

CS 229br Lecture 4: Robustness Boaz Barak



Yamin Bansal
Official TF



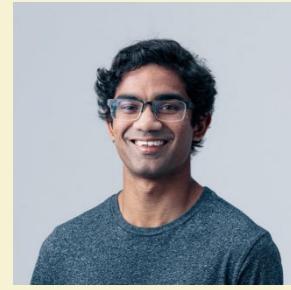
Javin Pombra
Official TF



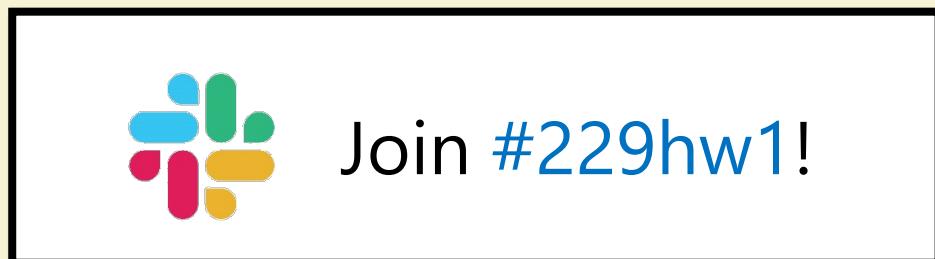
Dimitris Kalimeris
Unofficial TF



Gal Kaplun
Unofficial TF



Preetum Nakkiran
Unofficial TF



Plan

1. Math review
2. Digression – multiplicative weights / follow the regularized leader /...
3. Train-time robustness – robust mean estimation, data poisoning attacks
4. Test-time robustness – distribution shift and adversarial perturbations

KL refresher

$$\Delta_{KL}(p \parallel q) = \mathbb{E}_{x \sim p} \left[\log \frac{p(x)}{q(x)} \right] = \underbrace{\mathbb{E}_{x \sim p} [\log p(x)]}_{-H(p)} - \underbrace{\mathbb{E}_{x \sim p} [\log q(x)]}_{H(p, q)}$$

- $H(p)$

Negative entropy

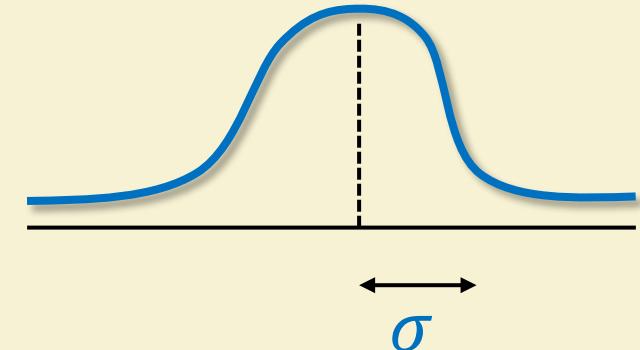
$H(p, q)$

Cross entropy

$$\Delta_{KL}(p \parallel q) \geq 0 \quad \Rightarrow \quad \forall p, q \quad H(p, q) \geq H(p)$$

Concentration

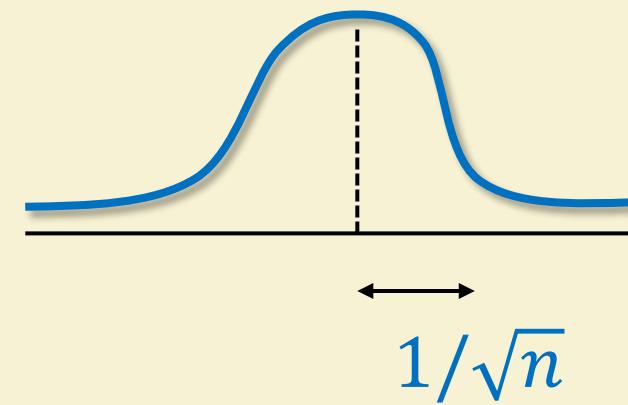
$X \sim N(\mu, \sigma^2)$ – normal mean μ std σ



$$\Pr[|X - \mu| \geq t\sigma] \approx \exp(-t^2)$$

If Y_1, \dots, Y_n are i.i.d bounded with expectation μ

$$\frac{1}{n} \sum Y_i \approx N(\mu, 1/n) = \frac{1}{\sqrt{n}} N(\mu, 1)$$



“Chernoff” / “Hoeffding” /
“Bernstein” / ...

$$\Pr[|\sum Y_i - \mu \cdot n| \geq \epsilon n] \approx \exp(-\epsilon^2 n)$$

* Dropping constants (even in exponents), assuming ϵ sufficiently small constant

Matrices

write as $M \geq 0$

A symmetric M is **psd** if all e-vals are non-negative.

Equivalently $v^T M v \geq 0$ for all vectors v

Def: $A \leq B$ if $v^T A v \leq v^T B v$ for all vectors v

say $A \in [a, b]I$ if $aI \leq A \leq bI$ Equivalently $\lambda_i(A) \in [a, b]$ for all i

Def: spectral norm of a matrix A is $\|A\| = \max_{\|v\|=1} \|Av\| = \lambda_{\max}(A)$

Def: Frobenius norm of A is $\|A\|_F = \sqrt{\sum A_{i,j}^2} = \|\text{vec}(A)\|$

Matrix/vector valued r.v.

Vector valued normals:

If $\mu \in \mathbb{R}^d$, $V \in \mathbb{R}^{d \times d}$ psd, $\mathbf{x} \sim N(\mu, V)$ normal over \mathbb{R}^d with

$$\mathbb{E}[x_i] = \mu_i , \quad \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)] = V_{i,j}$$

Standard vector-valued normal: $\mathbf{x} \sim N(0^d, I_d)$ (or $\mathbf{x} \sim N(0, I)$)

$$\mathbb{E}\mathbf{x} = \vec{0} , \quad \mathbb{E}[\mathbf{x}\mathbf{x}^\top] = I \quad \begin{pmatrix} \mathbb{E}x_i^2 = 1 \\ \mathbb{E}x_i x_j = 0 \text{ for } i \neq j \end{pmatrix}$$

Matrix concentration

Recall: If Y_1, \dots, Y_n are i.i.d over \mathbb{R} bounded with expectation μ

$$\Pr[|\sum Y_i - \mu \cdot n| \geq \epsilon n] \approx \exp(-\epsilon^2 n)$$

Matrix Bernstein inequality:

If Y_1, \dots, Y_n i.i.d symmetric matrices in $\mathbb{R}^{d \times d}$ with $\mathbb{E}Y_i = \mu$, $\|Y_i\| \leq O(1)$

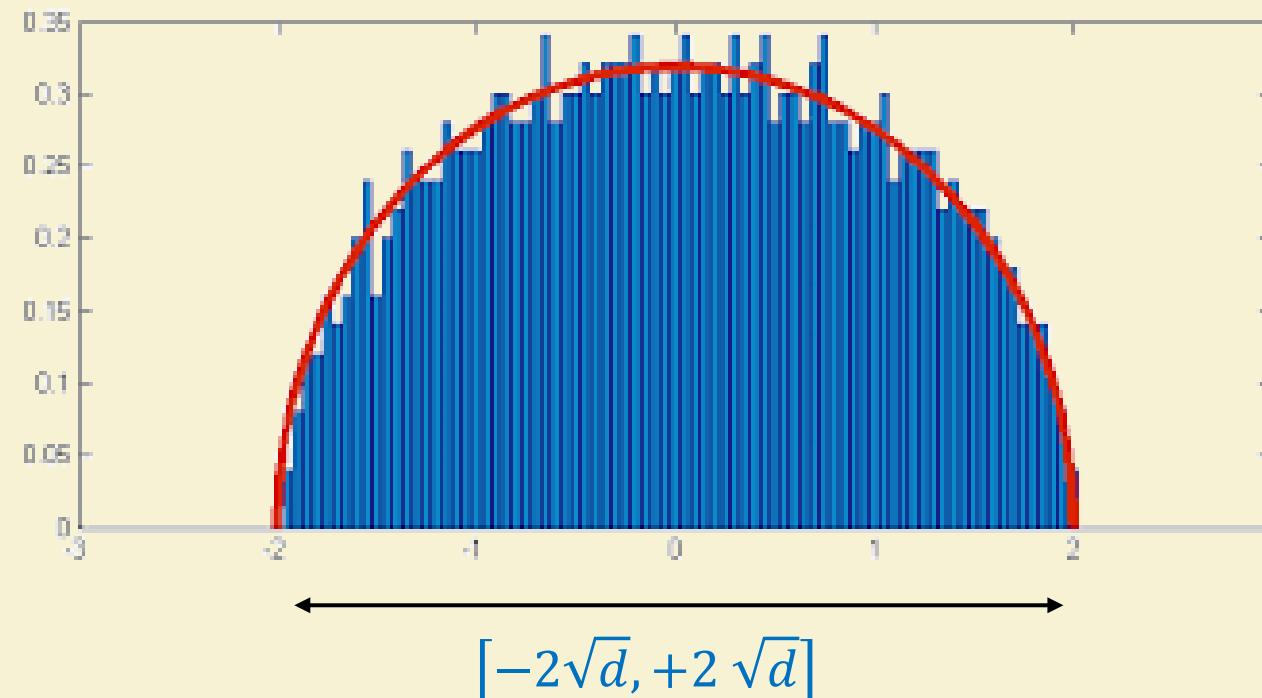
$$\Pr[\|\sum Y_i - \mu \cdot n\| \geq \epsilon n] \approx d \cdot \exp(-\epsilon^2 n)$$

$$\mathbb{E}\|\sum Y_i - \mu\| \leq O(\sqrt{n \log d})$$

Random matrices

A random $d \times d$ symmetric matrix with $A_{i,j} \sim N(0,1)$

Spectrum (eigenvalues) distributed according to Wigner Semi-Circle law

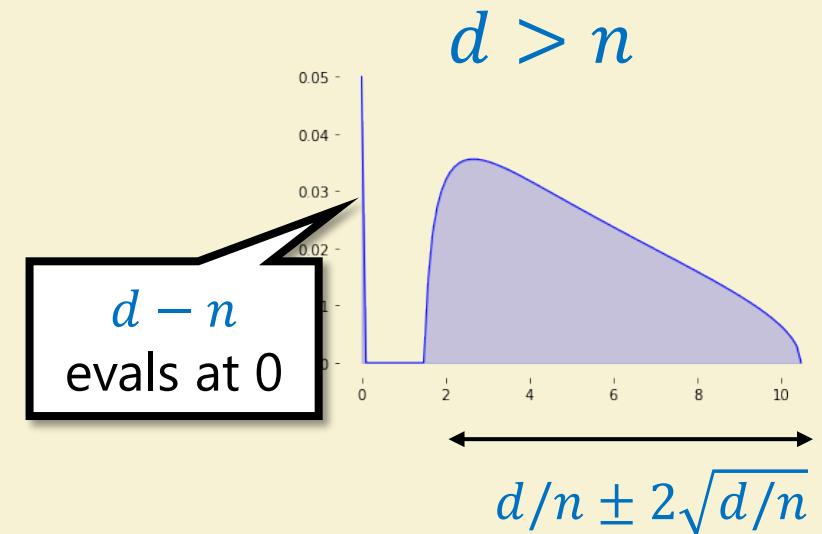
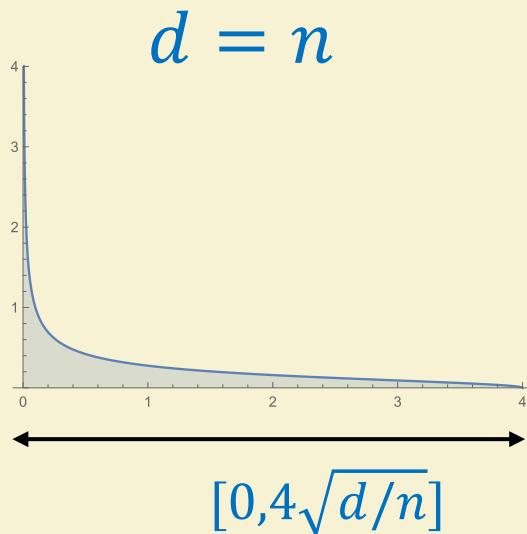
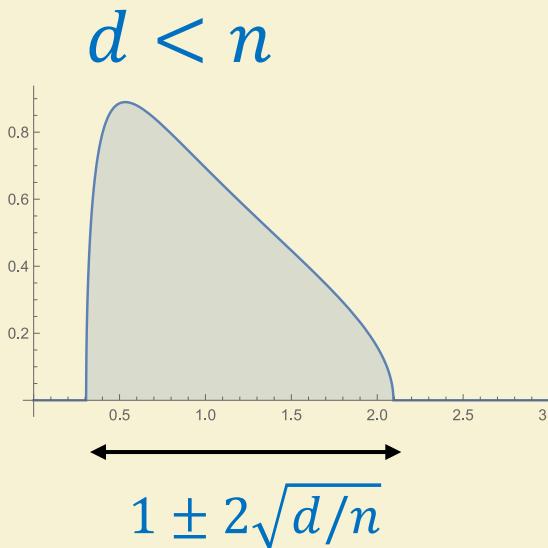


Random matrices

$$X = \frac{1}{n}AA^\top, A \in \mathbb{R}^{d \times n}, A_{i,j} \sim N(0,1)$$

X = empirical estimate for covariance of $x_1, \dots, x_n \sim N(0, I_d)$

Spectrum distributed according to **Marchenko-Pastur**



Digression:

Multiplicative Weights Update /
Follow The Regularized Leader /
Regret Minimization /
Mirror Descent

Digression: Multiplicative Weights

"Experts"
model

Setup: n possible actions a_1, \dots, a_n

At time $t = 1, 2, \dots, T$ learn loss $L_{i,t}$ for action i at time t

Approach:

