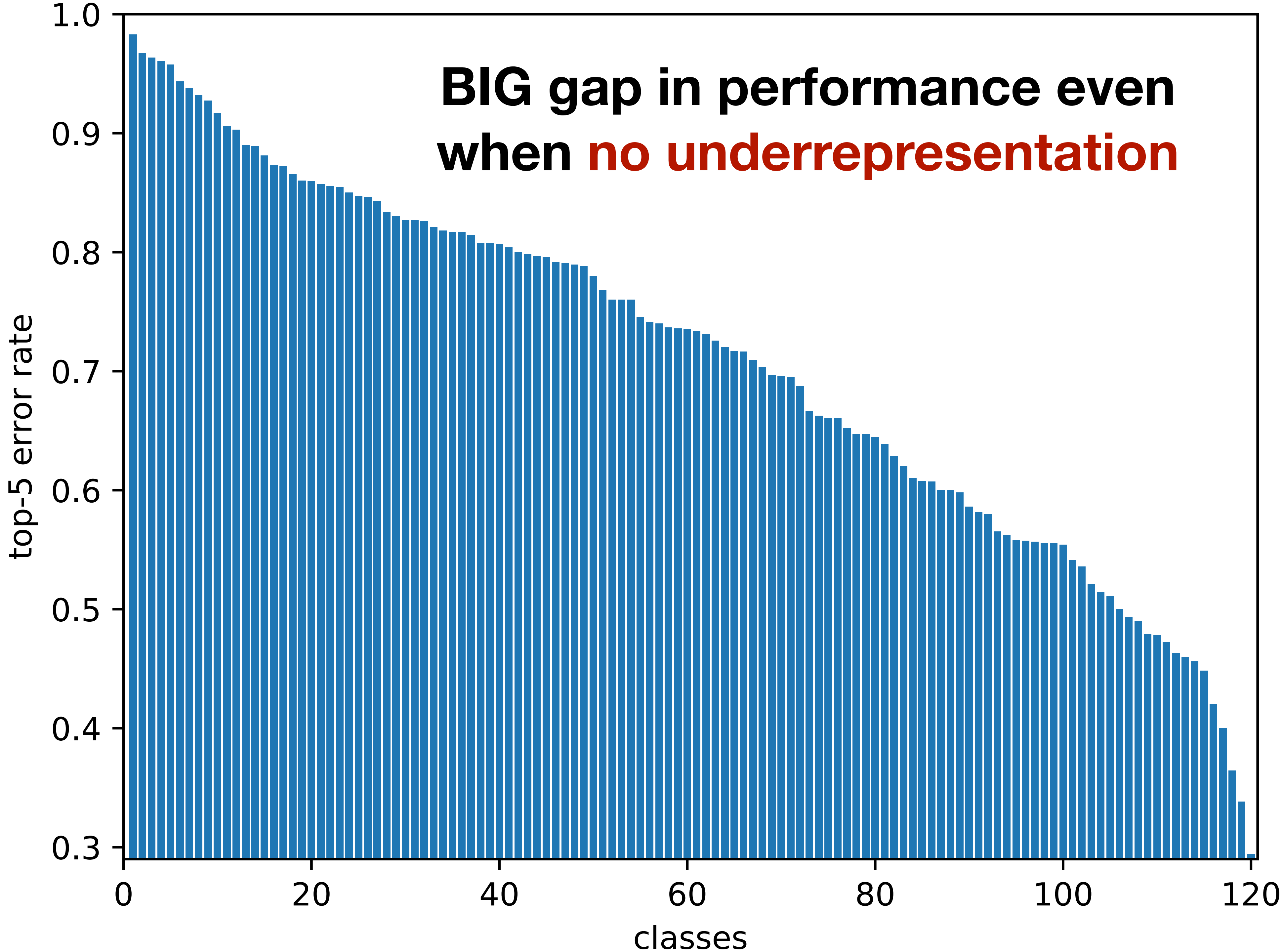
# Experiment 2: Fine-grained recognition

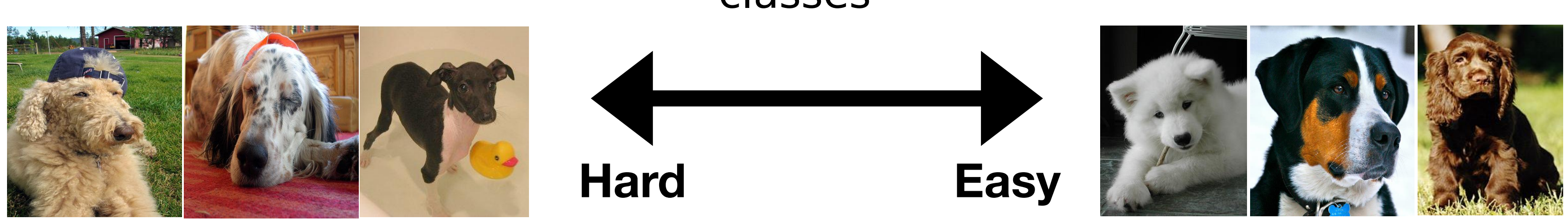
- Task: classify image of dog to breed (120 classes)
- Kernel features

Stanford Dogs Dataset [Khosla et al. '11]

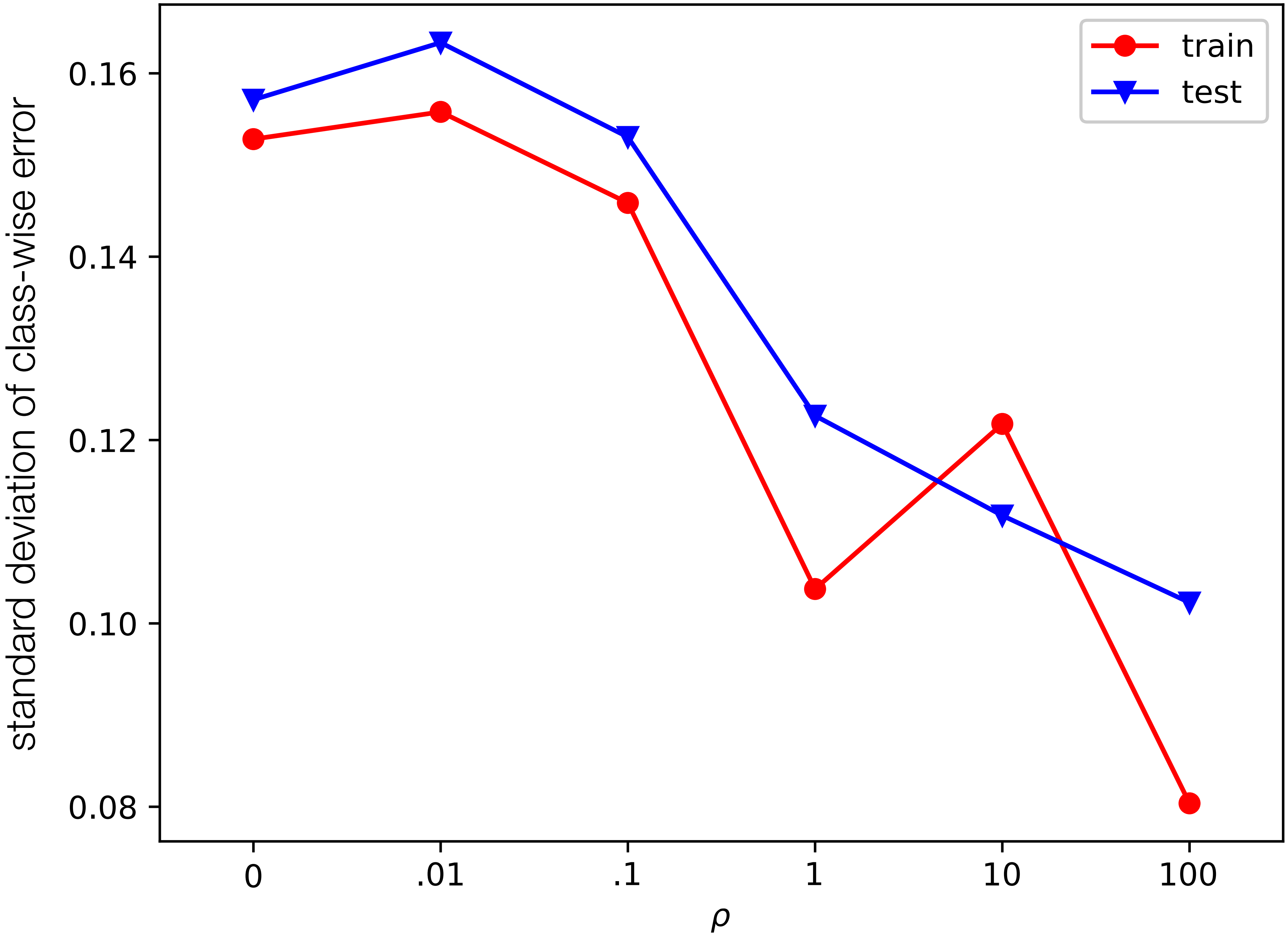
**No underrepresentation:**
**same number of images per class**

