n}{n} VC(\Pi)}$$

Note Can be extended to get:

- get rid of $\log n$
  (Chaining argument; see Pollard reading)
- replace bounded by subgaussian
- deal w/ nonbinary decision

Taking a step back

Usually counterfactuals are
not observed — only factuals $Y = Y(Z)$

$Y \mid Z = z$ vs. $Y(z)$

What to do?

Causal inference!

Next time: Controlled experiments
(A/B testing)

Create statistical ~~homogenei~~
homogeneity through randomization

Randomization $\Rightarrow$ no alternative explanations
for an observed association
other than a causal effect