- Initialize p_0 distribution over $[n]$
- p_{t+1} updates p_t by letting $p_{t+1}(i) \propto p_t(i) \exp(-\eta L_{i,t})$

Hope: Converge to "good aggregation"

Average
loss

$$\mathbb{E}_{t \sim [T]} \mathbb{E}_{i \sim p_t} L_{i,t} \approx \min_{p^*} \mathbb{E}_{t \sim [T]} \mathbb{E}_{i \sim p_t} L_{i,t}$$

Best loss in
hindsight

LHS – RHS = (average) **regret**

Multiplicative weights

- p_{t+1} updates p_t by letting $p_{t+1}(i) \propto p_t(i) \exp(-\eta L_{i,t})$

Hope: $\mathbb{E}_{t \sim [T]} \mathbb{E}_{i \sim p_t} L_{i,t} \approx \min_{p^*} \mathbb{E}_{t \sim [T]} \mathbb{E}_{i \sim p_t} L_{i,t}$

$$\text{THM: } (\mathbb{E}_t \mathbb{E}_{p_t} L - \mathbb{E}_t \mathbb{E}_{p^*} L) \leq \underbrace{\frac{\Delta_{KL}(p^* \| p_0)}{\eta \cdot T}}_{\text{Regret}} + \underbrace{O(\eta)}_{\text{Prior ignorance}} + \underbrace{o\left(\sqrt{\frac{\log n}{T}}\right)}_{\text{Sensitivity per step}}$$

* Assuming $|L_{i,t}| \leq O(1)$

Multiplicative weights

$$p_{t+1}(i) = p_t(i) \exp(-\eta L_{i,t}) / Z_t$$

- p_{t+1} updates p_t by letting $p_{t+1}(i) \propto p_t(i) \exp(-\eta L_{i,t})$

THM: $\underbrace{(\mathbb{E}_t \mathbb{E}_{p_t} L - \mathbb{E}_t \mathbb{E}_{p^*} L)}_{\text{Regret}} \leq \frac{\Delta_{KL}(p^* \| p_0)}{\eta \cdot T} + O(\eta)$

$$L_{t,i} = \frac{1}{\eta} \log \frac{p_t(i)}{p_{t+1}(i) \cdot Z_i}$$

PF:

$$\begin{aligned} \text{Regret} &= \frac{1}{T} \sum_t \left(- \mathbb{E}_{p_t} L - \mathbb{E}_{p^*} L \right) \\ &= \frac{1}{\eta \cdot T} [\mathbb{E}_{p^*} \log p_T - \mathbb{E}_{p^*} \log p_0] + \frac{1}{\eta \cdot T} \sum_t \Delta_{KL}(p_t, p_{t+1}) \end{aligned}$$

* Assuming $|L_{i,t}| \leq O(1)$

Multiplicative weights

- p_{t+1} updates p_t by letting $p_{t+1}(i) \propto p_t(i) \exp(-\eta L_{i,t})$

$$\text{THM: } \underbrace{(\mathbb{E}_t \mathbb{E}_{p_t} L - \mathbb{E}_t \mathbb{E}_{p^*} L)}_{\text{Regret}} \leq \frac{\Delta_{KL}(p^* \| p_0)}{\eta \cdot T} + O(\eta)$$

$$\begin{aligned} \text{PF: Regret} &= \frac{1}{\eta \cdot T} \left[\mathbb{E}_{p^*} \log p_T - \mathbb{E}_{p^*} \log p_0 \right] + \frac{1}{\eta \cdot T} \sum_t \Delta_{KL}(p_t, p_{t+1}) \\ &\quad || \\ &\quad \frac{1}{\eta \cdot T} [H(p_*, p_0) - H(p^*, p_T)] \\ &\leq \frac{1}{\eta \cdot T} [H(p_*, p_0) - H(p^*)] \\ &= \frac{1}{\eta \cdot T} \Delta_{KL}(p^* \| p_0) \end{aligned}$$

* Assuming $|L_{i,t}| \leq O(1)$

Multiplicative weights

- p_{t+1} updates p_t by letting $p_{t+1}(i) \propto p_t(i) \exp(-\eta L_{i,t})$

$$\text{THM: } \underbrace{(\mathbb{E}_t \mathbb{E}_{p_t} L - \mathbb{E}_t \mathbb{E}_{p^*} L)}_{\text{Regret}} \leq \frac{\Delta_{KL}(p^* \| p_0)}{\eta \cdot T} + O(\eta)$$

$$\text{PF: Regret} \leq \frac{1}{\eta \cdot T} \Delta_{KL}(p^* \| p_0) + \frac{1}{\eta \cdot T} \sum_t \Delta_{KL}(p_t, p_{t+1})$$

* Assuming $|L_{i,t}| \leq O(1)$

Multiplicative weights

- p_{t+1} updates p_t by letting $p_{t+1}(i) \propto p_t(i) \exp(-\eta L_{i,t})$

$$\text{THM: } \underbrace{(\mathbb{E}_t \mathbb{E}_{p_t} L - \mathbb{E}_t \mathbb{E}_{p^*} L)}_{\text{Regret}} \leq \frac{\Delta_{KL}(p^* \| p_0)}{\eta \cdot T} + O(\eta)$$

$$\text{PF: Regret} \leq \frac{\Delta_{KL}(p^* \| p_0)}{\eta \cdot T} + \frac{1}{\eta \cdot T} \sum_t \Delta_{KL}(p_t, p_{t+1})$$

CLM: If p, q s.t. $p(i) \propto q(i)\rho_i$ for $\rho_i \in [1 - \eta, 1 + \eta]$ then $\Delta_{KL}(p \| q) \leq O(\eta^2)$

PF: $p(i) = q(i)\rho_i/Z$ where $Z = \mathbb{E}_q \rho_i$

$$\Delta_{KL}(p \| q) = \mathbb{E}_p \log \rho_i - \log Z = \mathbb{E}_p \log \rho_i - \log \mathbb{E}_q \rho_i$$

$$\leq \mathbb{E}_p \log \rho_i - \mathbb{E}_q \log \rho_i \leq \max \log(\rho_i) \cdot \sum (p_i - q_i) \leq O(\eta^2)$$



* Assuming $|L_{i,t}| \leq O(1)$

Generalization: Follow The Regularized Leader

Set K of actions

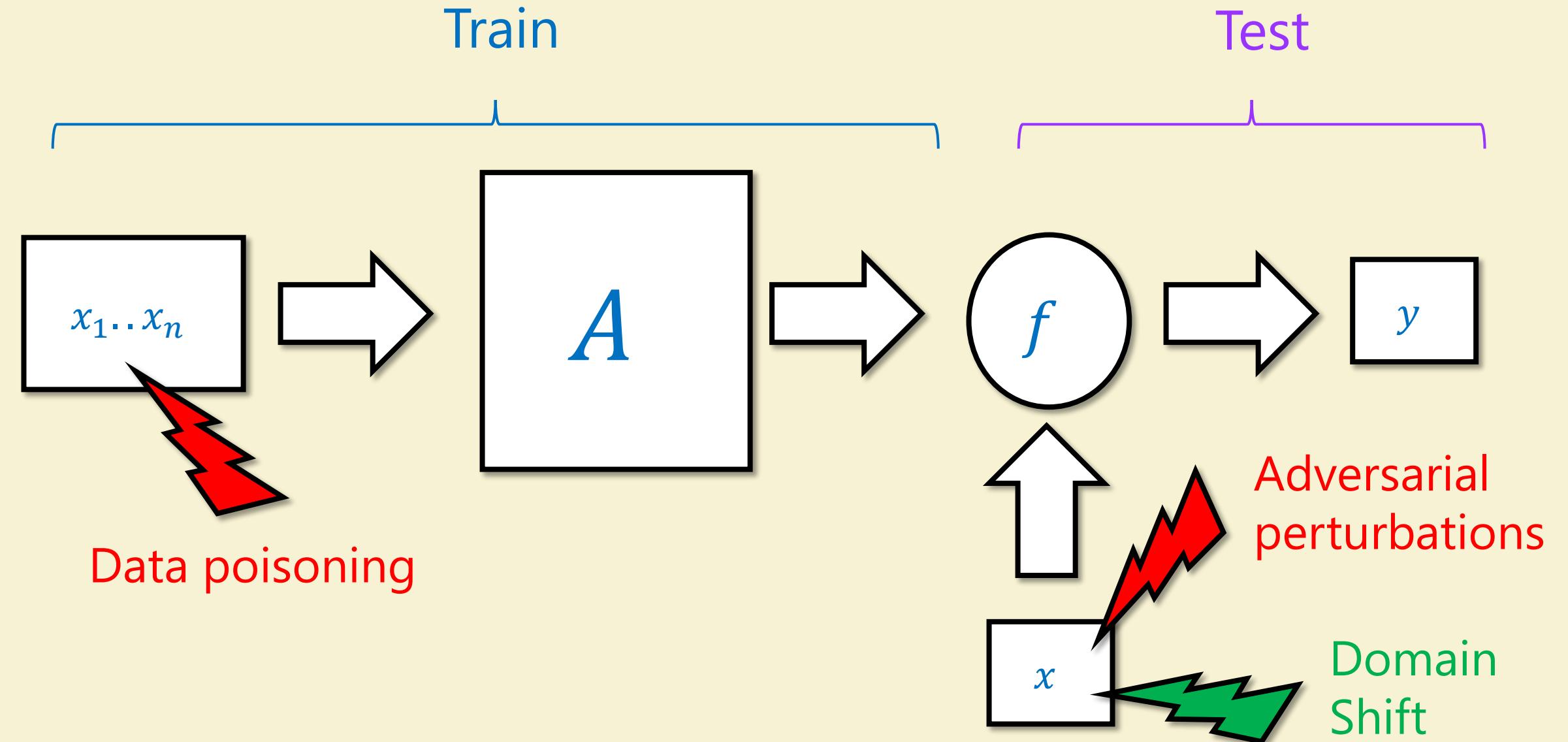
At time $t + 1$, make choice $x_{t+1} \in K$ and learn cost function $L_{t+1}: K \rightarrow \mathbb{R}$

$$\text{FTRL: } x_{t+1} = \arg \min_{x \in K} \left(R(x) + \sum_{i=1}^t L_i(x) \right)$$

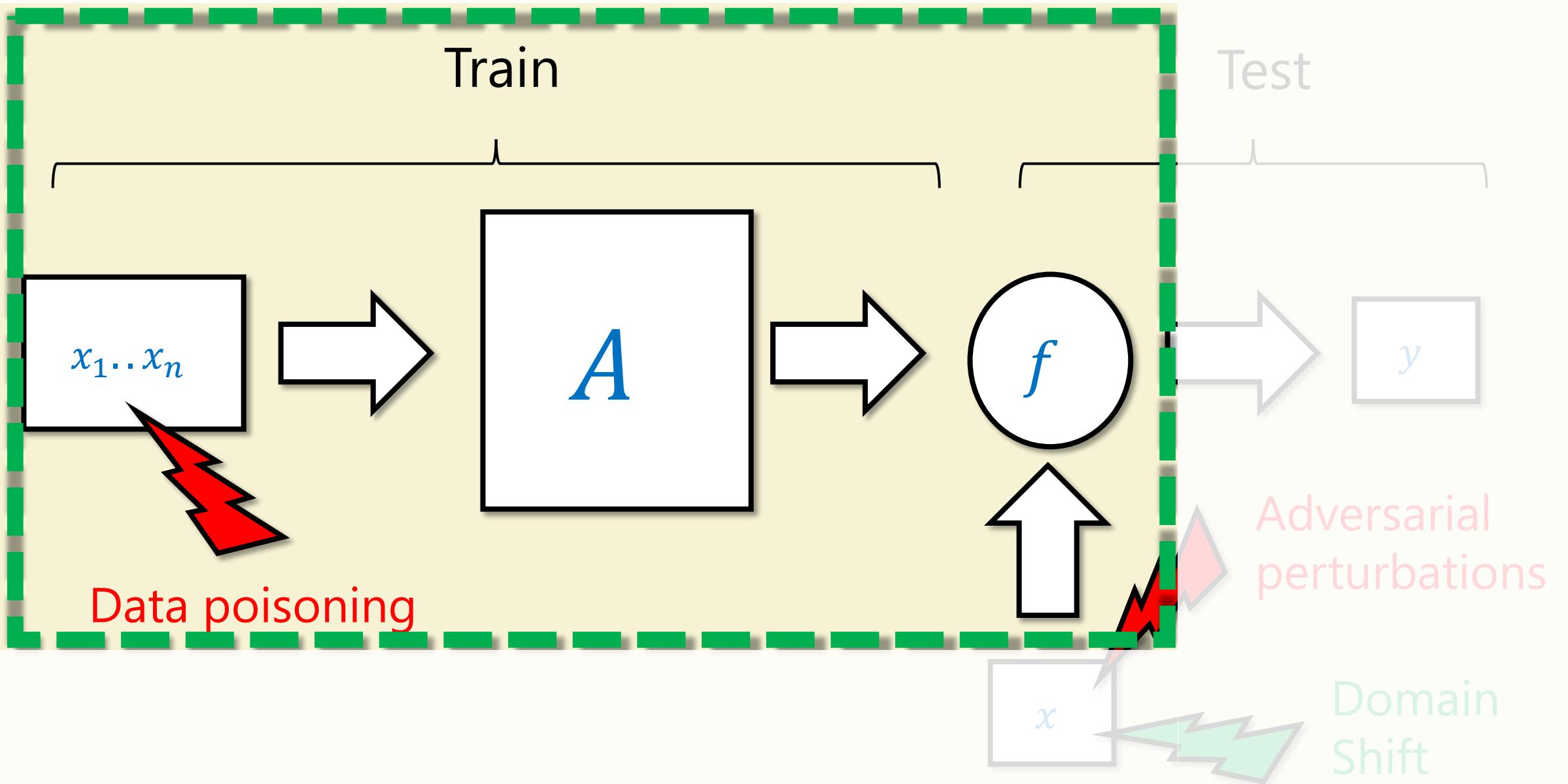
$$\text{THM: Mean regret at } T \leq \frac{1}{T} \left[\underbrace{R(x^*) - R(x_0)}_{\text{Prior ignorance}} + \underbrace{\sum_{t=1}^T (L_t(x_t) - L_t(x_{t+1}))}_{\text{Sensitivity per step}} \right]$$

Multiplicative Weights: $K = \{ \text{dists on } [n] \} , \quad R(x) = -\frac{1}{\eta} H(x)$

Robustness



Robustness



Train-Time Robustness.