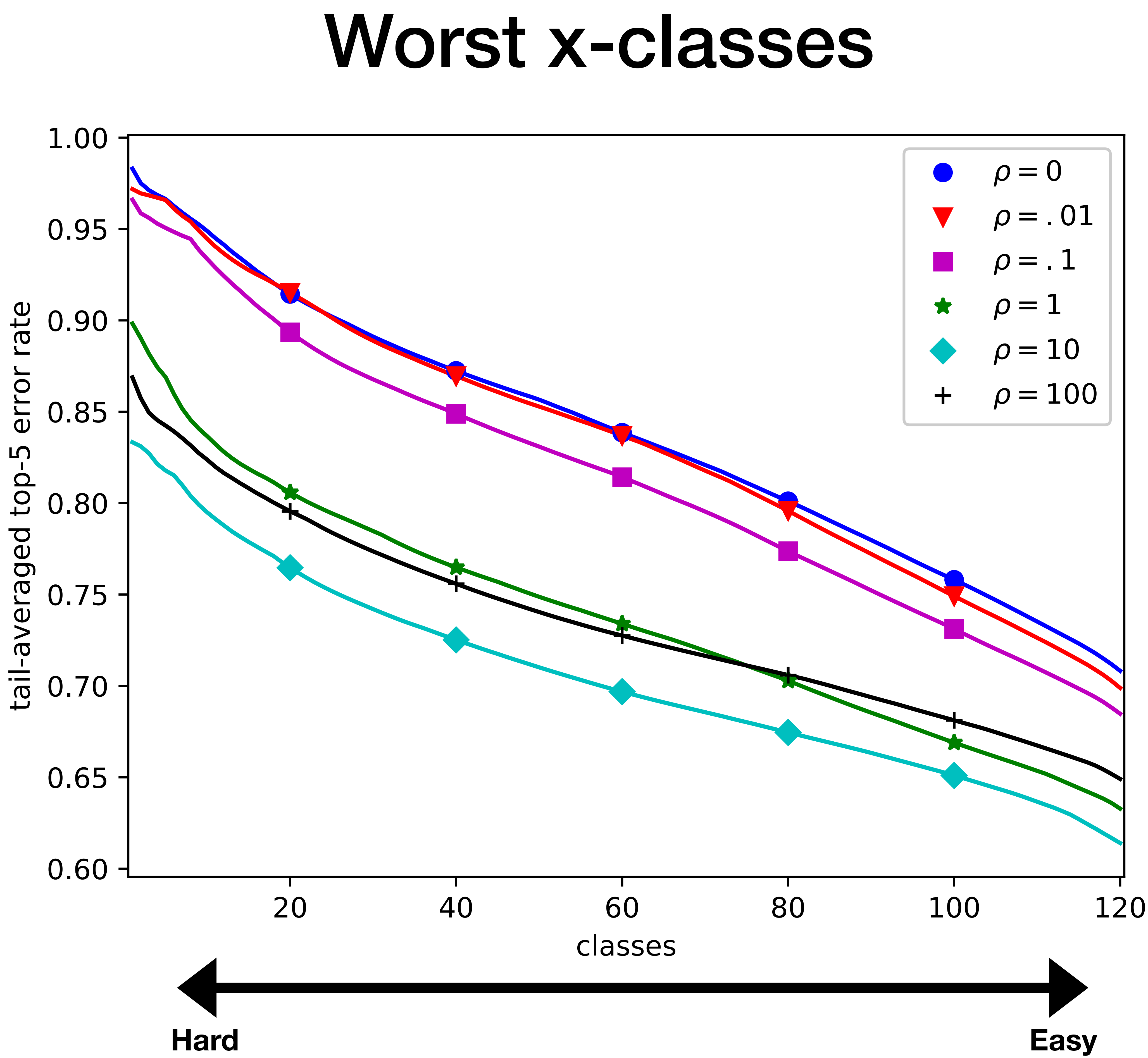
# ERM error rate



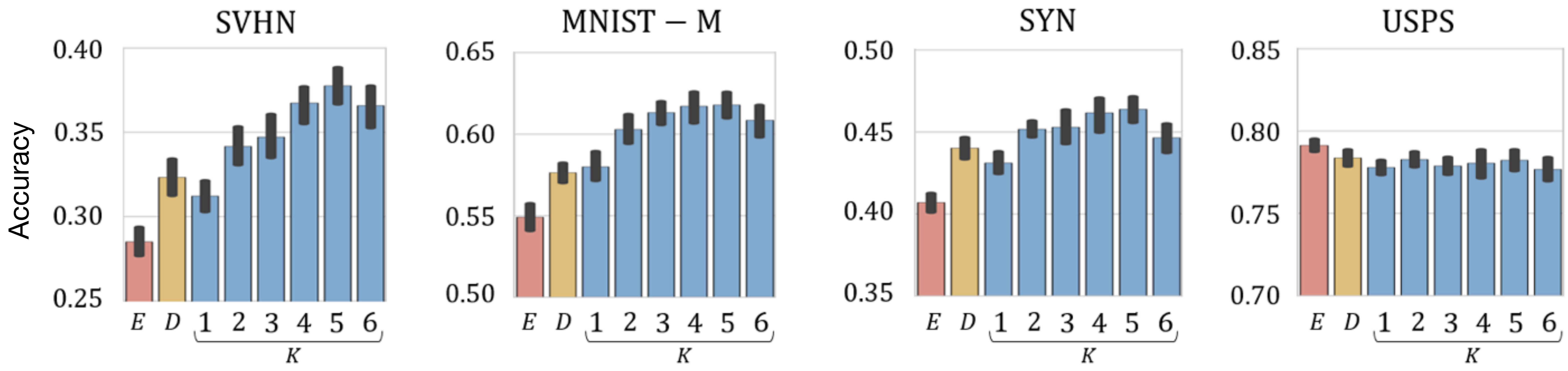
# Variation in error over 120 class





Worst x-classes
tail-averaged top-5 error rate
classes
$\rho = 0$
$\rho = .01$
$\rho = .1$
$\rho = 1$
$\rho = 10$
$\rho = 100$
Hard
Easy

# 30 seconds demo of Wasserstein DRO


$$\phi_\lambda(\theta; NN(x)) = \sup_{x'} \{ \ell(\theta; y, NN(x')) - \lambda \|NN(x) - NN(x')\|_2^2 \}$$

E = ERM with L2 regularization

D = Dropout regularization

K = number of Wasserstein DRO gradient ascent steps

Trained using lambda = 1.0, and an adaptive cost function defined on last hidden layer outputs of the neural network

# Distributional Robustness in Statistical Learning: A Few Vignettes

Hongseok Namkoong

June 2018

# Motivation

# Goal

We want machine-learned systems to
perform **reliably** when deployed in the real world

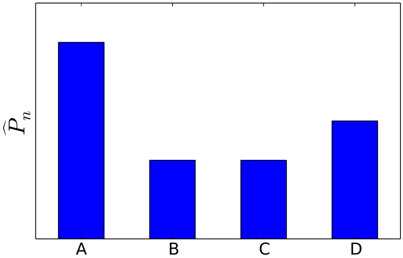
# Goal

We want machine-learned systems to
perform **reliably** when deployed in the real world

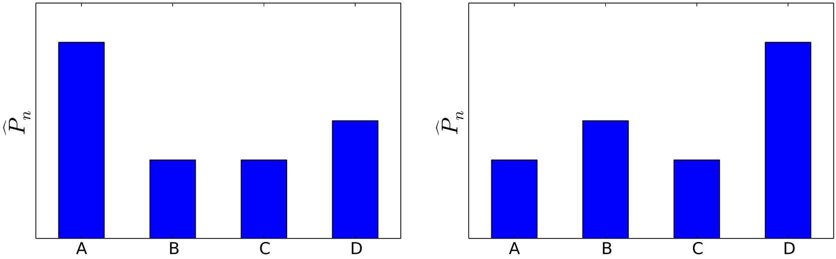
$\Rightarrow$ Uniformly good performance against distributional shifts

# Problem 0: Uncertainty in data

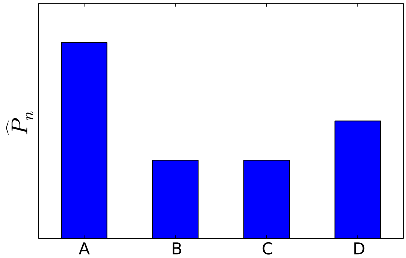
# Problem 0: Uncertainty in data

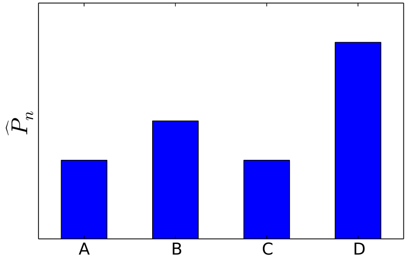


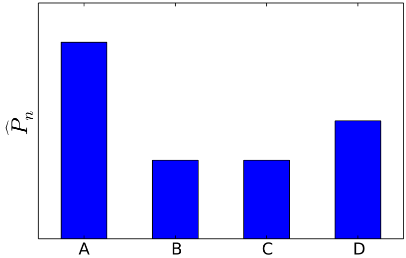
# Problem 0: Uncertainty in data



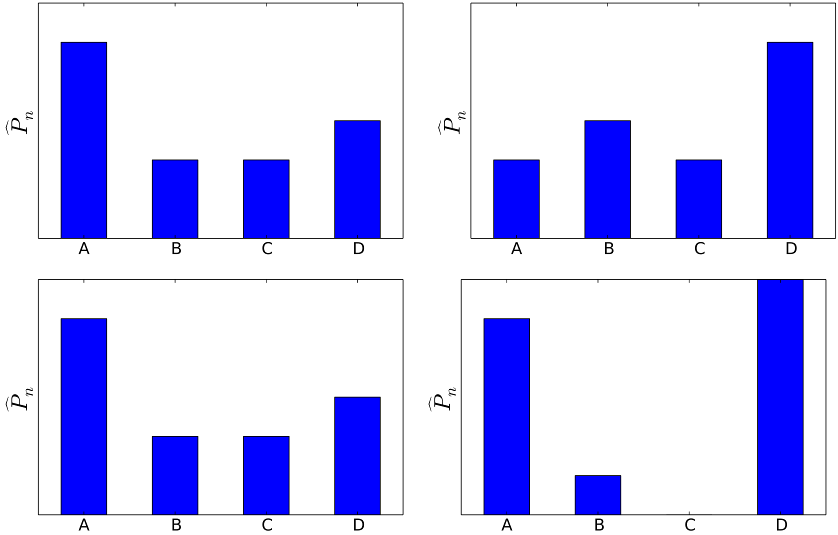
# Problem 0: Uncertainty in data



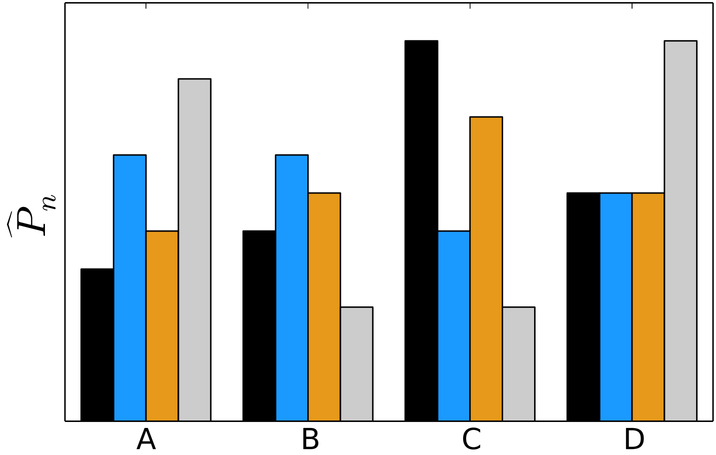



# Problem 0: Uncertainty in data



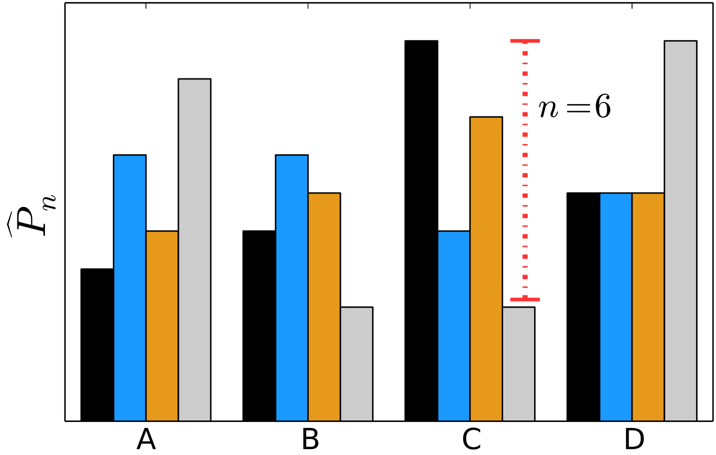
# Problem 0: Uncertainty in data

- Want to be robust to small perturbations in $\widehat{P}_n$

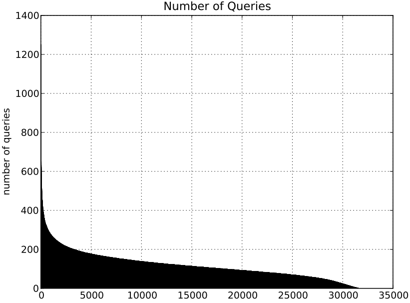


# Problem 0: Uncertainty in data

- Want to be robust to small perturbations in $\widehat{P}_n$



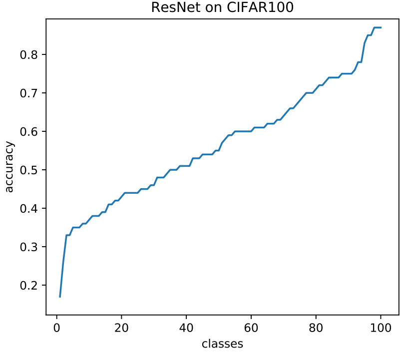
# Problem 1: Tail performance




MSR Learning to Rank

- Long-tailed data distribution
- At Google, a constant percentage of queries are new each day
- Rare queries determine quality of service