Assume $\|x_i\|^2 \approx 1$ for $i < (1 - \epsilon)n$

Model: $x_1, \dots, x_{(1-\epsilon)n} \sim X \subseteq \mathbb{R}^d$, $x_{(1-\epsilon)n+1}, \dots, x_n$ arbitrary.

Mean estimation: Estimate $\mu = \mathbb{E}X$

Noiseless case: Empirical mean $\hat{\mu} = \frac{1}{n} \sum x_i$

- $d = 1$: $|\hat{\mu} - \mu| \leq O(1/\sqrt{n})$
- General d : $\|\hat{\mu} - \mu\| \leq O(\sqrt{d/n})$

Adversarial case: Empirical mean arbitrarily bad

For $d = 1$ can use empirical median $\mu^* = \text{sort}(x_1, \dots, x_n)_{n/2}$

Guaranteed to lie in $(\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon)$ quantile of real data

Train-Time Robustness.

Model: $x_1, \dots, x_{(1-\epsilon)n} \sim X \subseteq \mathbb{R}^d$, $x_{(1-\epsilon)n+1}, \dots, x_n$ arbitrary.

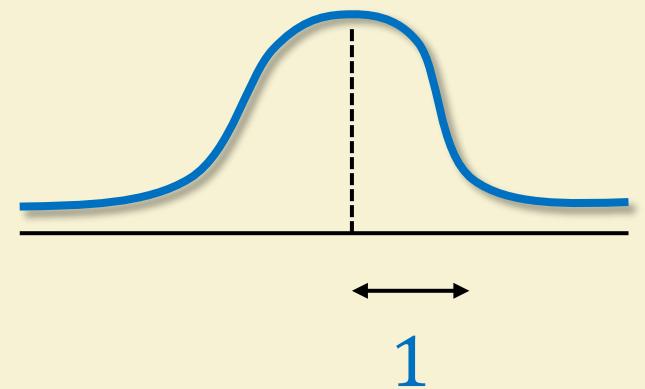
Mean estimation: Estimate $\mu = \mathbb{E}X$

For $d = 1$ can use empirical median $\mu^* = \text{sort}(x_1, \dots, x_n)_{n/2}$

Guaranteed to lie in $(\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon)$ quantile of real data

If $X \sim N(\mu, 1)$, $|\mu - \mu^*| \leq O(\epsilon + 1/\sqrt{n})$

Inherent!



Train-Time Robustness.

Model: $x_1, \dots, x_{(1-\epsilon)n} \sim X \subseteq \mathbb{R}^d$, $x_{(1-\epsilon)n+1}, \dots, x_n$ arbitrary.

Mean estimation: Estimate $\mu = \mathbb{E}X$

For $d = 1$ can use empirical median $\mu^* = \text{sort}(x_1, \dots, x_n)_{n/2}$

If $X \sim N(\mu, 1)$, $|\mu - \mu^*| \leq O(\epsilon + 1/\sqrt{n})$

Median of coordinates: ($d \geq 1$) $\mu_i^* = \text{sort}(x_{1,i}, \dots, x_{n,i})_{n/2}$

If $X \sim N(\mu, I)$, can make $\mu^* \approx (\epsilon, \dots, \epsilon) \Rightarrow \|\mu^* - \mu\| \approx \epsilon\sqrt{d}$

Can we do better?

John W. Tukey

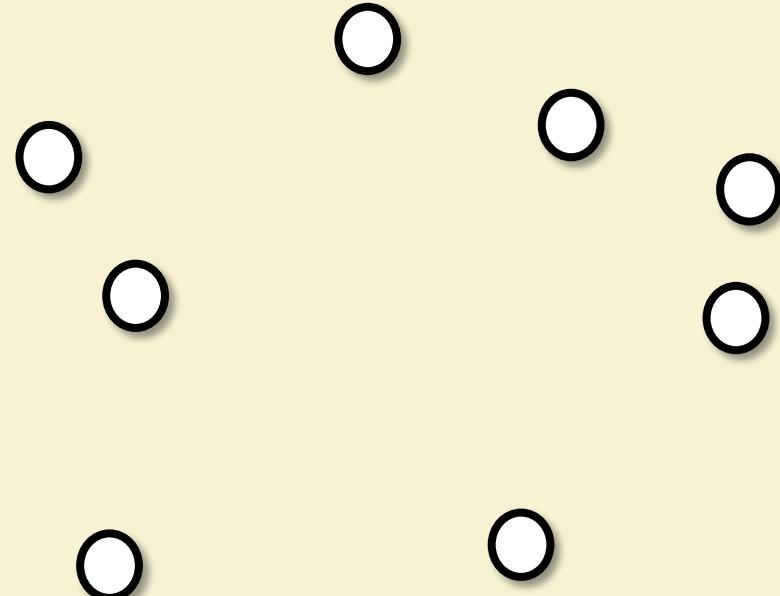
1. Introduction. Why am I writing on this topic? Partly because picturing of data is important. Partly because, if present trends continue, an increasing fraction of all mathematicians will touch—or come close to touching—data during the next few decades. Mathematicians have many advantages in approaching data—and one major disadvantage. Those mathematicians who might come close to data need to know their advantages from their disadvantages.

Experience and facility with clear thinking—and with varied sorts of calculi that lead step-by-step from start to conclusion—knowledge of a variety of mathematical structures—even some of the more abstract are sometimes relevant to data—these are great advantages. The habit of building one technique on another—of assembling procedures like something made of erector-set parts—can be especially useful in dealing with data. So too is looking at the same thing in many ways or many things in the same way; an ability to generalize in profitable ways and a liking for a massive search for order. Mathematicians understand how subtle assumptions can make great differences and are used to trying to trace the paths by which this occurs. The mathematician's great disadvantage in approaching data is his—or her—attitude toward the words “hypothesis” and “hypotheses”.

Tukey Median

A Tukey median* of x_1, \dots, x_n is μ^* s.t. for every nonzero $v \in \mathbb{R}^d$

$$\# i \text{ s.t. } \langle x_i - \mu^*, v \rangle > 0 \text{ is in } \left(\frac{1}{2} \pm 3\epsilon\right)n$$

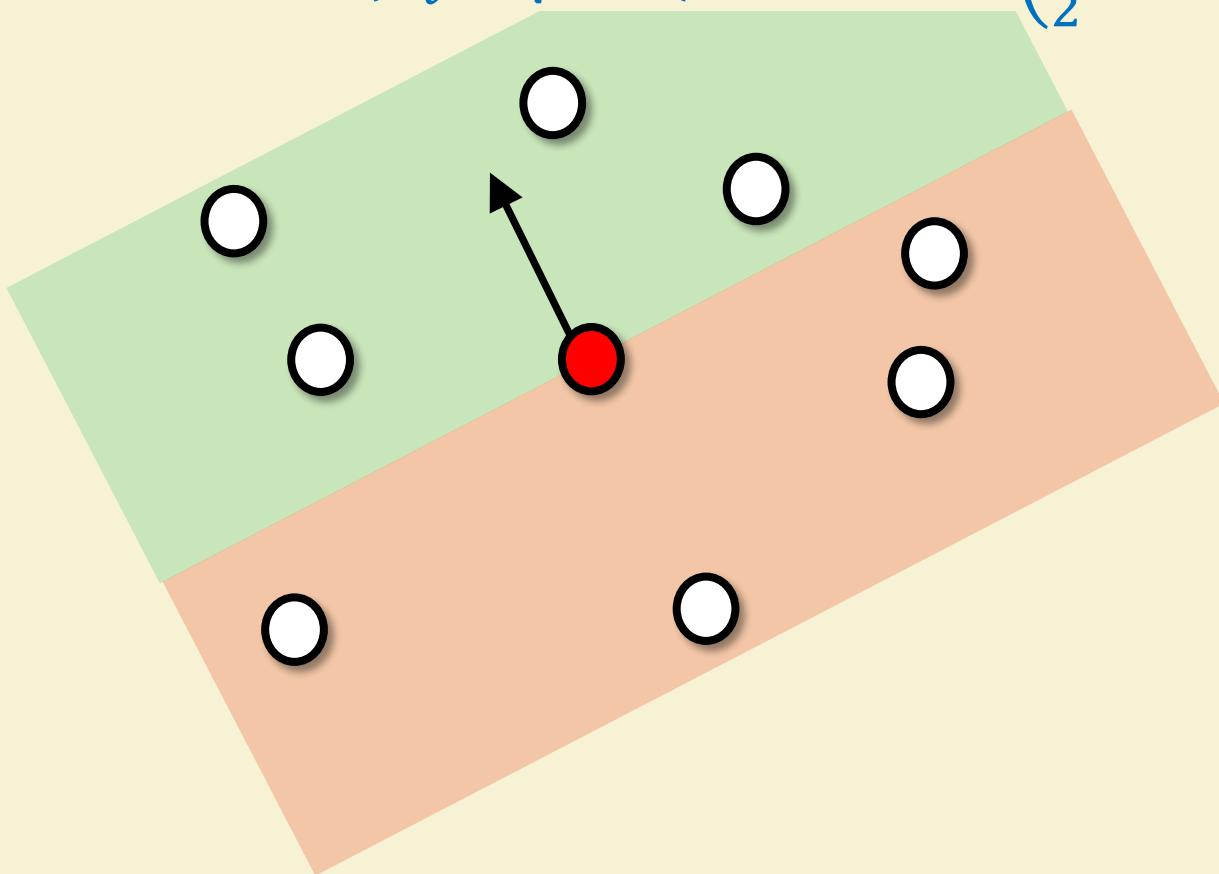


* Formal def: Tukey median is $\arg \max_{\mu^*} \min_{v \neq 0} |\{i | \langle x_i - \mu^*, v \rangle > 0\}|$

Tukey Median

A Tukey median* of x_1, \dots, x_n is μ^* s.t. for every nonzero $v \in \mathbb{R}^d$

$$\# i \text{ s.t. } \langle x_i - \mu^*, v \rangle > 0 \text{ is in } \left(\frac{1}{2} \pm 3\epsilon\right)n$$



Tukey Median

A Tukey median* of x_1, \dots, x_n is μ^* s.t. for every nonzero $v \in \mathbb{R}^d$

$$\# i \text{ s.t. } \langle x_i - \mu^*, v \rangle > 0 \text{ is in } \left(\frac{1}{2} \pm 3\epsilon\right)n$$

THM: If $x_1, \dots, x_{(1-\epsilon)n} \sim N(\mu, I)$ and $\sqrt{d/n} \ll \epsilon$ then

- 1) Exists Tukey median μ^*
- 2) $\|\mu - \mu^*\| \leq O(\epsilon)$

A Tukey median* of x_1, \dots, x_n is μ^* s.t. for every nonzero $v \in \mathbb{R}^d$

$$\# i \text{ s.t. } \langle x_i - \mu^*, v \rangle > 0 \text{ is in } \left(\frac{1}{2} \pm 3\epsilon\right)n$$

THM: If $x_1, \dots, x_{(1-\epsilon)n} \sim N(\mu, I)$ and $\sqrt{d/n} \ll \epsilon$ then

- 1) Exists Tukey median μ^*
- 2) $\|\mu - \mu^*\| \leq O(\epsilon)$

PF OF 1: μ is Tukey median:

$$\forall v \neq 0, Y_1 = \text{sign}(\langle x_1 - \mu, v \rangle), \dots, Y_k = \text{sign}(\langle x_k - \mu, v \rangle) \text{ i.i.d } \pm 1 \text{ vars}$$

$$\Pr[\sum Y_i > \epsilon k] < \exp(-\epsilon^2 n) \ll \exp(-d)$$

$$k = (1 - \epsilon)n$$

\Rightarrow can "enumerate" over all (unit) $v \in \mathbb{R}^d$



A Tukey median* of x_1, \dots, x_n is μ^* s.t. for every nonzero $v \in \mathbb{R}^d$

$$\#\ i \text{ s.t. } \langle x_i - \mu^*, v \rangle > 0 \text{ is in } \left(\frac{1}{2} \pm 3\epsilon\right)n$$

THM: If $x_1, \dots, x_{(1-\epsilon)n} \sim N(\mu, I)$ and $\sqrt{d/n} \ll \epsilon$ then

- 1) Exists Tukey median μ^*
- 2) $\|\mu - \mu^*\| \leq O(\epsilon)$

PF OF 2: Suppose $\|\mu - \mu^*\| \geq 10\epsilon$ Let $v = \mu - \mu^*$

$N(0, 1)$

$$\langle x_i - \mu^*, v/\|v\| \rangle = \langle x_i - \mu + v, v/\|v\| \rangle = \langle x_i - \mu, v/\|v\| \rangle + 10\epsilon$$

Let $Y_i = \text{sign}(\langle x_i - \mu^*, v/\|v\| \rangle)$

$$\Pr[Y_i = -1] = \Pr[N(0, 1) \leq -10\epsilon] \leq 1/2 - 5\epsilon$$

\Rightarrow whp $< n/2 - 4\epsilon n$ i 's such that $Y_i = -1$

Efficient Algorithms

- :(Computing Tukey median is NP hard
- :) Efficient algorithms for robust mean estimation of normals, other distributions