# Problem 1: Tail performance




class-wise test accuracy

- Same number of training examples for each class
- Average accuracy is around $60 - 70\%$
- Low performance on certain classes

# Problem 2: Changes in environment

Driving in California

# Problem 2: Changes in environment

Driving in California


Not driving in California

# Problem 3: Fairness

- Data collection almost always contains demographic, geographic, behavioral, temporal biases
- Pre-existing biases in language
  - Bias in word representations (word2vec) [Bolukbasi et al (2016)]
    man − woman ≈ computer programmer − homemaker

# Problem 3: Fairness

- Data collection almost always contains demographic, geographic, behavioral, temporal biases
- Pre-existing biases in language
  - Bias in word representations (word2vec) [Bolukbasi et al (2016)]
    man − woman ≈ computer programmer − homemaker
- Representation disparity for minority groups
  ⇒ disparate performance over different demographic groups
  - e.g. race, gender, age
- Speech recognition, facial recognition, automatic video captioning, language identification, academic recommender systems etc
  [Amodei et al (2016), Grother et al (2010), Hovy et al (2015), Blodgett et al (2016), Sapiezynski et al (2017), Tatman (2017)]

# Problem 3: Fairness

**Criminal Justice System**

- Predict if defendant should receive bail (crime recidivism)
- Higher false positive for African Americans

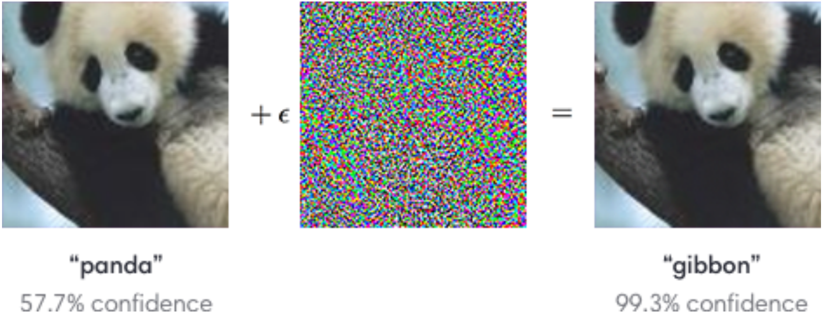
Table: ProPublica Analysis of COMPAS

| | Caucasian | African American |
|---|---|---|
| False High-Risk | 23.5% | 44.9% |
| False Low-Risk | 47.7% | 28.0% |

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

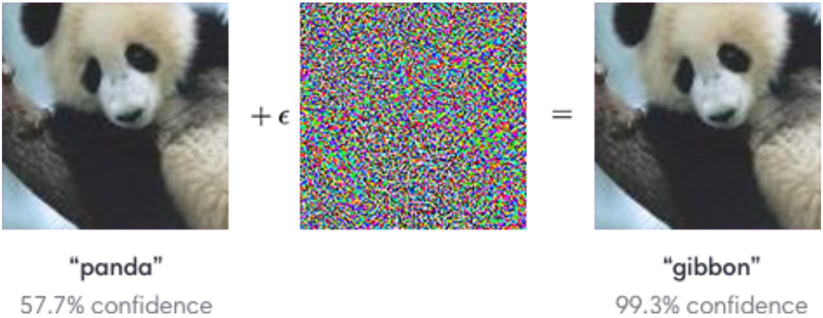
- More likely to wrongly deny African Americans bail!
- Used state-wide in New York, Wisconsin.

# Problem 4: Adversaries




[Goodfellow et al. 15]

# Problem 4: Adversaries




[Goodfellow et al. 15]

**Paraphrased Quote:**

We could put a transparent film on a stop sign, essentially imperceptible to a human, and a computer would see the stop sign as air (Dan Boneh)

# Risk-averseness

- Distributional Robustness = Risk-averseness (coherent risk measures) [Shapiro et al (2009)]
- *Risk-averse* decision making is standard in OR, economics, finance
- Optimizing *average-case* performance is still common in stats/ML
  - empirical risk minimization (ERM), maximum likelihood estimation

# Risk-averseness

- Distributional Robustness = Risk-averseness (coherent risk measures) [Shapiro et al (2009)]
- *Risk-averse* decision making is standard in OR, economics, finance
- Optimizing *average-case* performance is still common in stats/ML
  - empirical risk minimization (ERM), maximum likelihood estimation

Can we be risk-averse in statistics and machine learning?

# Small perturbations to data

# Stochastic optimization problems

Data $X$ and parameters $\theta$ to learn, with loss $\ell(\theta, X)$

Minimize the population expected loss

$$\underset{\theta \in \Theta}{\text{minimize}} \left\{ R(\theta) := \mathbb{E}_{P_0}[\ell(\theta, X)] = \int \ell(\theta, x) dP_0(x) \right\}$$

given an i.i.d. sample $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} P_0$

# Empirical Risk Minimization

**Standard approach:** Solve

# Empirical Risk Minimization

**Standard approach:** Solve

$$\widehat{\theta}^{\text{erm}} \in \operatorname*{argmin}_{\theta \in \Theta} \widehat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; X_i) \approx \mathbb{E}_{P_0}[\ell(\theta; X)].$$

**Goal:** Can we hedge against uncertainty in data?

# Empirical Risk Minimization

**Standard approach:** Solve

$$\widehat{\theta}^{\rm erm} \in \mathop{\rm argmin}_{\theta \in \Theta} \widehat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^n \ell(\theta; X_i) \underbrace{\approx \mathbb{E}_{P_0}[\ell(\theta; X)]}_{\text{Hopefully!}}.$$

**Goal:** Can we hedge against uncertainty in data?

# Point of departure: bias/variance tradeoff

# Point of departure: bias/variance tradeoff

- Any learning algorithm has *bias* (approximation error) and *variance* (estimation error)

# Point of departure: bias/variance tradeoff

- Any learning algorithm has *bias* (approximation error) and *variance* (estimation error)
- From empirical Bernstein's inequality, with probability $1 - \delta$

$$R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)] \le \underbrace{\widehat{R}_n(\theta)}_{\text{bias}} + \underbrace{\sqrt{\frac{2\text{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}}}_{\text{variance}} + \frac{C \log \frac{1}{\delta}}{n}$$

# Point of departure: bias/variance tradeoff

- Any learning algorithm has *bias* (approximation error) and *variance* (estimation error)
- From empirical Bernstein's inequality, with probability $1 - \delta$

$$R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)] \le \underbrace{\widehat{R}_n(\theta)}_{\text{bias}} + \underbrace{\sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}}}_{\text{variance}} + \frac{C \log \frac{1}{\delta}}{n}$$

- Can be made uniform in $\theta \in \Theta$ [Maurer & Pontil 09]

# Point of departure: bias/variance tradeoff

- Any learning algorithm has *bias* (approximation error) and *variance* (estimation error)
- From empirical Bernstein's inequality, with probability $1-\delta$

$$R(\theta) = \mathbb{E}_{P_0}[\ell(\theta;X)] \le \underbrace{\widehat{R}_n(\theta)}_{\text{bias}} + \underbrace{\sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}(\ell(\theta;X))}{n}}}_{\text{variance}} + \frac{C\log\frac{1}{\delta}}{n}$$

- Can be made uniform in $\theta \in \Theta$ [Maurer & Pontil 09]

**Goal:** Trade between these automatically and optimally by solving

$$\widehat{\theta}^{\mathrm{var}} \in \operatorname*{argmin}_{\theta\in\Theta}\left\{\widehat{R}_n(\theta) + \sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}(\ell(\theta;X))}{n}}\right\}.$$

# Optimizing for bias and variance

**Good idea:** Directly minimize bias + variance, certify optimality!

# Optimizing for bias and variance

**Good idea:** Directly minimize bias + variance, certify optimality!
Minor issue: variance is wildly non-convex

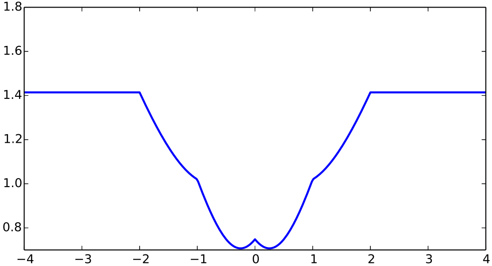



Figure: Variance of $|\theta - X|$

# Distributionally Robust Optimization

**Goal:**

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

# Distributionally Robust Optimization

**Goal:**

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

Solve empirical risk minimization problem

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; \sum_{i=1}^{n} \frac{1}{n} \ell(\theta; X_i)$$

# Distributionally Robust Optimization

**Goal:**

$$\underset{\theta \in \Theta}{\text{minimize}} \; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

Solve empirical risk minimization problem

$$\underset{\theta \in \Theta}{\text{minimize}} \; \sum_{i=1}^{n} \frac{1}{n} \ell(\theta; X_i)$$

# Distributionally Robust Optimization

**Goal:**

$$\operatorname*{minimize}_{\theta \in \Theta} \; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

Instead, solve *distributionally robust optimization (RO) problem*

$$\operatorname*{minimize}_{\theta \in \Theta} \; \sup_{p \in \mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

where $\mathcal{P}_{n,\rho}$ is some appropriately chosen set of vectors

# Distributionally Robust Optimization

**Goal:**

$$\operatorname*{minimize}_{\theta\in\Theta} \; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

Instead, solve *distributionally robust optimization (RO) problem*

$$\operatorname*{minimize}_{\theta\in\Theta} \; \sup_{p\in\mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

where $\mathcal{P}_{n,\rho}$ is some appropriately chosen set of vectors

[Scarf 58, Dupacová 87, Yue et al. 06, Popescu 07, Delage & Ye 10, Ben-Tal et al. 13, Bertsimas et al. 17, and many others]

# Empirical likelihood

**Idea:** Instead of using empirical distribution $\widehat{P}_n$ on sample $X_1, \ldots, X_n$, look at all distributions "near" it.

# Empirical likelihood

**Idea:** Instead of using empirical distribution $\widehat{P}_n$ on sample $X_1, \ldots, X_n$, look at all distributions "near" it.

- The $f$*-divergence* between distributions $P$ and $Q$ is

$$D_f\left(P \| Q\right) := \int f\left(\frac{dP}{dQ}\right) dQ$$

  where $f$ is some convex function with $f(1) = 0$.
  (w.l.o.g. can take $f'(1) = 0$ too)

# Empirical likelihood

**Idea:** Instead of using empirical distribution $\widehat{P}_n$ on sample $X_1, \dots, X_n$, look at all distributions "near" it.

- The $f$-*divergence* between distributions $P$ and $Q$ is

$$D_f\left(P \| Q\right) := \int f\left(\frac{dP}{dQ}\right) dQ$$

where $f$ is some convex function with $f(1) = 0$.
(w.l.o.g. can take $f'(1) = 0$ too)

- Measures of closeness we use: $f(t) = \frac{1}{2}(t-1)^2$

$$D_{\chi^2}\left(P \| Q\right) = \frac{1}{2} \sum_x \frac{(p(x) - q(x))^2}{q(x)} \qquad \text{Chi-square}$$

(Owen (1990): original empirical likelihood $f(t) = -\log t$)

# Empirical likelihood

$$E_n(\rho) := \left\{ \sum_{i=1}^{n} p_i Z_i : D_{\chi^2}\left(p \| \mathbf{1}/n\right) \le \frac{\rho}{n} \right\}$$

then *independently of distribution* on $Z \in \mathbb{R}^k$

$$\mathbb{P}(\mathbb{E}[Z] \in E_n(\rho)) \to \mathbb{P}(\chi_k^2 \le \rho).$$

ellipse [Owen 90, Baggerly 98, Newey and Smith 01, Imbens 02]

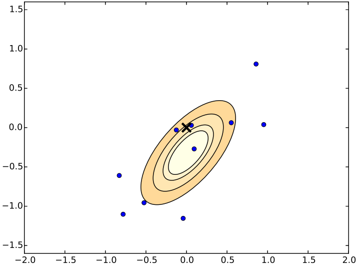
# Empirical likelihood

$$E_n(\rho) := \left\{ \sum_{i=1}^{n} p_i Z_i : D_{\chi^2}\left(p \| \mathbf{1}/n\right) \le \frac{\rho}{n} \right\}$$

then *independently of distribution* on $Z \in \mathbb{R}^k$

$$\mathbb{P}(\mathbb{E}[Z] \in E_n(\rho)) \to \mathbb{P}(\chi_k^2 \le \rho).$$

ellipse [Owen 90, Baggerly 98, Newey and Smith 01, Imbens 02]



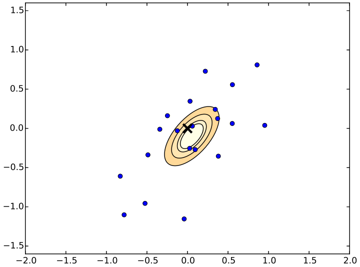
# Empirical likelihood

$$E_n(\rho) := \left\{ \sum_{i=1}^{n} p_i Z_i : D_{\chi^2}\left(p \| \mathbf{1}/n\right) \le \frac{\rho}{n} \right\}$$

then *independently of distribution* on $Z \in \mathbb{R}^k$

$$\mathbb{P}(\mathbb{E}[Z] \in E_n(\rho)) \to \mathbb{P}(\chi^2_k \le \rho).$$

ellipse [Owen 90, Baggerly 98, Newey and Smith 01, Imbens 02]



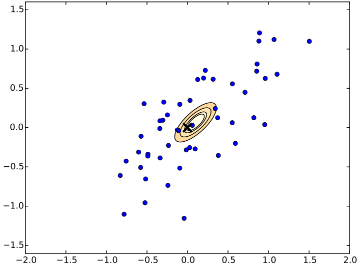
# Empirical likelihood

$$E_n(\rho) := \left\{ \sum_{i=1}^n p_i Z_i : D_{\chi^2}\left(p \| \mathbf{1}/n\right) \le \frac{\rho}{n} \right\}$$

then *independently of distribution* on $Z \in \mathbb{R}^k$

$$\mathbb{P}(\mathbb{E}[Z] \in E_n(\rho)) \to \mathbb{P}(\chi_k^2 \le \rho).$$

ellipse [Owen 90, Baggerly 98, Newey and Smith 01, Imbens 02]



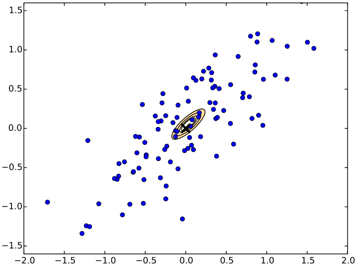
# Empirical likelihood

$$E_n(\rho) := \left\{ \sum_{i=1}^{n} p_i Z_i : D_{\chi^2}\left(p \| \mathbf{1}/n\right) \le \frac{\rho}{n} \right\}$$

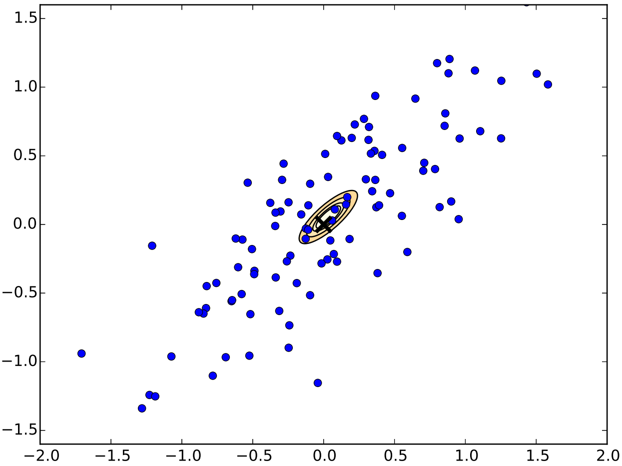
then *independently of distribution* on $Z \in \mathbb{R}^k$

$$\mathbb{P}(\mathbb{E}[Z] \in E_n(\rho)) \to \mathbb{P}(\chi^2_k \le \rho).$$

ellipse [Owen 90, Baggerly 98, Newey and Smith 01, Imbens 02]



# Empirical likelihood




**Idea:** Leverage this in stochastic optimization

# Robust Optimization

**Idea:** Optimize over *uncertainty set* of possible distributions,

$$\mathcal{P}_{n,\rho} := \left\{ \text{Distributions } P \text{ such that } D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n} \right\}$$

for some $\rho > 0$.

# Robust Optimization

**Idea:** Optimize over *uncertainty set* of possible distributions,

$$\mathcal{P}_{n,\rho} := \left\{ \text{Distributions } P \text{ such that } D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n} \right\}$$

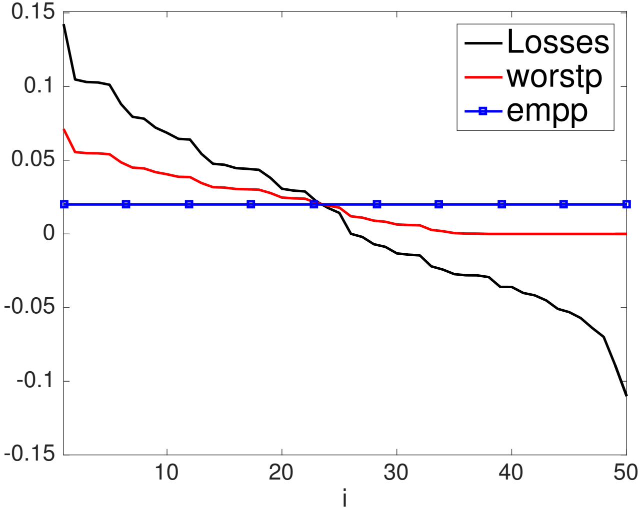
for some $\rho > 0$.

Define (and optimize) *empirical likelihood upper confidence bound*

$$R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P: D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] = \max_{p: D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

[Ben-Tal et al. 13, Bertsimas et al. 16, Lam & Zhou 16]

# Visualization of worst-case



# Optimization

Solve

$$\widehat{\theta}^{\text{rob}} := \operatorname*{argmin}_{\theta \in \Theta} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] \right\}.$$

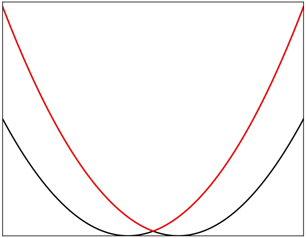
# Optimization

Solve

$$\widehat{\theta}^{\text{rob}} := \operatorname*{argmin}_{\theta \in \Theta} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] \right\}.$$

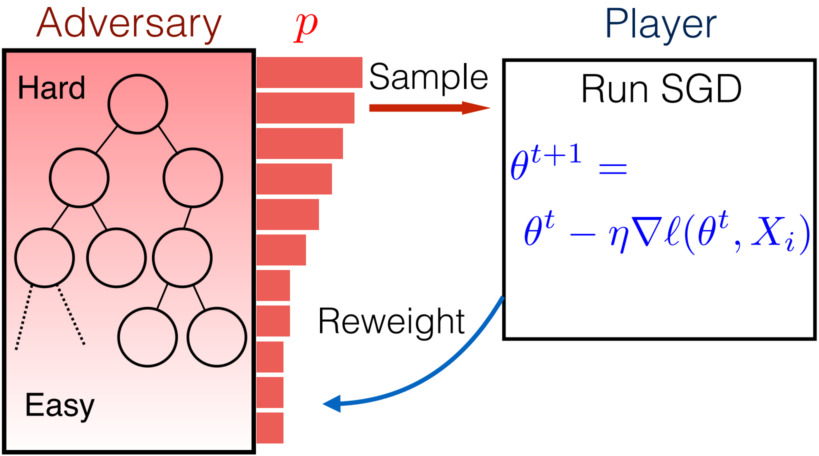
**Nice properties:**

- Convex optimization problem.
- Solve dual reformulation using interior point methods [Ben-Tal et al. 13]
- For large $n$ and $d$, efficient solution methods as fast as stochastic gradient descent [N. & Duchi, 16] 1



Play a **two-player stochastic game** [N. & Duchi 16]

$$\min_{\theta \in \Theta} \max_{p \in \mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$



# Robust Optimization $\approx$ Variance Regularization

Theorem (Duchi, Glynn & N. 2016)