Spectral Signatures

Let $x_1, \dots, x_{(1-\epsilon)n} \sim N(\mu, I)$ and $x_{(1-\epsilon)n+1}, \dots, x_n$ arbitrary

Let $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^\top$

Empirical mean

Empirical covariance matrix

Claim: $\|\hat{\mu} - \mu\| \leq O\left(\sqrt{\frac{d}{n}} + \sqrt{\epsilon}\|\hat{\Sigma}\|\right)$

Unknown quantity

Known quantity

* If all from $N(\mu, I)$ then $\|\hat{\Sigma}\| = O(\sqrt{d/n})$

Let $x_1, \dots, x_{(1-\epsilon)n} \sim N(\mu, I)$ and $x_{(1-\epsilon)n+1}, \dots, x_n$ arbitrary

Let $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^\top$

Claim: $\|\hat{\mu} - \mu\| \leq O\left(\sqrt{\frac{d}{n}} + \sqrt{\epsilon} \|\hat{\Sigma}\|\right)$

PF: ($\mu = 0$)

$$\|\hat{\mu}\|^2 = \langle \hat{\mu}, \hat{\mu} \rangle = \langle \hat{\mu}, \frac{1}{n} \sum x_i \rangle = \langle \hat{\mu}, \frac{1}{n} \sum_G x_i \rangle + \langle \hat{\mu}, \frac{1}{n} \sum_B x_i \rangle$$

$$\leq O\left(\sqrt{\frac{d}{n}} \cdot \|\hat{\mu}\|\right) + \frac{1}{n} \sum_B \langle \hat{\mu}, x_i - \hat{\mu} \rangle + \frac{1}{n} \sum_B \langle \hat{\mu}, \hat{\mu} \rangle$$

$$= \epsilon \|\hat{\mu}\|^2$$

$$(CS) \leq \sqrt{\frac{\epsilon}{n} \sum_B \langle \hat{\mu}, x_i - \hat{\mu} \rangle^2} \leq \sqrt{\epsilon \cdot \|\hat{\Sigma}\| \cdot \|\hat{\mu}\|^2} = \sqrt{\epsilon \cdot \|\hat{\Sigma}\| \cdot \|\hat{\mu}\|}$$

Claim: $\|\hat{\mu} - \mu\| \leq O\left(\sqrt{\frac{d}{n}} + \sqrt{\epsilon}\|\hat{\Sigma}\|\right)$

PF: ($\mu = 0$)

$$\|\hat{\mu}\|^2 = \langle \hat{\mu}, \hat{\mu} \rangle = \langle \hat{\mu}, \frac{1}{n} \sum x_i \rangle = \langle \hat{\mu}, \frac{1}{n} \sum_G x_i \rangle + \langle \hat{\mu}, \frac{1}{n} \sum_B x_i \rangle$$

$$\leq O\left(\sqrt{\frac{d}{n}} \cdot \|\hat{\mu}\|\right) + \frac{1}{n} \sum_B \langle \hat{\mu}, x_i - \hat{\mu} \rangle + \frac{1}{n} \sum_B \langle \hat{\mu}, \hat{\mu} \rangle$$

$$= \epsilon \|\hat{\mu}\|^2$$

$$(CS) \leq \sqrt{\frac{\epsilon}{n} \sum_B \langle \hat{\mu}, x_i - \hat{\mu} \rangle^2} \leq \sqrt{\epsilon \cdot \|\hat{\Sigma}\| \cdot \|\hat{\mu}\|^2} = \sqrt{\epsilon \cdot \|\hat{\Sigma}\|} \cdot \|\hat{\mu}\|$$

$$\Rightarrow (1 - \epsilon) \|\hat{\mu}\| \leq O\left(\sqrt{\frac{d}{n}} + \sqrt{\epsilon \cdot \|\hat{\Sigma}\|}\right)$$

Filtering

Let $x_1, \dots, x_{(1-\epsilon)n} \sim N(\mu, I)$ and $x_{(1-\epsilon)n+1}, \dots, x_n$ arbitrary

$$\max p_i \leq O(1/n)$$

\forall “flat” dist $p_1 \dots p_n$ let $\hat{\mu}(p) = \sum_{i=1}^n p_i x_i$ and $\hat{\Sigma}(p) = \sum_{i=1}^n p_i (x_i - \hat{\mu})(x_i - \hat{\mu})^\top$

Claim: $\|\hat{\mu}(p) - \mu\| \leq O\left(\sqrt{\frac{d}{n}} + \sqrt{\epsilon} \|\hat{\Sigma}(p)\|\right)$

Robust mean estimation:

1. Let $p_0 = (1/n, \dots, 1/n)$, $t = 0$
2. If $\|\hat{\Sigma}(p_t)\| \leq C$: return $\hat{\mu}(p_t)$
3. Otherwise let $v = v_{\max}(\hat{\Sigma}(p_t))$, $L_i = \langle x_i, v \rangle^2$
Let $p_{t+1}(i) \propto p_{t+1}(i) \cdot \exp(-\eta L_i)$ and go back to 2

Can do better:
See Jerry Li /
Jacob Steinhardt
lecture notes

Robust statistics via SoS

Mean Estimation
Moment Estimation
Linear Regression
Clustering Spherical Mixtures

} “certifiable
subgaussianity”

Clustering Non-Spherical Mixtures

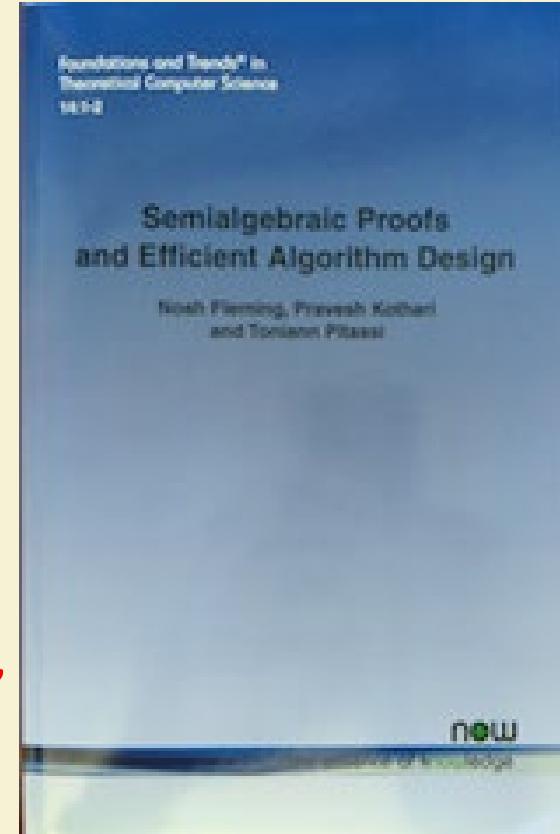
} “certifiable hypercontractivity”
“certifiable anti-concentration”

Heavy-Tailed Estimation
List-decodable mean estimation

} “certifiable
subgaussianity”

List-decodable regression
List-decodable subspace clustering

} “certifiable hypercontractivity”
“certifiable anti-concentration”



Refs

Mean Estimation [LRV'16],[DKKLMS'16], ...,[Hopkins-Li'18],[Kothari-Steurer'18]

Moment Estimation [Kothari-Steurer'18]

Linear Regression [Klivans-Meka-Kothari'18], [Diakonikolas, Kamath, Kane, Li, Steinhardt, Stewart'18],
[Prasad, Suggala, Balakrishnan and Ravikumar'18][Bakshi-Prasad'20]

Clustering Spherical Mixtures [Hopkins-Li'18],[Diakonikolas-Kane-Stewart'18],[Kothari-Steinhardt'18]

Clustering Non-Spherical Mixtures [Bakshi-Kothari'20],[Diakonikolas-Hopkins-Kane-Karmalkar'20]

Learning Arbitrary Mixtures [Bakshi-Diakonikolas-Jia-Kane-Kothari-Vempala'20],[Liu-Moitra'20]

Heavy-Tailed Estimation [Hopkins'19],[Cherapanamjeri-Hopkins-Kathuria-Raghavendra-Tripuraneni'20]

List-decodable mean estimation [Diakonikolas-Kane-Stewart'18],[Kothari-Steinhardt'18]

List-decodable regression [Karmalkar-Klivans-Kothari'19],[Raghavendra-Yau'19]

List-decodable subspace clustering [Bakshi-Kothari'20],[Raghavendra-Yau'20]

In Neural Networks

Embedding: r : data $\rightarrow \mathbb{R}^m$

(semi) cartoon:

2 std devs
certainty of dog

$$r(\text{dog}) = (2, 0.6, \dots, \dots, \dots, \dots)$$

$$r(\text{cat}) = (0.1, 1.5, \dots, \dots, \dots, \dots, \dots)$$



Dog
dimension

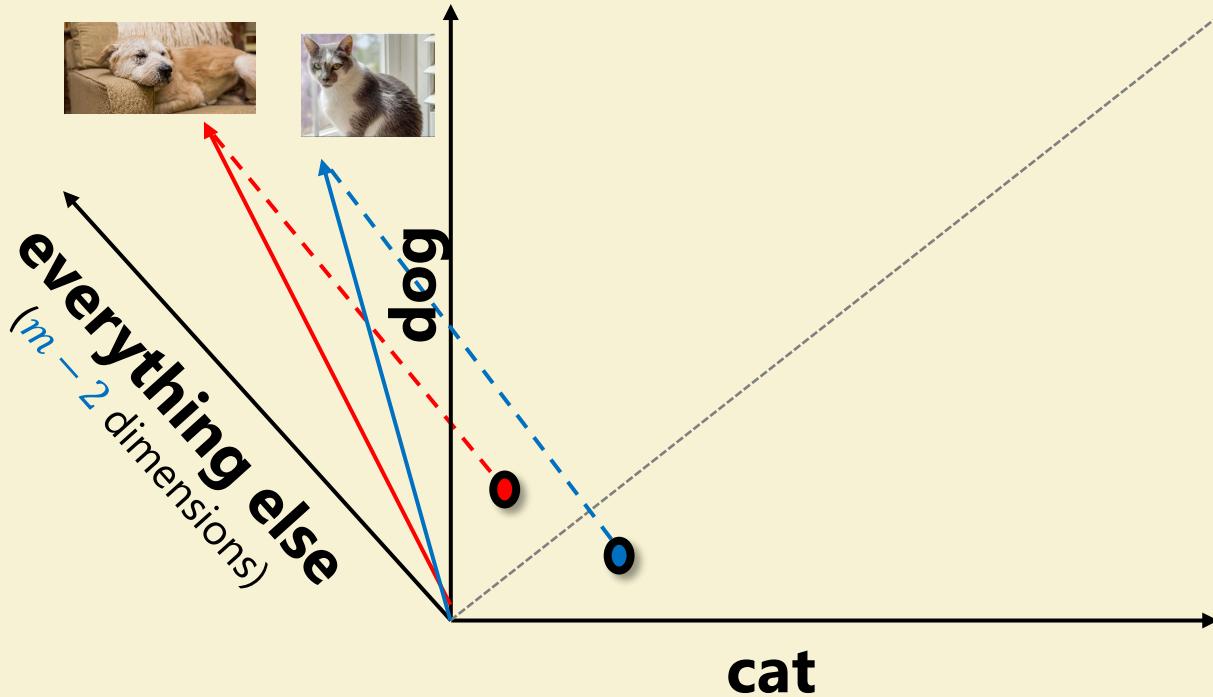


Cat
dimension

$m - 2$ other dimensions

Most norm is here

In Neural Networks

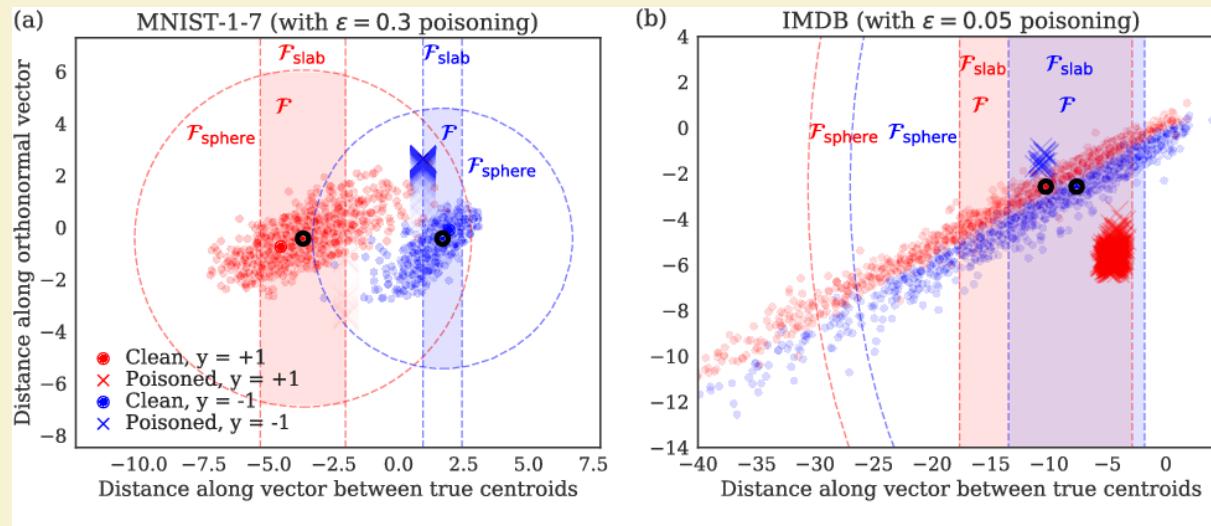
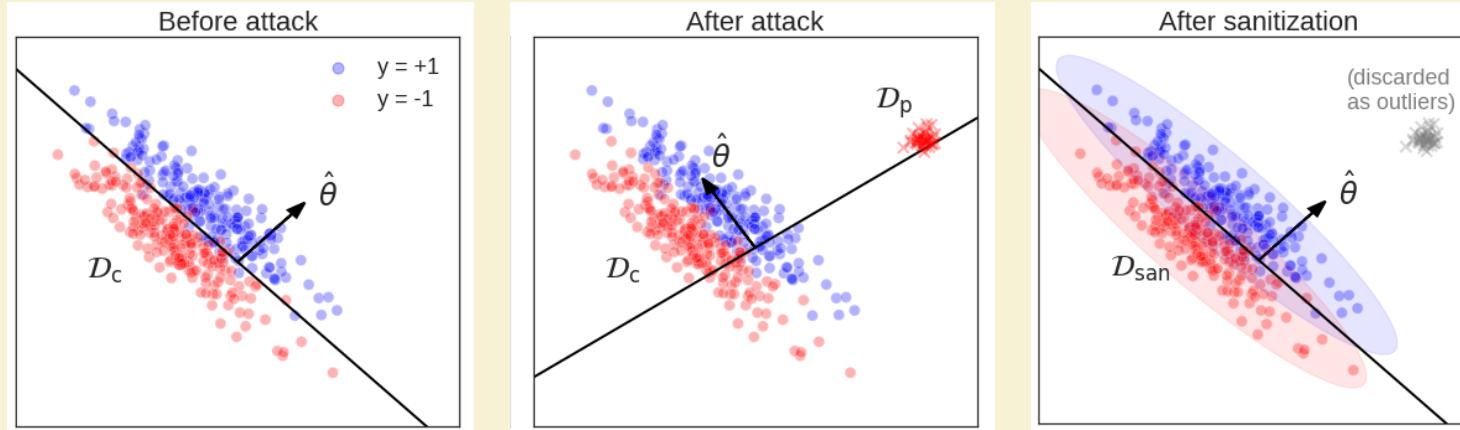