*For general $f$-divergences,*

$$R_n(\theta; \mathcal{P}_{n,\rho}) = \widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}} + \mathit{Rem}_n(\theta).$$

- *If $\sigma^2(\theta) < \infty$, then $\sqrt{n}\mathit{Rem}_n(\theta) \stackrel{P^*}{\to} 0$*
- *If $\{\ell(\theta; \cdot) : \theta \in \Theta\}$ is $P_0$-Donsker, then $\sqrt{n} \sup_{\theta \in \Theta} \mathit{Rem}_n(\theta) \stackrel{P^*}{\to} 0$*

# Robust Optimization $\approx$ Variance Regularization

Theorem (Duchi, Glynn & N. 2016)

*For general $f$-divergences,*

$$R_n(\theta; \mathcal{P}_{n,\rho}) = \widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}} + \mathit{Rem}_n(\theta).$$

- *If $\sigma^2(\theta) < \infty$, then $\sqrt{n}\mathit{Rem}_n(\theta) \stackrel{P^*}{\to} 0$*
- *If $\{\ell(\theta; \cdot) : \theta \in \Theta\}$ is $P_0$-Donsker, then $\sqrt{n} \sup_{\theta \in \Theta} \mathit{Rem}_n(\theta) \stackrel{P^*}{\to} 0$*
- *[Lam (2013), Gotoh et al (2015), Lam and Zhao (2017)]*

# Robust Optimization $\approx$ Variance Regularization

Theorem (Duchi & N. 2016)

*Assume that* $\ell(\theta; X) \le M$. *Let* $\sigma^2(\theta) := \mathrm{Var}(\ell(\theta; X))$.

$$R_n(\theta; \mathcal{P}_{n,\rho}) = \widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}} + \mathit{Rem}_n(\theta).$$

- $\mathit{Rem}_n(\theta) \le \frac{\sqrt{12}\rho M}{n}$
- $\mathit{Rem}_n(\theta) = 0$ *with probability at least* $1 - \exp(-\frac{n\sigma^2(\theta)}{36M^2})$ proof

# Robust Optimization $\approx$ Variance Regularization

Theorem (Duchi & N. 2016)

*Assume that* $\ell(\theta; X) \le M$. *Let* $\sigma^2(\theta) := \mathrm{Var}(\ell(\theta; X))$.

$$R_n(\theta; \mathcal{P}_{n,\rho}) = \widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}} + \mathit{Rem}_n(\theta).$$

- $\mathit{Rem}_n(\theta) \le \frac{\sqrt{12}\rho M}{n}$
- $\mathit{Rem}_n(\theta) = 0$ *with probability at least* $1 - \exp(-\frac{n\sigma^2(\theta)}{36M^2})$ proof
- *Let* $N(\mathcal{F}, \tau, \|\cdot\|_{L^\infty})$ *be the* $\tau$*-covering number with respect to the supremum norm.*

$$\mathbb{P}\left(\mathit{Rem}_n(\theta) = 0 \text{ for all } \theta \in \Theta \text{ s.t. } \sigma^2(\theta) \ge \tau^2\right)$$
$$\ge 1 - cN(\mathcal{F}, \tau, \|\cdot\|_{L^\infty}) \exp(-\frac{n\tau^2}{M^2}).$$

# Robust Optimization ≈ Variance Regularization

With high probability,

$$\underbrace{R_n(\theta; \mathcal{P}_{n,\rho})}_{\text{Robust}} = \underbrace{\widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}}}_{\text{VarReg}}$$

# Robust Optimization ≈ Variance Regularization

With high probability,

$$\underbrace{R_n(\theta; \mathcal{P}_{n,\rho})}_{\text{Robust}} = \underbrace{\widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}}}_{\text{VarReg}}$$

- Robust is empirical likelihood UCB and VarReg is normal UCB

# Robust Optimization $\approx$ Variance Regularization

With high probability,

$$\underbrace{R_n(\theta; \mathcal{P}_{n,\rho})}_{\text{Robust}} = \underbrace{\widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}}}_{\text{VarReg}}$$

- Robust is empirical likelihood UCB and VarReg is normal UCB
- Robust is convex, VarReg is non-convex

# Robust Optimization ≈ Variance Regularization

With high probability,

$$\underbrace{R_n(\theta; \mathcal{P}_{n,\rho})}_{\text{Robust}} = \underbrace{\widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}}}_{\text{VarReg}}$$

- Robust is empirical likelihood UCB and VarReg is normal UCB
- Robust is convex, VarReg is non-convex
- Robust **only** penalizes upward (bad) deviations in the loss whereas VarReg penalizes downward (good) deviations along with the upward (bad) deviations

# Robust Optimization ≈ Variance Regularization

With high probability,

$$\underbrace{R_n(\theta; \mathcal{P}_{n,\rho})}_{\text{Robust}} = \underbrace{\widehat{R}_n(\theta) + \sqrt{\frac{2\rho \text{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}}}_{\text{VarReg}}$$

- Robust is empirical likelihood UCB and VarReg is normal UCB
- Robust is convex, VarReg is non-convex
- Robust **only** penalizes upward (bad) deviations in the loss whereas VarReg penalizes downward (good) deviations along with the upward (bad) deviations
- Robust is a coherent risk measure (i.e. it is a sensible negative utility)

# Empirical likehood for stochastic optimization

Solve

$$\widehat{\theta}^{\rm rob} := \mathop{\rm argmin}_{\theta \in \Theta} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] \right\}.$$

# Empirical likehood for stochastic optimization

Solve

$$\widehat{\theta}^{\rm rob} := \mathop{\rm argmin}_{\theta\in\Theta} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] \right\}.$$

Assume that $\{\ell(\theta; \cdot) : \theta \in \Theta\}$ is $P_0$-Donsker

e.g. $\Theta \subset \mathbb{R}^d$ compact and $\ell(\cdot; X)$ is $M(X)$-Lipschitz with $\mathbb{E}M(X)^2 < \infty$.

# Empirical likehood for stochastic optimization

Solve

$$\widehat{\theta}^{\text{rob}} := \mathop{\text{argmin}}_{\theta \in \Theta} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] \right\}.$$

Assume that $\{\ell(\theta; \cdot) : \theta \in \Theta\}$ is $P_0$-Donsker

e.g. $\Theta \subset \mathbb{R}^d$ compact and $\ell(\cdot; X)$ is $M(X)$-Lipschitz with $\mathbb{E}M(X)^2 < \infty$.

Theorem (Duchi, Glynn & N. 16 [1])

*If* $\theta^\star := \text{argmin}_{\theta \in \Theta} R(\theta)$ *is unique, then*

$$\lim_{n \to \infty} \mathbb{P}\left( \inf_{\theta \in \Theta} R(\theta) \le R_n(\widehat{\theta}^{\text{rob}}, \mathcal{P}_{n,\rho}) \right) = \mathbb{P}\left( N(0,1) \ge -\sqrt{2\rho} \right).$$

Can be extended to Harris recurrent Markov chains that mix suitably fast

# Optimal bias variance tradeoff

Solve

$$\widehat{\theta}^{\mathrm{rob}} := \operatorname*{argmin}_{\theta \in \Theta} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] \right\}.$$

# Optimal bias variance tradeoff

Solve

$$\widehat{\theta}^{\mathrm{rob}} := \operatorname*{argmin}_{\theta \in \Theta} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] \right\}.$$

Let $\ell(\cdot; X)$ is $M$-Lipschitz and $\mathrm{diam}(\Theta) = r$

# Optimal bias variance tradeoff

Solve

$$\widehat{\theta}^{\text{rob}} := \operatorname*{argmin}_{\theta\in\Theta} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P: D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] \right\}.$$

Let $\ell(\cdot; X)$ is $M$-Lipschitz and $\text{diam}(\Theta) = r$

Theorem (Duchi & N. 2016)

*Let $\rho = \log\frac{1}{\delta} + d\log n$. Then with probability at least $1-\delta$,*

$$R(\widehat{\theta}^{\text{rob}}) \le \underbrace{R_n(\widehat{\theta}^{\text{rob}}, \mathcal{P}_{n,\rho})}_{\text{optimality certificate}} + \frac{crM}{n}\rho$$

$$\le \underbrace{\min_{\theta\in\Theta} \left\{ R(\theta) + 2\sqrt{\frac{2\rho \text{Var}(\ell(\theta, \xi))}{n}} \right\}}_{\text{optimal tradeoff}} + \frac{crM}{n}\rho$$

# Fast rates from optimal tradeoff

- Let $\rho \approx \mathfrak{Comp}_n(\Theta)$. If $\ell(\theta; X) \in [0, M]$, then with high prob,

$$R(\widehat{\theta}^{\mathrm{rob}}) \le \underbrace{\min_{\theta \in \Theta} \left\{ R(\theta) + 2\sqrt{\frac{2\rho \mathrm{Var}(\ell(\theta, \xi))}{n}} \right\}}_{\text{optimal tradeoff}} + \frac{CM\rho}{n}$$

# Fast rates from optimal tradeoff

- Let $\rho \approx \mathfrak{Comp}_n(\Theta)$. If $\ell(\theta; X) \in [0, M]$, then with high prob,

$$R(\widehat{\theta}^{\rm rob}) \le \underbrace{\min_{\theta \in \Theta} \left\{ R(\theta) + 2\sqrt{\frac{2\rho \mathrm{Var}(\ell(\theta, \xi))}{n}} \right\}}_{\text{optimal tradeoff}} + \frac{CM\rho}{n}$$

- ERM: For $R(\theta^\star) = \inf_{\theta \in \Theta} R(\theta)$, with high probability,

$$R(\widehat{\theta}^{\rm erm}) \le R(\theta^\star) + \sqrt{\frac{2\rho M R(\theta^\star)}{n}} + \frac{CM\rho}{n}$$

# Fast rates from optimal tradeoff

- Let $\rho \approx \mathfrak{Comp}_n(\Theta)$. If $\ell(\theta; X) \in [0, M]$, then with high prob,

$$R(\widehat{\theta}^{\text{rob}}) \le \underbrace{\min_{\theta \in \Theta} \left\{ R(\theta) + 2\sqrt{\frac{2\rho \text{Var}(\ell(\theta, \xi))}{n}} \right\}}_{\text{optimal tradeoff}} + \frac{CM\rho}{n}$$

- ERM: For $R(\theta^\star) = \inf_{\theta \in \Theta} R(\theta)$, with high probability,

$$R(\widehat{\theta}^{\text{erm}}) \le R(\theta^\star) + \sqrt{\frac{2\rho M R(\theta^\star)}{n}} + \frac{CM\rho}{n}$$

- If $\text{Var}(\ell(\theta^\star; X)) \ll M R(\theta^\star)$, first bound is **tighter**
  - See paper for an **explicit example** where

$$R(\widehat{\theta}^{\text{rob}}) \le R(\theta^\star) + \frac{C_1}{n} \quad \text{but} \quad R(\widehat{\theta}^{\text{erm}}) \ge R(\theta^\star) + \frac{C_2}{\sqrt{n}}$$

# Experiment: Coverage Rates

- Portfolio optimization $\ell(\theta; X) = \theta^\top X$
- Conditional Value-at-Risk $\ell(\theta; X) = \frac{1}{1-\alpha}(X - \theta)_+ + \theta$
- Newsvendor problem $\ell(\theta; X) = b^\top (\theta - X)_+ + s^\top (X - \theta)_+$.

# Experiment: Coverage Rates

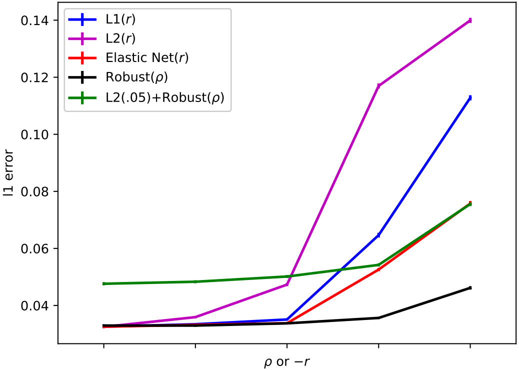
- Portfolio optimization $\ell(\theta; X) = \theta^\top X$
- Conditional Value-at-Risk $\ell(\theta; X) = \frac{1}{1-\alpha}(X - \theta)_+ + \theta$
- Newsvendor problem $\ell(\theta; X) = b^\top(\theta - X)_+ + s^\top(X - \theta)_+$.

Figure: Coverage Rates (nominal = 95%)

| % | Portfolio | | CVaR | | Newsvendor | |
|---|---|---|---|---|---|---|
| sample size | EL | Normal | EL | Normal | EL | Normal |
| 20 | 75.16 | 89.2 | 30.1 | 91.38 | 91.78 | 95.02 |
| 200 | 92.96 | 93.68 | 86.73 | 95.27 | 94.64 | 95.26 |
| 2000 | 95.48 | 95.25 | 93.73 | 95.25 | 94.92 | 95.04 |
| 10000 | 96.43 | 95.51 | 94.71 | 94.85 | 94.43 | 94.43 |

# Experiment: Regression

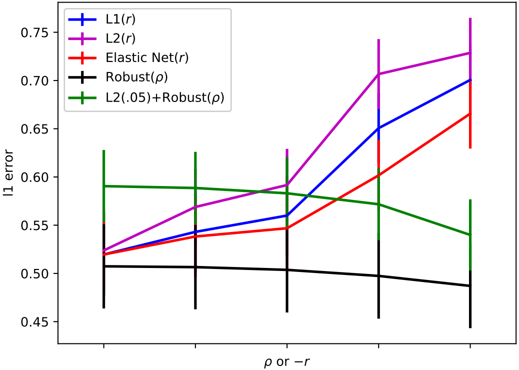
**Problem:** Predict crime rate $Y$, given feature vector describing community




Median test loss $\ell(\theta; (W, Y)) = |\theta^\top W - Y|$

# Experiment: Regression

**Problem:** Predict crime rate, given feature vector on community




Maximal test loss $\ell(\theta; (X, Y)) = |\theta^\top X - Y|$

# Experiment: Reuters Corpus (multi-label)

**Problem:** Classify documents as a **subset** of the 4 categories:

$$\left\{\text{Corporate, Economics, Government, Markets}\right\}$$

- Data: pairs $x \in \mathbb{R}^d$ represents document, $y \in \{-1, 1\}^4$ where $y_j = 1$ indicating $x$ belongs $j$-th category.
- Logistic loss, with $\Theta = \left\{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1000\right\}$
- $d = 47,236$, $n = 804,414$. 10-fold cross-validation.
- Use precision and recall to evaluate performance

$$\text{Precision} = \frac{\#\ \text{Correct}}{\#\ \text{Guessed Positive}} \qquad \text{Recall} = \frac{\#\ \text{Correct}}{\#\ \text{Actually Positive}}$$

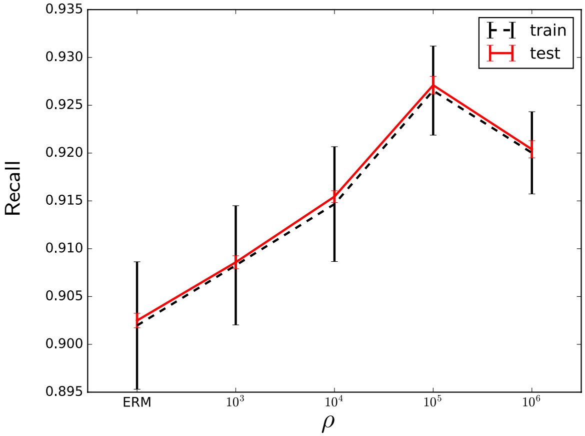
# Experiment: Reuters Corpus (multi-label)

Table: Reuters Number of Examples

| Corporate | Economics | Government | Markets |
|---|---|---|---|
| 381,327 | 119,920 | 239,267 | 204,820 |

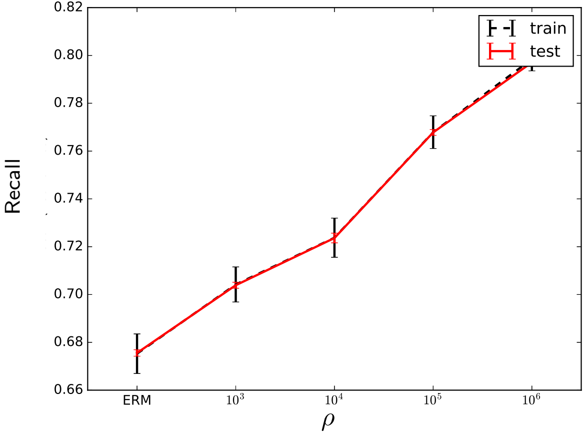
# Experiment: Reuters Corpus (multi-label)

Figure: Recall on common category (Corporate)



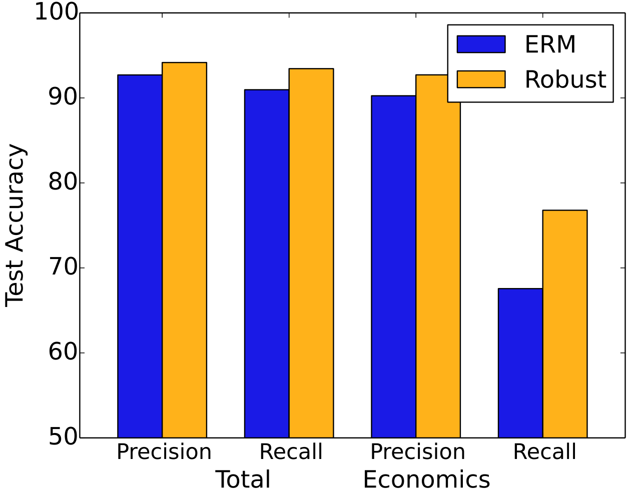
# Experiment: Reuters Corpus (multi-label)

Figure: Recall on rare category (Economics)



# Experiment: Reuters Corpus (multi-label)

Do well **almost all** the time intead of just on average!



# Perturbations to population distribution

# Distributionally robust optimization

**Idea:** Replace data-generating distribution $P_0$ with "uncertainty" set $\mathcal{P}$ of possible distributions around $P_0$

$$\operatorname*{minimize}_{\theta \in \Theta} \; \mathbb{E}_{P_0}[\ell(\theta, X)]$$

# Distributionally robust optimization

**Idea:** Replace data-generating distribution $P_0$ with "uncertainty" set $\mathcal{P}$ of possible distributions around $P_0$

$$\mathop{\text{minimize}}_{\theta \in \Theta} \left\{ R(\theta; P_0) := \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\theta, X)] \right\}$$

**Intuition:** We want $\mathcal{P}$ to contain "hard" subpopulations, minority groups, domain changes, and even adversarial shifts.

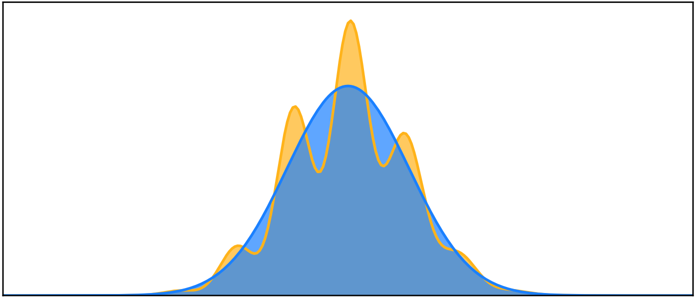
# Divergence-based uncertainty sets

The $f$-*divergence* between distributions $P$ and $Q$ is

$$D_f(P\|Q) := \int f\left(\frac{dP}{dQ}\right) dQ$$

where $f$ is some convex function with $f(1) = 0$.
Use **non-parametric** uncertainty region

$$\mathcal{P} := \{P : D_f(P\|P_0) \le \rho\}$$



# Curvature of $f$

- Curvature of $t \mapsto f(t)$ around $1$ determines size of uncertainty region
- Cressie-Read family [Cressie and Read (1998)] for $k \in (1, \infty)$

$$f_k(t) = \frac{1}{k(k-1)}(t^k - kt + k - 1),$$

where $\mathcal{P}_k := \left\{ P : D_{f_k}\left(P \| P_0\right) = \int f_k\left(\frac{dP}{dP_0}\right) dP_0 \le \rho \right\}$

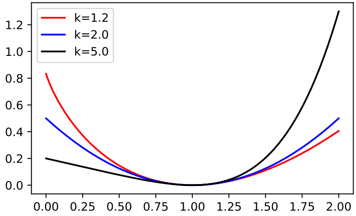
# Curvature of $f$

- Curvature of $t \mapsto f(t)$ around $1$ determines size of uncertainty region
- Cressie-Read family [Cressie and Read (1998)] for $k \in (1, \infty)$

$$f_k(t) = \frac{1}{k(k-1)}(t^k - kt + k - 1),$$

where $\mathcal{P}_k := \left\{ P : D_{f_k}(P \| P_0) = \int f_k\left(\frac{dP}{dP_0}\right) dP_0 \le \rho \right\}$

- Curvature $k$ controls size of $\mathcal{P}_k$.
- As $k \to 1$,
  - $D_f(P \| P_0)$ grows smaller
  - Uncertain set $\mathcal{P}_k$ grows larger
  - DRO is more risk-averse



# Distributionally robust optimization

**Formulation:** For divergence given by $f_k(t) \propto t^k - 1$, solve

$$\operatorname*{minimize}_{\theta \in \Theta} \left\{ R_k(\theta; P_0) := \sup_{P} \left\{ \mathbb{E}_P[\ell(\theta, X)] : D_{f_k}(P \| P_0) \le \rho \right\} \right\}$$

# Distributionally robust optimization

**Formulation:** For divergence given by $f_k(t) \propto t^k - 1$, solve

$$\underset{\theta \in \Theta}{\text{minimize}} \left\{ R_k(\theta; P_0) := \sup_P \{\mathbb{E}_P[\ell(\theta, X)] : D_{f_k}(P \| P_0) \le \rho\} \right\}$$

**Empirical plug-in:** For the empirical measure $\widehat{P}_n$, solve the plug-in

$$\underset{\theta \in \Theta}{\text{minimize}} \left\{ R_k(\theta, \widehat{P}_n) := \sup_P \left\{\mathbb{E}_P[\ell(\theta, X)] : D_{f_k}\left(P \| \widehat{P}_n\right) \le \rho\right\} \right\}$$

Contrast to previous formulation with shrinking robustness $\rho/n$.