$$\langle v, \text{dog} \rangle, \langle v, \text{cat} \rangle \ll \|v\|$$

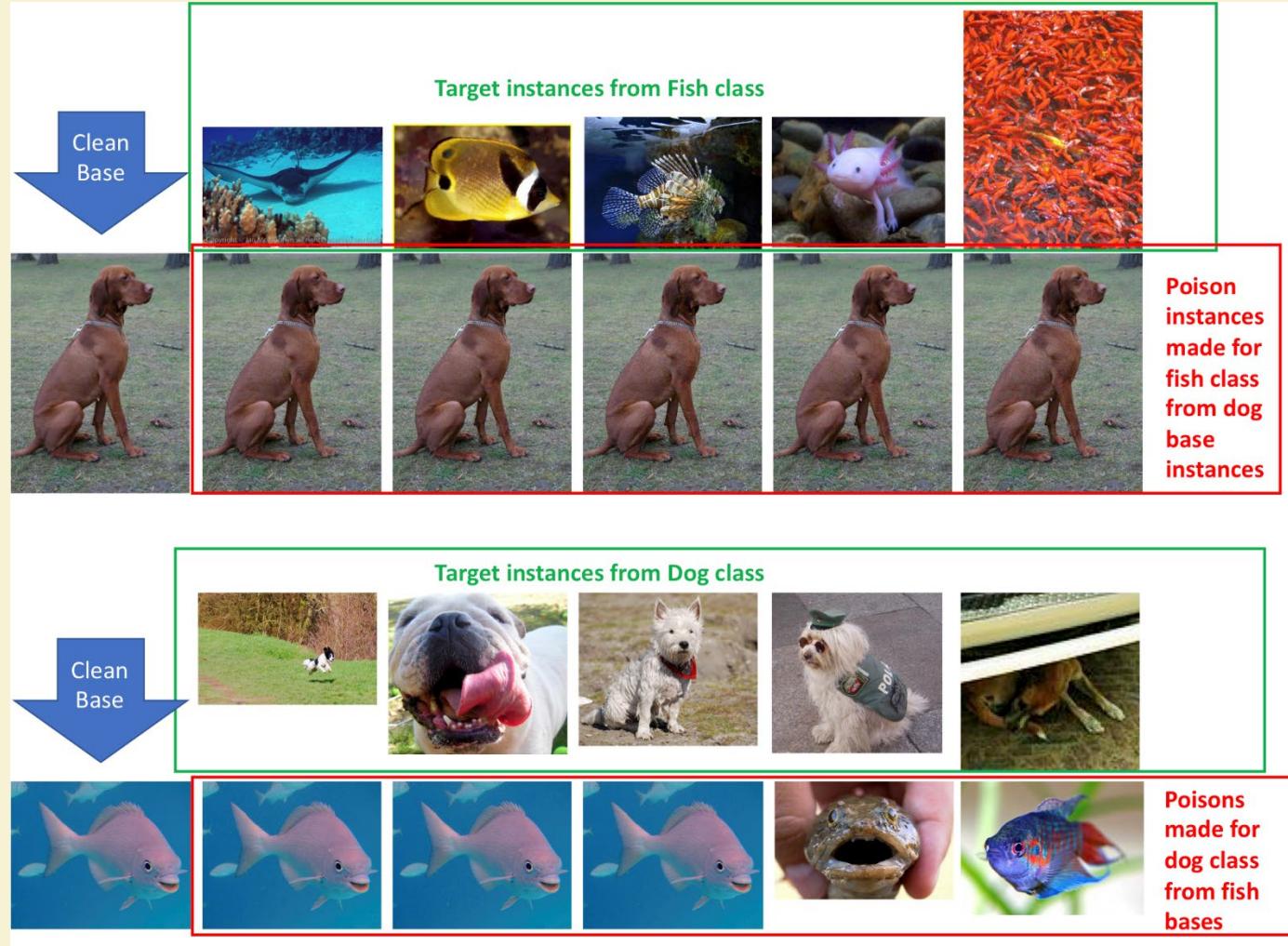
Slightly perturb dog : data poisoning (train time)

Slightly perturb v : adversarial perturbation (test time)

Data Poisoning attacks and defenses



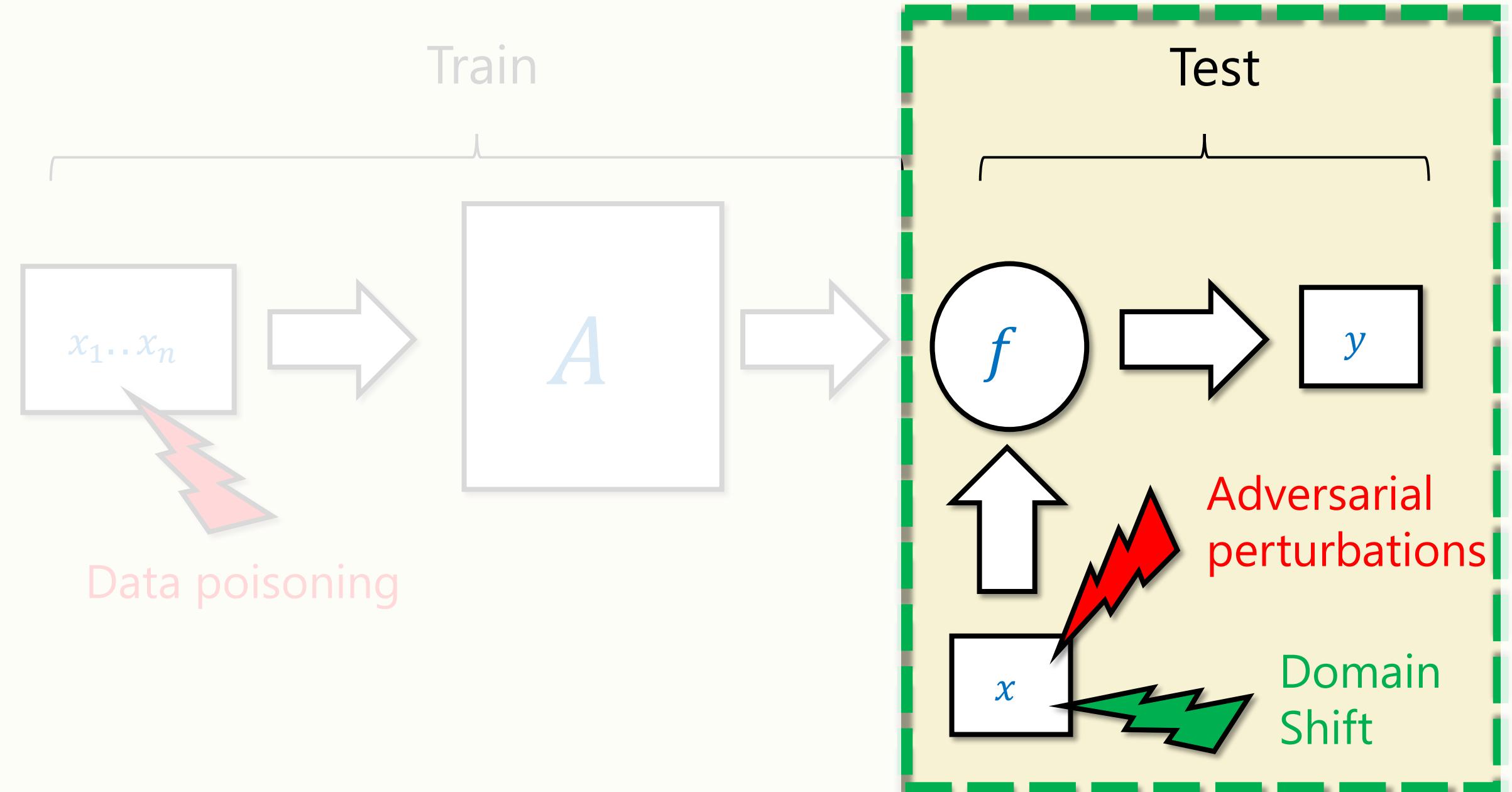
Data Poisoning attacks and defenses



Train Time Robustness

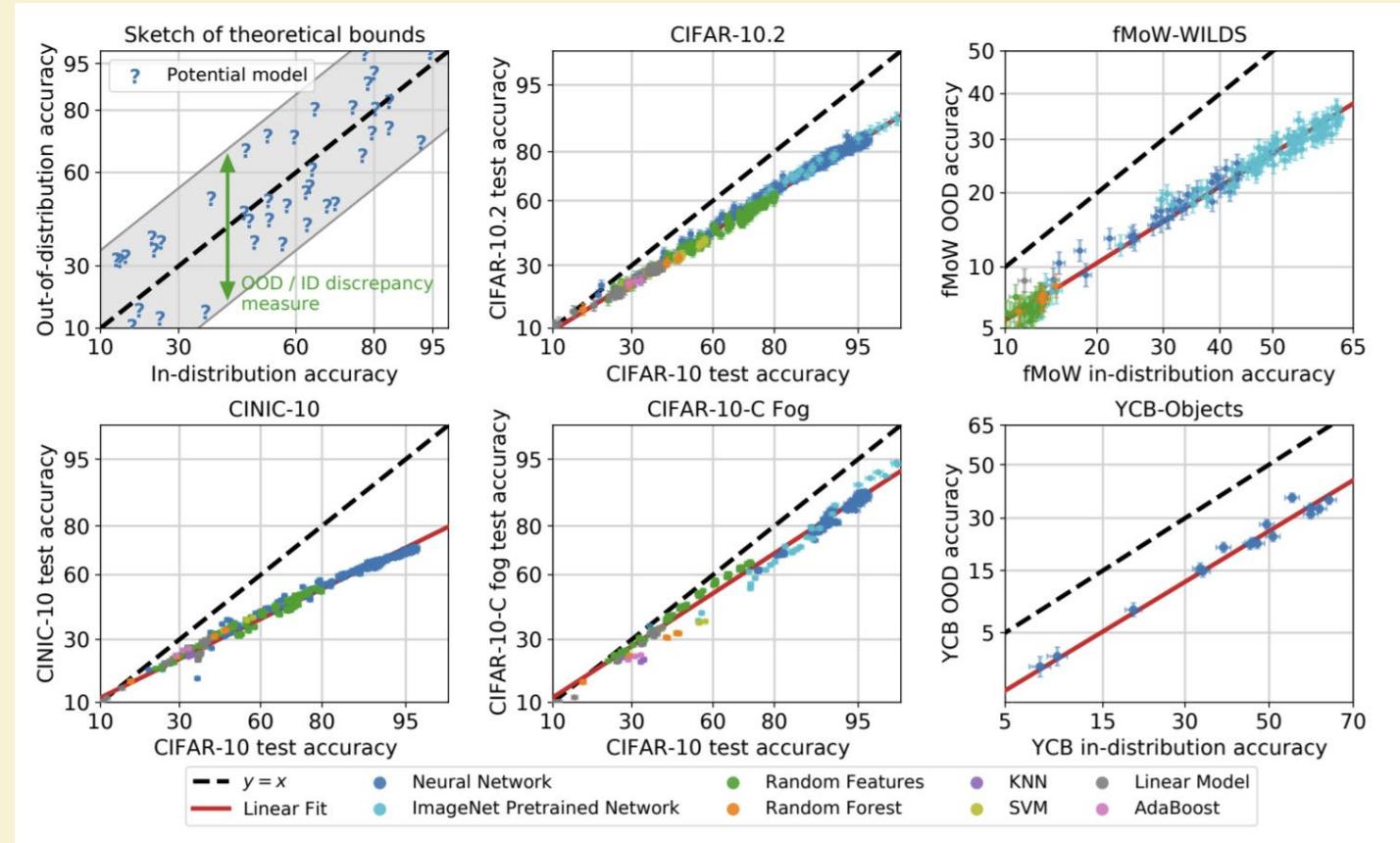
- Out of distribution
- Adversarial perturbation