# Minimax bounds for $\min_{\theta\in\Theta} R_k(\theta; P_0)$

Recall $R_k(\theta; P_0) := \sup_P \{\mathbb{E}_P[\ell(\theta, X)] : D_{f_k}(P\|P_0) \le \rho\}$

Theorem (Duchi & N. 2018)

*For* $k, k_* = \frac{k}{k-1} \in (1, \infty)$, *and* $\ell(\theta; X) \in [-M, M]$

$$\inf_{\widehat{\theta}} \sup_{P_0} \mathbb{E}_{P_0}\left[R_k(\widehat{\theta}; P_0) - \inf_{\theta\in\Theta} R_k(\theta; P_0)\right] \asymp n^{-\frac{1}{(k_* \vee 2)}}$$

*where infimum is over all measurable functions* $\widehat{\theta} \in \sigma(X_1, \ldots, X_n)$, *and supremum is over all distributions.*

# Minimax bounds for $\min_{\theta\in\Theta} R_k(\theta; P_0)$

Recall $R_k(\theta; P_0) := \sup_P \{\mathbb{E}_P[\ell(\theta, X)] : D_{f_k}(P\|P_0) \le \rho\}$

Theorem (Duchi & N. 2018)

*For* $k, k_* = \frac{k}{k-1} \in (1, \infty)$, *and* $\ell(\theta; X) \in [-M, M]$

$$\inf_{\widehat{\theta}} \sup_{P_0} \mathbb{E}_{P_0}\left[R_k(\widehat{\theta}; P_0) - \inf_{\theta\in\Theta} R_k(\theta; P_0)\right] \approx n^{-\frac{1}{(k_* \vee 2)}}$$

*where infimum is over all measurable functions* $\widehat{\theta} \in \sigma(X_1, \ldots, X_n)$, *and supremum is over all distributions.*

- Upper bound attained by plug-in estimator
- Lower bound shows fudamental statistical cost of robustness

# Upper bound

Recall $k, k_* = \frac{k}{k-1} \in (1, \infty)$, and the plug-in

$$\widehat{\theta}_{k,n} = \operatorname*{argmin}_{\theta \in \Theta} \left\{ R_k(\theta, \widehat{P}_n) := \sup_{P} \left\{ \mathbb{E}_P[\ell(\theta, X)] : D_{f_k}\left(P \| \widehat{P}_n\right) \le \rho \right\} \right\}$$

Theorem (Duchi & N. 2018)

*Let* $\theta \mapsto \ell(\theta; x)$ *be* $L$*-Lipschitz,* $D := \sup_{\theta, \theta' \in \Theta} \|\theta - \theta'\| < \infty$*, and* $\inf_{\theta \in \Theta} \ell(\theta; X) = 0$*. Then, w.p.* $\ge 1 - 2\exp\left(-t + d \log\left(1 + \frac{3DL}{t}\right)\right)$

$$R_k(\widehat{\theta}_{k,n}; P_0) \le \inf_{\theta \in \Theta} R_k(\theta; P_0) + 2C_{k,\rho} DL\sqrt{t} n^{-\frac{1}{(k_* \vee 2)}}$$

*for a constant* $C_{k,\rho} > 0$ *that depends only on* $k$ *and* $\rho$*.*

# Lower bound

Theorem (Duchi & N. 2018)

*Let* $\ell(\theta; X) = \theta X$ *with* $\theta \in \Theta = [-M, M]$ *and* $\xi \in [-1, 1]$*. Then, for a constant* $c_{k,\rho}$ *that only depends on* $k$ *and* $\rho$

$$\inf_{\widehat{\theta}} \sup_{P_0} \mathbb{E}_{P_0} \left[ R_f(\widehat{\theta}; P_0) - \inf_{\theta \in \Theta} R_k(\theta; P_0) \right] \geq c_{k,\rho} M n^{-\frac{1}{(k_* \vee 2)}}$$

*where infimum is over* $\sigma(X_1, \ldots, X_n)$*-measurable mappings, and supremum is over all probability distributions.*

# Lower bound

Theorem (Duchi & N. 2018)

*Let $\ell(\theta; X) = \theta X$ with $\theta \in \Theta = [-M, M]$ and $\xi \in [-1, 1]$. Then, for a constant $c_{k,\rho}$ that only depends on $k$ and $\rho$*

$$\inf_{\widehat{\theta}} \sup_{P_0} \mathbb{E}_{P_0}\left[R_f(\widehat{\theta}; P_0) - \inf_{\theta \in \Theta} R_k(\theta; P_0)\right] \geq c_{k,\rho} M n^{-\frac{1}{(k_* \vee 2)}}$$

*where infimum is over $\sigma(X_1, \ldots, X_n)$-measurable mappings, and supremum is over all probability distributions.*

- Worst than parametric rate for $k \in (1, 2)$ and $k_* = k/(k-1) \in (2, \infty)$
- Statistical cost of distributional robustness
- Lower bound applies to any $f$-divergence $f(t) \propto t^k - 1$.

# Remarks

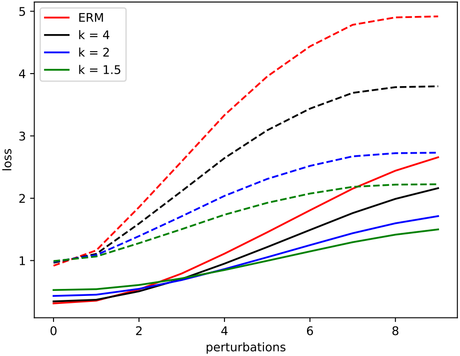
- Our upper and lower bounds are tight up to dimension dependent constants
- Lower bound can be loose in high dimensions
- Central limit theorem: under suitable conditions,

$$\sqrt{n}(\widehat{\theta}_{k,n} - \theta^\star) \overset{d}{\rightsquigarrow} N(0, A)$$

  where $\widehat{\theta}_{k,n}$ is empirical plug-in, and $A$ can be fully-specified.
- Worst-case rate different from asymptotic rate

# Experiment: SVM sanity check

Test on distributions with adversarially shifted true classifier




$$\ell(\theta; (w, y)) = \left(1 - y w^\top \theta\right)_+$$

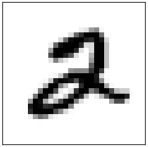
# Experiment: Domain Generalization

**Problem:** Given an hand-written or type-written digit, classify it

- Majority group: hand-written, minority group: type-written
- Data: MNIST hand-written training dataset comprising of $n_{\text{train}} = 60,000$ digits with $\{0, 6, 10, 60, 100, 600\}$ images per digit replaced with a type-written dataset (with the same label).
- Multiclass logistic loss

2

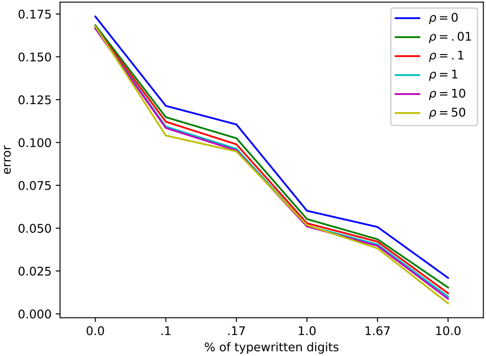
Type-written data



Hand-written data

# Experiment: Domain Generalization

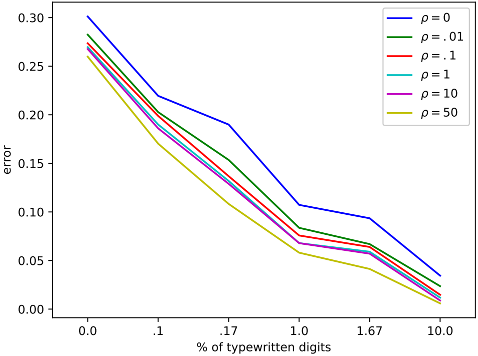
Performance on minority group




Test error on type-written all digits

# Experiment: Domain Generalization

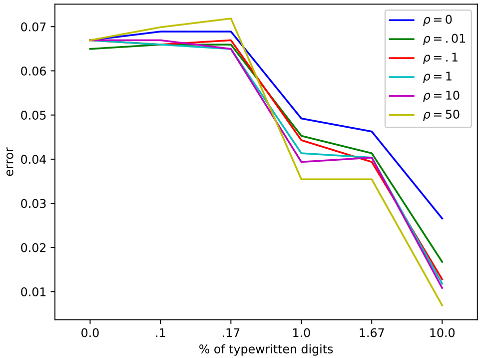
Performance on "hard" digit in minority group




Test error on type-written digit $9$

# Experiment: Domain Generalization

Performance on "easy" digit in minority group




Test error on type-written digit $3$

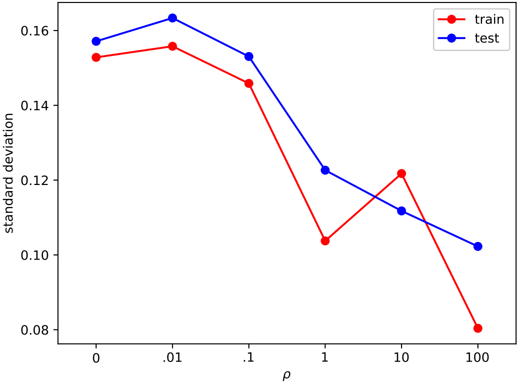
# Experiment: fine-grained recognition

- 120 distinct classes (all dog breeds) [Khosla et al. 11]