Robustness



Domain Shift

- Train on data $(x_1, y_1), \dots, (x_n, y_n) \sim D$
- Test on $(x, y) \sim D'$



CIFAR 10 (Krizhevsky, Nair, Hinton 2009)

50K+10K 32x32 color images, 10 classes

Base: [Tiny Images](#) (Torralba, Fergus, Freeman '08)

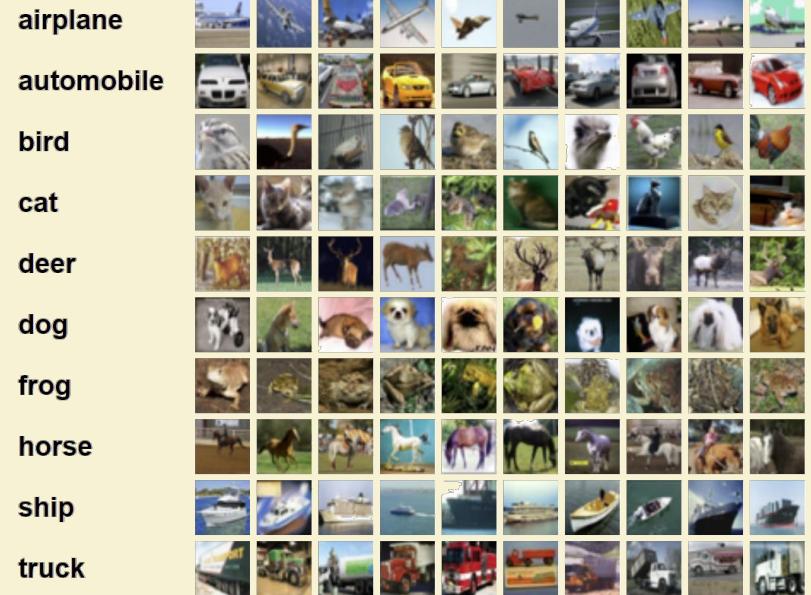
Extracted 75K nouns from WordNet

Used these to search images from 7 search engines
(Altavista, Ask, Flickr, Cydral, Google, Picsearch and Webshots)

Up to 3K images per search term, subsample to 32x32

Construction:

- Student labelers considered all images corresponding to class + hyponym
- Filtered to 6K images per class, split to 5K vs 1K



CIFAR 10.1 and 10.2

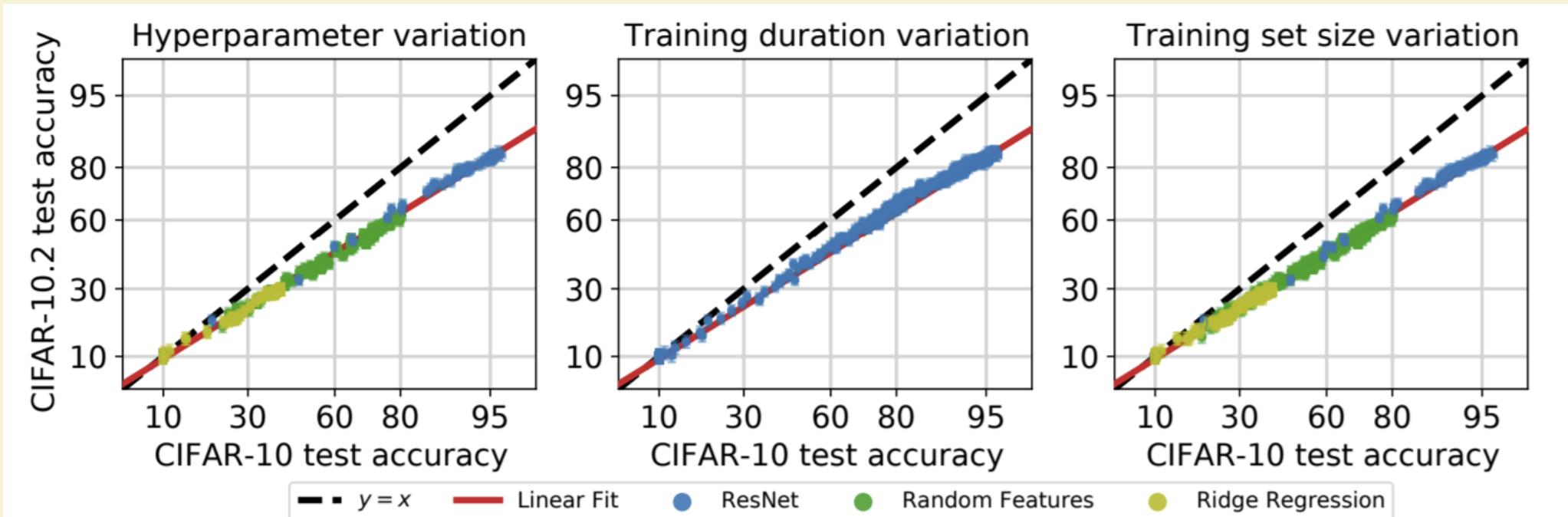
Recht-Roelofs-Schmidt-Shankar '19

Lu-Nott-Olson-Todeschini-Vahabi-Carmon-Schmidt '20

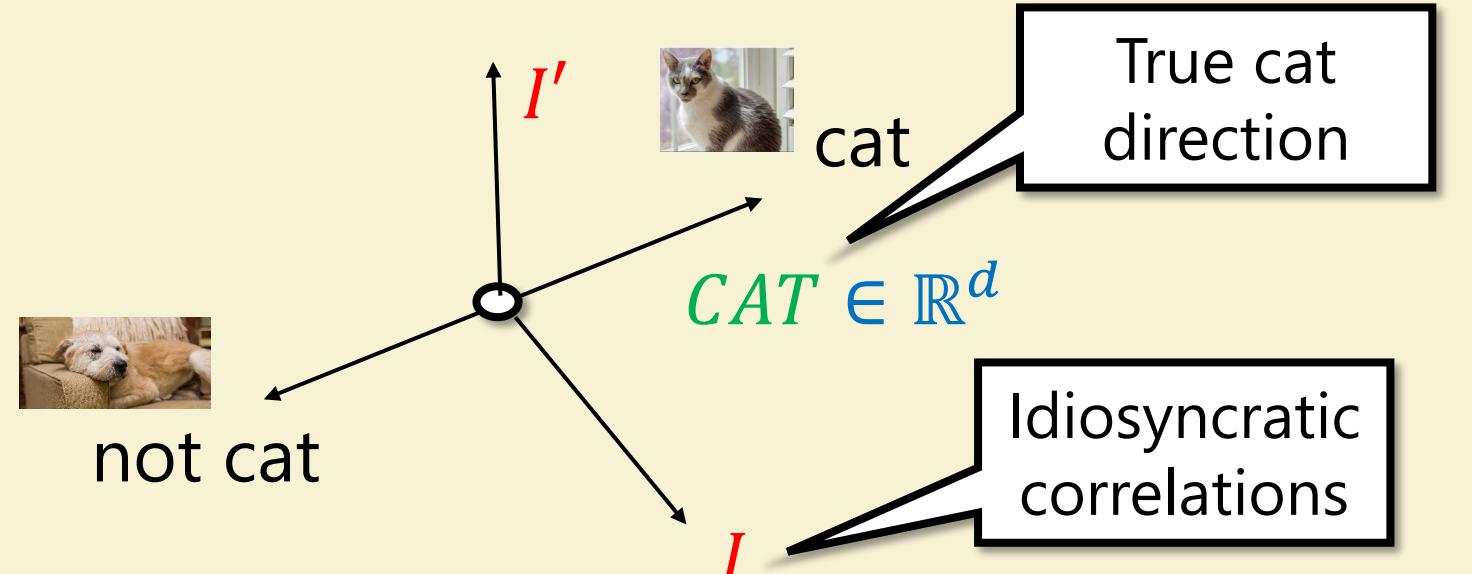
Replicated datasets can be easier/harder
<http://gradientscience.org/data rep bias/>

Followed same procedure but smaller size

- **CIFAR 10.1:** Test set only (2K images)
- **CIFAR 10.2:** Train + test (10K + 2K)



Theoretical model for domain shift



Dataset D : $\Pr[x \text{ labeled cat}] \propto \exp(\beta \langle x, \text{CAT} + \alpha I \rangle)$

Dataset D' : $\Pr[x \text{ labeled cat}] \propto \exp(\beta' \langle x, \text{CAT} + \alpha' I' \rangle)$

$$Acc_D(C) = \beta \langle C, \text{CAT} \rangle + \beta \alpha \langle C, I \rangle$$

$$Acc_{D'}(C) = \beta' \langle C, \text{CAT} \rangle + \beta' \alpha' \langle C, I' \rangle$$

If C trained on D
assume $\langle C, I' \rangle \approx 0$

Theoretical model for domain shift

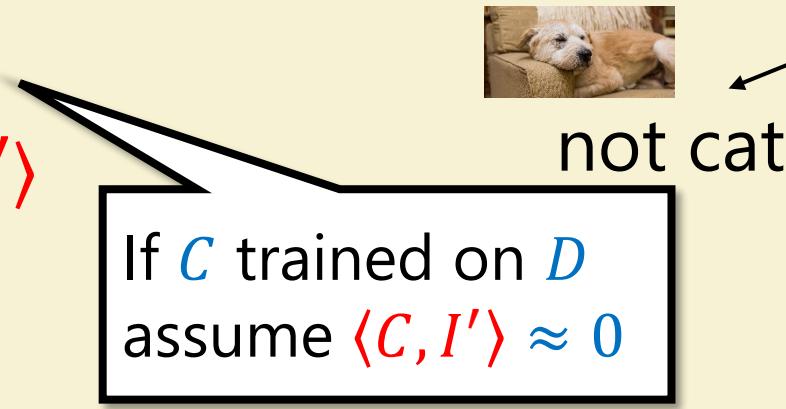


cat

$$Acc_D(C) = \beta \langle C, CAT \rangle + \beta \alpha \langle C, I \rangle$$

$$Acc_{D'}(C) = \beta' \langle C, CAT \rangle + \beta' \alpha' \langle C, I' \rangle$$

If learn by grad descent



$$\nabla(\beta \langle C, CAT \rangle + \alpha \beta \langle C, I \rangle) = \beta \cdot CAT + \beta \alpha \cdot I$$

$$\text{If } C \text{ trained on } D, C \propto CAT + \alpha \cdot I + \text{Noise} = \frac{\gamma}{\|CAT + \alpha \cdot I\|^2} (CAT + \alpha \cdot I) + \text{Noise}$$

$$Acc_D(C) = \beta \langle C, CAT + \alpha \cdot I \rangle = \beta \gamma$$

$$Acc_{D'}(C) = \beta' \langle C, CAT + \alpha \cdot I' \rangle = \beta' \gamma \cdot \frac{\|CAT\|^2}{\|CAT\|^2 + \alpha^2 \|I\|^2}$$

Theoretical model for domain shift



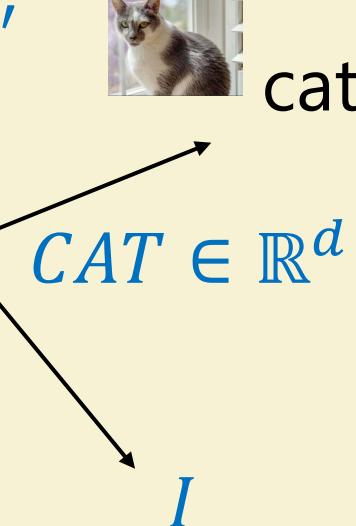
cat

$$Acc_D(C) = \beta \langle C, CAT + \alpha \cdot I \rangle = \beta \gamma$$

$$Acc_{D'}(C) = \beta' \langle C, CAT + \alpha \cdot I' \rangle = \beta' \gamma \cdot \frac{\|CAT\|^2}{\|CAT\|^2 + \alpha^2 \|I'\|^2}$$

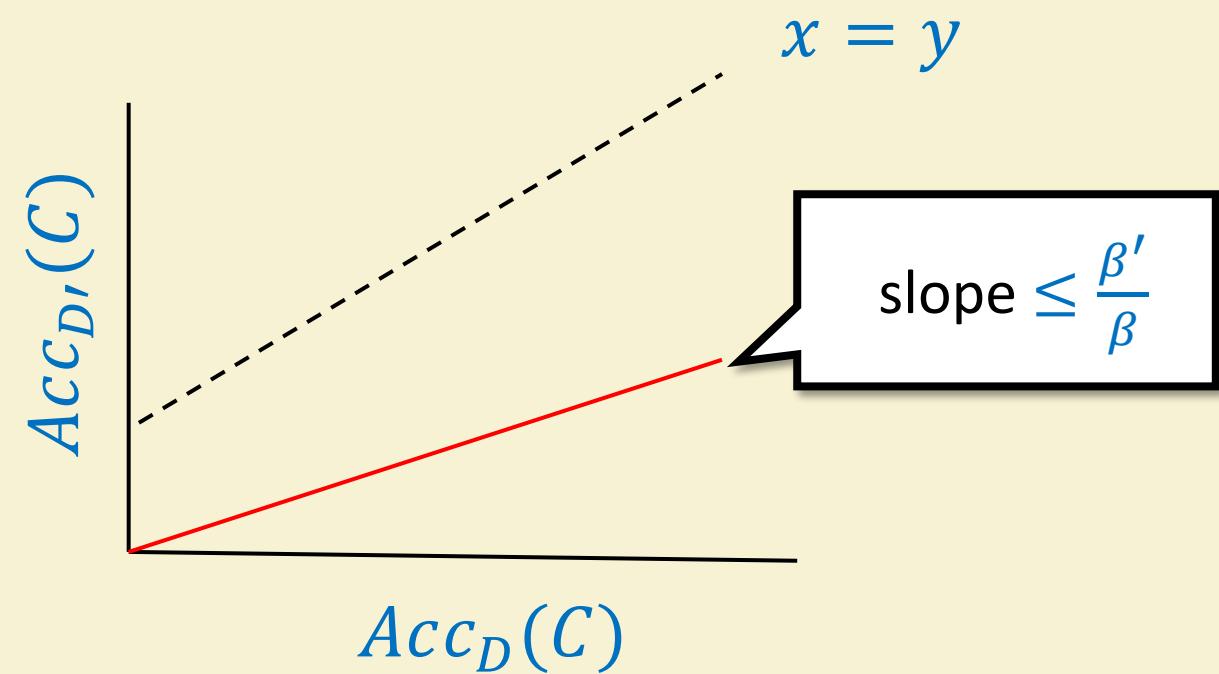


not cat



$$Acc_{D'}(C) = \frac{\beta'}{\beta(1 + \theta^2)} \cdot Acc_D(C)$$

- $\beta'/\beta < 1$ iff D' harder than D
- θ^2 grows with idiosyncratic component of D



Theoretical model for domain shift



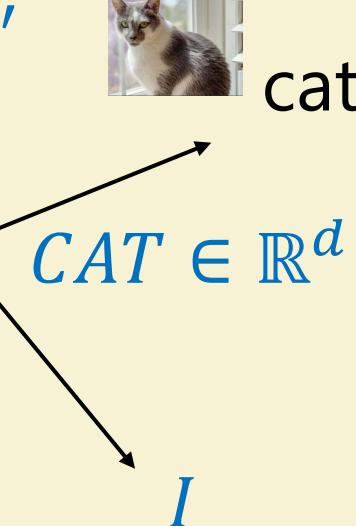
cat

$$Acc_D(C) = \beta \langle C, CAT + \alpha \cdot I \rangle = \beta \gamma$$

$$Acc_{D'}(C) = \beta' \langle C, CAT + \alpha \cdot I' \rangle = \beta' \gamma \cdot \frac{\|CAT\|^2}{\|CAT\|^2 + \alpha^2 \|I'\|^2}$$



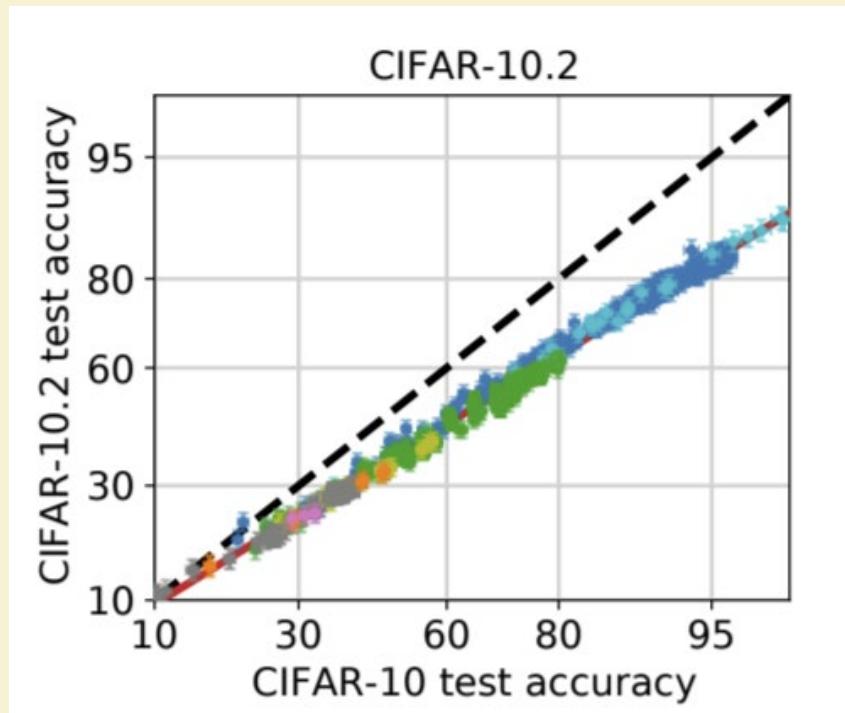
not cat



$$Acc_{D'}(C) = \frac{\beta'}{\beta(1 + \theta^2)} \cdot Acc_D(C)$$

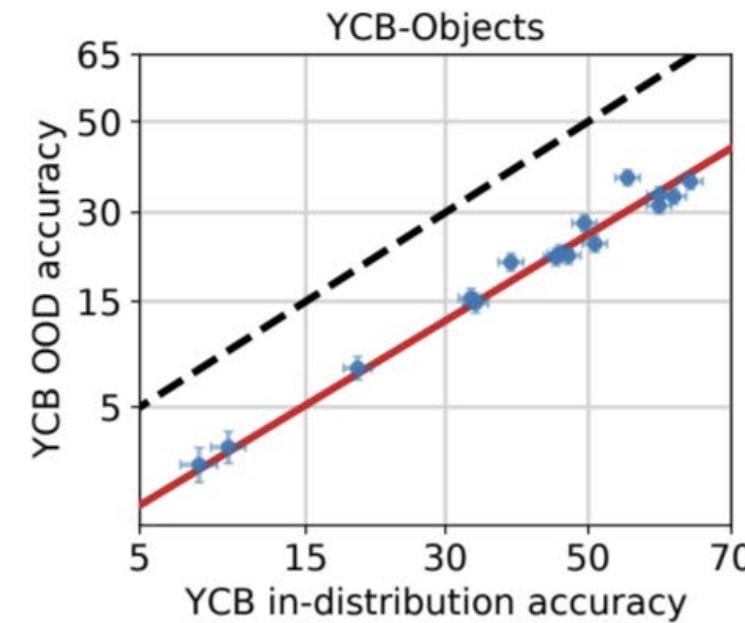
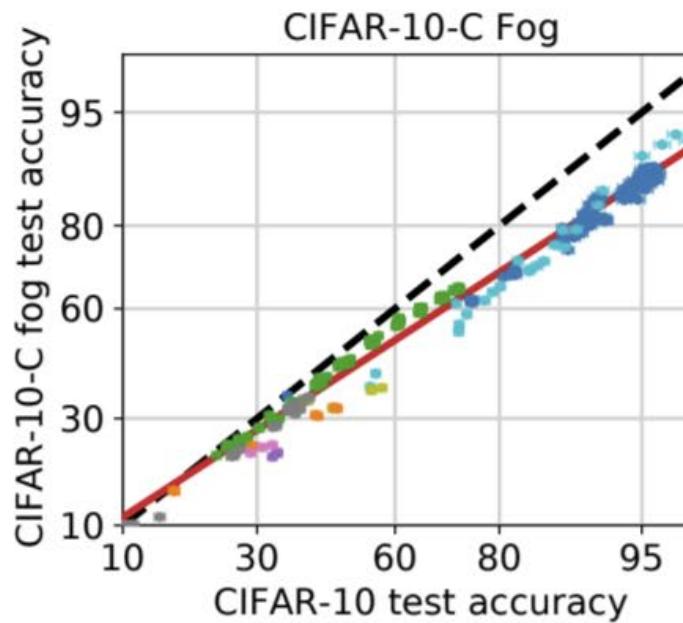
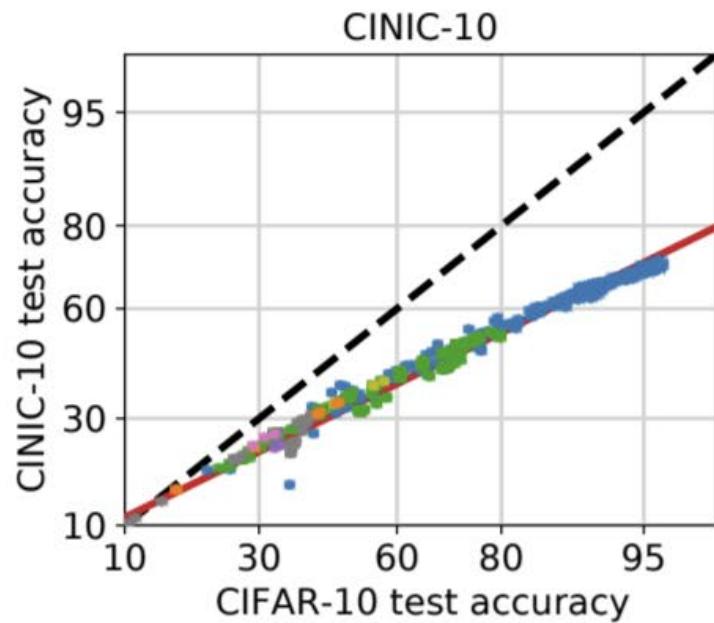
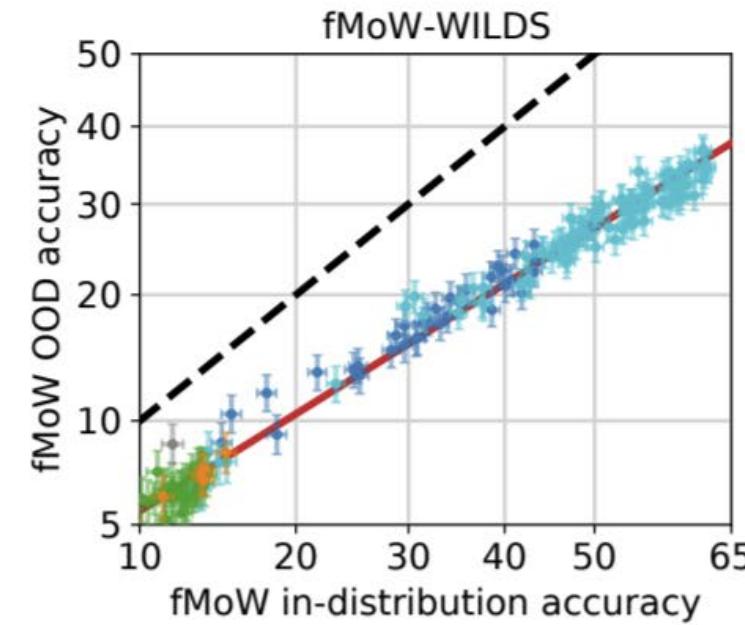
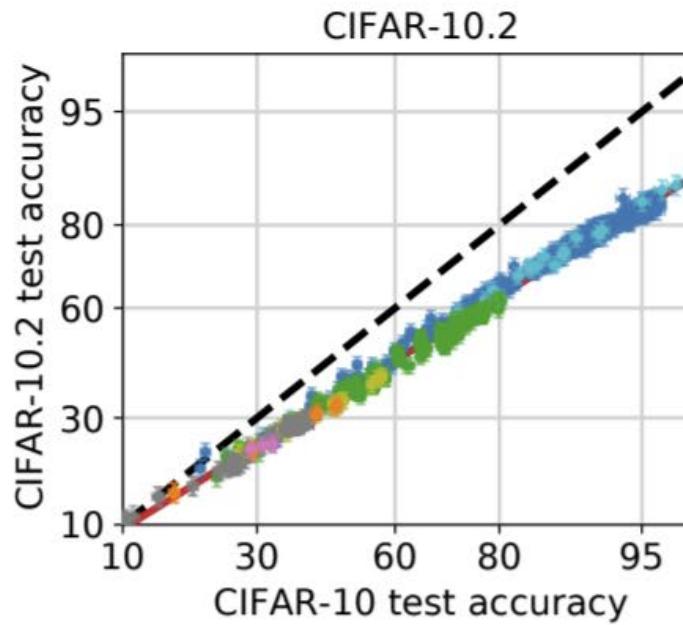
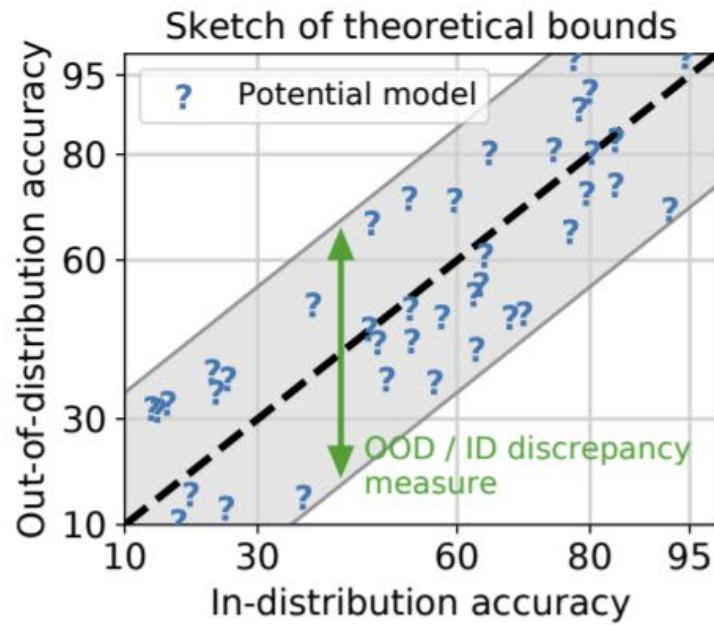
- $\beta'/\beta < 1$ iff D' harder than D
- θ^2 grows with idiosyncratic component of D

$$Acc_{D'}(C)$$



$$Acc_D(C)$$

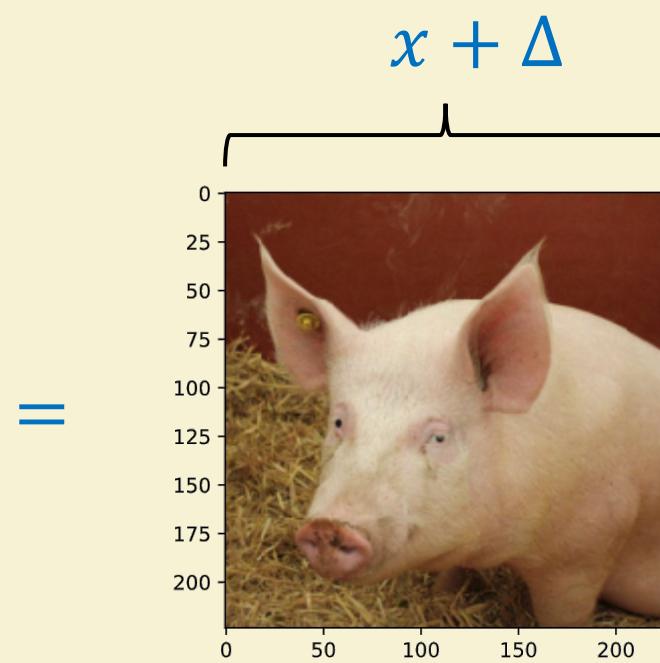
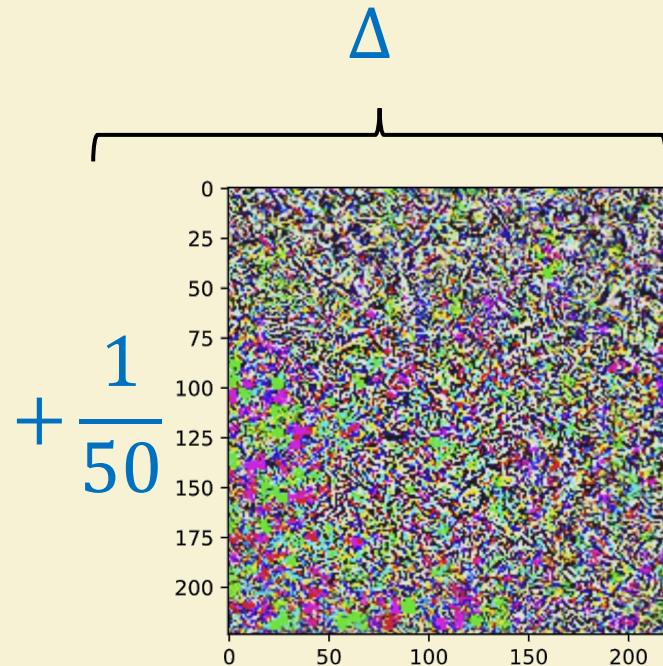
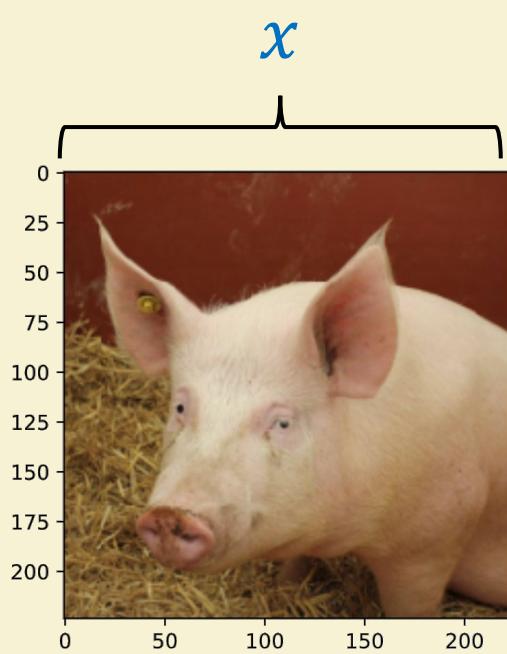
$$\text{so } \leq \frac{\beta'}{\beta}$$



— $y = x$	● Neural Network	● Random Features	● KNN	— $y = x$	● Linear Model
— Linear Fit	● ImageNet Pretrained Network	● Random Forest	● SVM	— Linear Fit	● AdaBoost

Adversarial perturbations

<https://adversarial-ml-tutorial.org/>



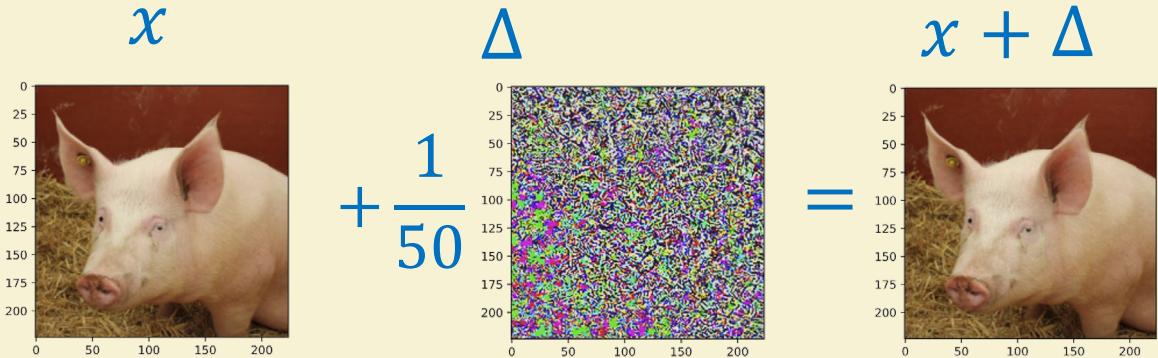
```
# 341 is the class index corresponding to "hog"
print(nn.CrossEntropyLoss()(model(norm(pig_tensor)),torch.LongTensor([341])).item())
0.0038814544677734375
```

$\Pr[x \text{ is hog}] \approx 99.6\%$

```
max_class = pred.max(dim=1)[1].item()
print("Predicted class: ", imagenet_classes[max_class])
print("Predicted probability:", nn.Softmax(dim=1)(pred)[0,max_class].item())
Predicted class: wombat
Predicted probability: 0.9997960925102234
```

$\Pr[x + \Delta \text{ is hog}] \approx 0.001\%$

Adversarial perturbations



$\Pr[x \text{ is hog}] \approx 99.6\%$

$\Pr[x + \Delta \text{ is hog}] \approx 0.001\%$

Should we be surprised? $|\Delta_i| \approx \frac{1}{64} |x_i| \quad \|\Delta\| \approx \frac{1}{64} \|x\|$

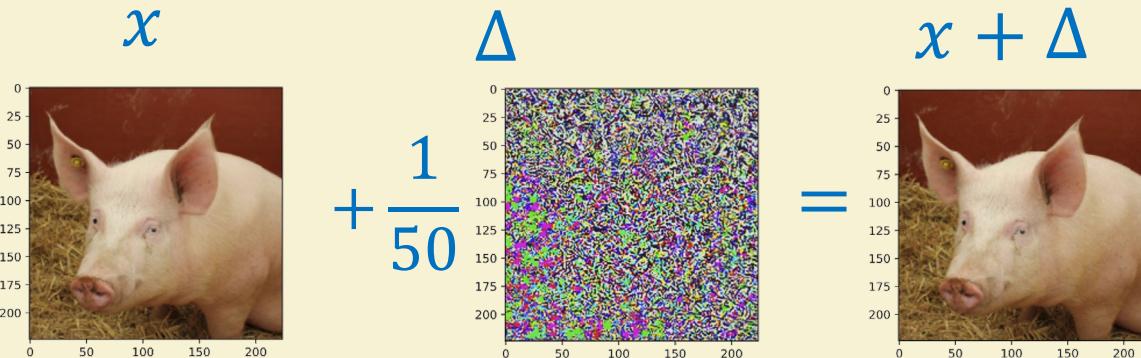
Resnet50 penultimate layer has $d = 2048$ params

$r(x) \in \mathbb{R}^d$, scale s.t. $\|x\| \approx \|r(x)\| \approx \sqrt{d}$

$$\|r(x + \Delta) - r(x)\| \approx L \|\Delta\|$$

L : Lipschitz "constant"

Adversarial perturbations



$\Pr[x \text{ is hog}] \approx 99.6\%$

$\Pr[x + \Delta \text{ is hog}] \approx 0.001\%$

Should we be surprised? $|\Delta_i| \approx \frac{1}{64} |x_i|$ $\|\Delta\| \approx \frac{1}{64} \|x\|$ $r(x) \in \mathbb{R}^{d=2048}$

$$\|r(x + \Delta) - r(x)\| \approx L\|\Delta\|$$

$$r(x) = C \cdot HOG + \sqrt{1 - c^2/d} N(0, I)$$

$$\Pr[x \text{ is not hog}] = \frac{\exp(-C)}{\exp(C) + \exp(-C)}$$

$$\langle r(x), HOG \rangle^2 \approx 3/d \approx 1/700$$

$$C \approx 3$$

HOG: unit vector s.t.
 $\Pr[x \text{ is hog}] \propto \langle HOG, r(x) \rangle$

Not surprising if
 $L \gg \frac{64}{25} \approx 2.5$

$$\|r(x)_{HOG}\| \approx \frac{1}{25} \|r(x)\|$$

Robust loss

Given

- set of transformation $\mathcal{T}: X \rightarrow X$
- loss function $\mathcal{L}: Y \times Y \rightarrow \mathbb{R}$
- classifier $f: X \rightarrow Y$

e.g. $t_\Delta(x) = x + \Delta$,
 $\mathcal{T} = \{t_\Delta : \|\Delta\|_\infty \leq \epsilon\}$

Robust loss of f at point (x, y) is

$$\mathcal{L}_{\mathcal{T}, x, y}(f) = \max_{t \in \mathcal{T}} \mathcal{L}(f(t(x)), y)$$

Robust training: Given $x_1, y_1, \dots, x_n, y_n$

$$f = \arg \min_{f \in \mathcal{F}} \sum \mathcal{L}_{\mathcal{T}, x_i, y_i}(f) = \arg \min_{f \in \mathcal{F}} \sum \max_{t \in \mathcal{T}} \mathcal{L}(f(t(x_i)), y_i)$$

Minimizing robust loss

Robust training: Given $x_1, y_1, \dots, x_n, y_n$

$$f = \arg \min_{f \in \mathcal{F}} \sum \mathcal{L}_{\mathcal{T}, x_i, y_i}(f) = \arg \min_{f \in \mathcal{F}} \sum \max_{t \in \mathcal{T}} \mathcal{L}(f(t(x_i)), y_i)$$

Goal: Find $\nabla_f \max_{t \in \mathcal{T}} \mathcal{L}(f(t(x_i)), y_i)$

Danskin's Theorem*: If $g(f, t)$ is nice (diff, continuous), \mathcal{T} compact then

$$\nabla_f \max_{t \in \mathcal{T}} g(f, t) = \nabla_f g(f, t^*(f))$$

where $t^*(f) = \arg \max_{t \in \mathcal{T}} g(f, t)$

To find gradient of outer,
enough to solve inner.

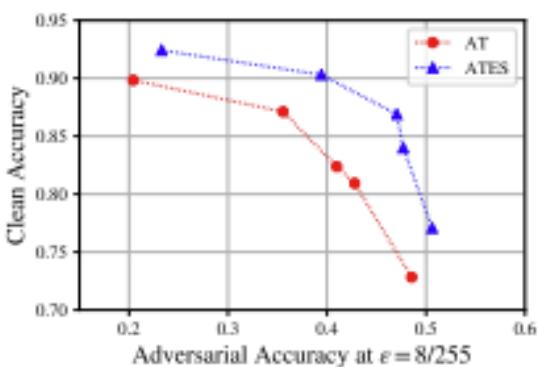
* Extends when $t^*(f)$ non unique though there is other fine print.
See Appendix A, Madry,Makelov,Schmidt,Tsipras, Vladu 2017

Clean vs Adversarial accuracy tradeoffs

Pareto curve – improving adversarial robustness decreases “clean” accuracy

CIFAR-10				
Defense	Natural	FGSM	PGD ²⁰	CW _∞
Standard	84.44	61.89	47.55	45.98
MMA	84.76	62.08	48.33	45.77
Dynamic	83.33	62.47	49.40	46.94
TRADES	82.90	62.82	50.25	48.29
MART	83.07	65.65	55.57	54.87

Wang , Zou, Yi, Bailey, Ma,Gu ‘20



Balaji, Goldstein, Hoffman’ 19
Sitawarin, Chakraborty, Wagner ’20

Robustness May Be at Odds with Accuracy

Dimitris Tsipras*
MIT
tsipras@mit.edu Shibani Santurkar*
MIT
sibani@mit.edu Logan Engstrom*
MIT
engstrom@mit.edu

Alexander Turner
MIT
turneram@mit.edu Aleksander Mądry
MIT
madry@mit.edu

Adversarial Examples Are Not Bugs, They Are Features

Andrew Ilyas*
MIT
aileyas@mit.edu Shibani Santurkar*
MIT
sibani@mit.edu Dimitris Tsipras*
MIT
tsipras@mit.edu

Logan Engstrom*
MIT
engstrom@mit.edu Brandon Tran
MIT
btran115@mit.edu Aleksander Mądry
MIT
madry@mit.edu

Theoretically Principled Trade-off between Robustness and Accuracy

Hongyang Zhang*
CMU & TTIC
hongyanz@cs.cmu.edu Yaodong Yu†
University of Virginia
yy8ms@virginia.edu Jiantao Jiao
UC Berkeley
jiantao@eecs.berkeley.edu

Eric P. Xing
CMU & Petuum Inc.
epxing@cs.cmu.edu Laurent El Ghaoui
UC Berkeley
elghaoui@berkeley.edu Michael I. Jordan
UC Berkeley
jordan@cs.berkeley.edu

<https://distill.pub/2019/advex-bugs-discussion/>

Inherent?

Puzzle: Augmentation vs Robustness

Robust loss:

$$\mathcal{L}_{\mathcal{T},x,y}(f) = \max_{t \in \mathcal{T}} \mathcal{L}(f(t(x)), y)$$

Training tends to **decrease** “clean” accuracy

Augmented loss:

$$\mathcal{L}_{\mathcal{T},x,y}(f) = \mathbb{E}_{t \sim \mathcal{T}} \mathcal{L}(f(t(x)), y)$$

Training tends to **increase** “clean” accuracy