Cairn

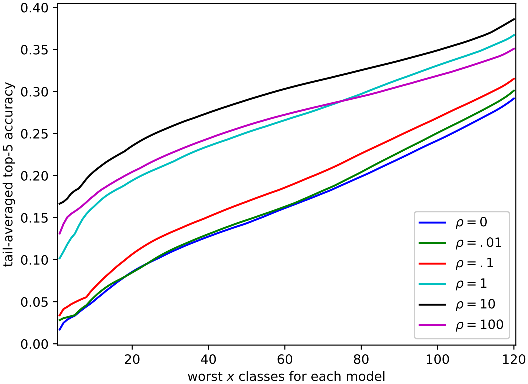
Border

# Experiment: fine-grained recognition



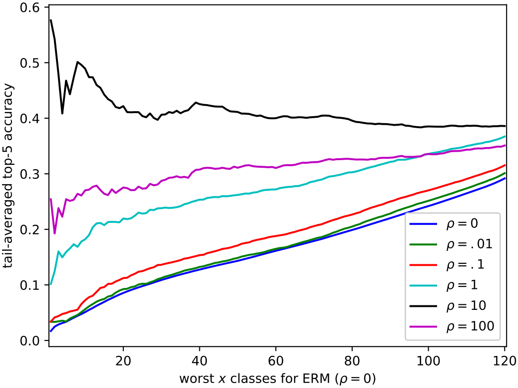

Variation of top-$5$ accuracy across 120 classes

# Experiment: fine-grained recognition




Test top-5 accuracy evaluated on worst $x$ classes for each model

# Experiment: fine-grained recognition



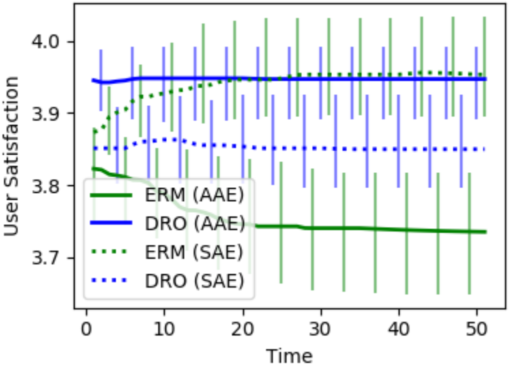

Test top-5 accuracy evaluated on worst $x$ classes for empirical risk minimization

# Representation Disparity Amplification

Problem: Users may drop out of service if low performance

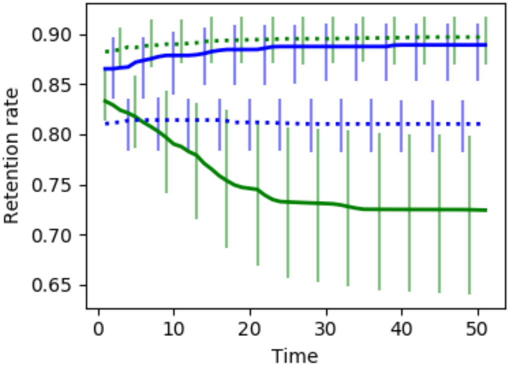
- Evaluate user satisfaction and retention on Mechanical Turk
- Corpora (tweets) from two demographic groups: Caucasians (SAE), African Americans (AAE)
- Task: autocomplete 10 tweets
- Use satisfaction survey to estimate user retention, repeat with changed demographic proportions
- See [Hashimoto, Srivastava, N., Liang 18] for details

# Representation Disparity Amplification




Green: ERM, Blue: DRO, real-line: AAE (minority), dotted-line: SAE

# Representation Disparity Amplification




Green: ERM, Blue: DRO, real-line: AAE (minority), dotted-line: SAE

# Revisiting choice of uncertainty region

Distributionally robust formulations depend heavily on uncertainty region

$$\underset{\theta \in \Theta}{\text{minimize}} \quad \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\theta, X)]$$

# Revisiting choice of uncertainty region

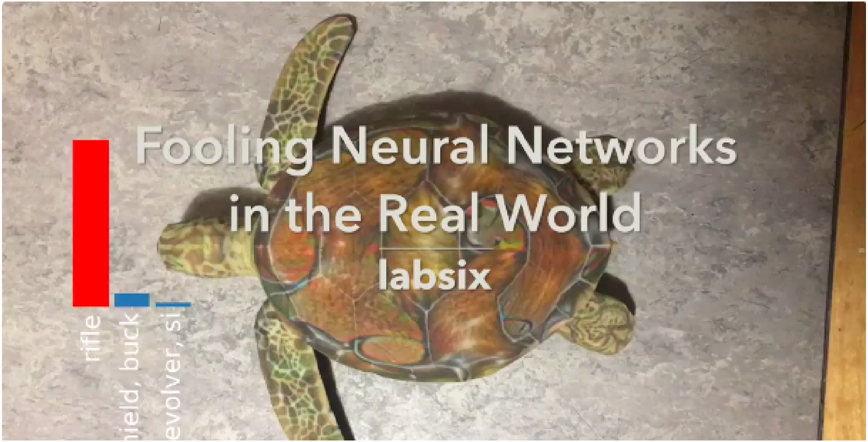
Distributionally robust formulations depend heavily on uncertainty region

$$\underset{\theta \in \Theta}{\text{minimize}} \; \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\theta, X)]$$

Q: Are there better choices of uncertainty sets $\mathcal{P}$, especially for over-parameterized models such as deep nets?

# Why changing support is important

- Deep networks are not robust




Athalye et al. (2017)

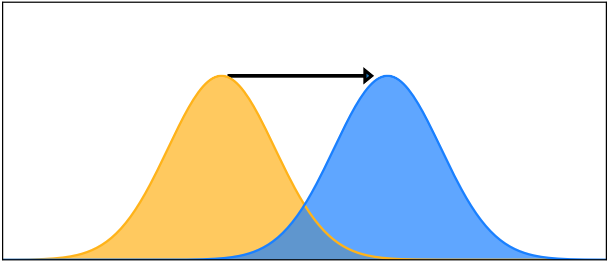
# Wasserstein-based robustness sets

Define *Wasserstein distance* from a (convex) transportation cost function $c$

$$W_c(P, Q) := \max_{h} \left\{ \int h(x) \left[p(x) - q(x)\right] dz \mid h(x) - h(x') \le c(x, x') \right\}$$

Use uncertainty region

$$\mathcal{P}_\rho := \{P : W_c(P, P_0) \le \rho\}$$



# Wasserstein robustness

Look at distributionally robust risk

$$\underset{\theta \in \Theta}{\text{minimize}} \sup_{P} \{\mathbb{E}_P[\ell(\theta; Z)] \mid P \in \mathcal{P}\}$$

# Wasserstein robustness

Look at distributionally robust risk defined for $\rho \geq 0$

$$R(\theta, \rho) := \sup_{P} \{\mathbb{E}_P[\ell(\theta; Z)] \;\; \text{s.t. } W_c(P, P_0) \leq \rho\}$$

# Wasserstein robustness

Look at distributionally robust risk defined for $\rho \geq 0$

$$R(\theta, \rho) := \sup_{P} \{\mathbb{E}_P[\ell(\theta; Z)] \ \text{ s.t. } W_c(P, P_0) \leq \rho\}$$

- Allows *changing support* to harder distributions [Shafieezadeh-Abadeh et al. 15, Esfahani & Kuhn 15, Blanchet and Murthy 16, Blanchet et al 16]

**Example** (Linear models): If loss $\ell(\theta, x, y) = \phi(\theta^T xy)$ for some $\phi$, then

- if $c(x, x') = \|x - x'\|_\infty$, yields data-dependent $\ell_1$-regularization
- if $c(x, x') = \|x - x'\|_2$, yields data-dependent $\ell_2$-regularization

# Wasserstein robustness

Look at distributionally robust risk defined for $\rho \geq 0$

$$R(\theta, \rho) := \sup_{P} \{\mathbb{E}_P[\ell(\theta; Z)] \text{ s.t. } W_c(P, P_0) \leq \rho\}$$

- Allows *changing support* to harder distributions [Shafieezadeh-Abadeh et al. 15, Esfahani & Kuhn 15, Blanchet and Murthy 16, Blanchet et al 16]

**Example** (Linear models): If loss $\ell(\theta, x, y) = \phi(\theta^T xy)$ for some $\phi$, then

- if $c(x, x') = \|x - x'\|_\infty$, yields data-dependent $\ell_1$-regularization
- if $c(x, x') = \|x - x'\|_2$, yields data-dependent $\ell_2$-regularization

**Minor issue:** Often NP-hard when not simple linear model

# Duality and robustness

Theorem (Blanchet and Murthy (2016))

*Let $P_0$ be any distribution on $\mathcal{Z}$ and $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ be any function. Then*

$$\sup_{W_c(P,P_0)\le\rho} \mathbb{E}_P[\ell(\theta; Z)] = \inf_{\lambda\ge 0} \left\{ \int \sup_{z'} \left\{ \ell(\theta; z') - \lambda c(z', z) \right\} dP_0(z) + \lambda\rho \right\}$$

$$= \inf_{\lambda\ge 0} \left\{ \mathbb{E}_{P_0} \left[ \ell_\lambda(\theta; Z) \right] + \lambda\rho \right\}.$$

# Duality and robustness

Theorem (Blanchet and Murthy (2016))

*Let $P_0$ be any distribution on $\mathcal{Z}$ and $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ be any function. Then*

$$\sup_{W_c(P,P_0)\le\rho} \mathbb{E}_P[\ell(\theta; Z)] = \inf_{\lambda\ge 0} \left\{ \int \sup_{z'} \left\{ \ell(\theta; z') - \lambda c(z', z) \right\} dP_0(z) + \lambda\rho \right\}$$
$$= \inf_{\lambda\ge 0} \left\{ \mathbb{E}_{P_0}\left[\ell_\lambda(\theta; Z)\right] + \lambda\rho \right\}.$$

**Computational Idea:** Pick a large enough $\lambda$, and “solve”

$$\operatorname*{minimize}_{\theta}\ \mathbb{E}_{P_0}\left[\ell_\lambda(\theta; Z)\right]$$

# A first idea

(Simple) insight: If $\ell(\theta, z)$ is smooth in $\theta$ and $z$, then life gets a bit easier

# A first idea

(Simple) insight: If $\ell(\theta, z)$ is smooth in $\theta$ and $z$, then life gets a bit easier

The function

$$\ell_\lambda(\theta; z) := \sup_{\Delta} \left\{ \ell(\theta; z + \Delta) - \frac{\lambda}{2} \|\Delta\|_2^2 \right\}$$

is efficient to compute (and differentiable, etc.) for *large enough* $\lambda$

# Stochastic gradient algorithm

$$\underset{\theta}{\text{minimize}} \;\; \mathbb{E}_{P_0}[\ell_\lambda(\theta; Z)] = \mathbb{E}_{P_0}\left[\sup_{\Delta}\left\{\ell(\theta; Z + \Delta) - \frac{\lambda}{2}\|\Delta\|_2^2\right\}\right]$$

**Repeat:**

1. Draw $Z_k \stackrel{\text{iid}}{\sim} P$
2. Compute (approximate) maximizer

$$\widehat{Z}_k \approx \operatorname*{argmax}_z \left\{\ell(\theta; z) - \frac{\lambda}{2}\|z - Z_k\|_2^2\right\}$$

3. For a stepsize $\alpha_k$, update

$$\theta_{k+1} := \theta_k - \alpha_k \nabla_\theta \ell(\theta_k; \widehat{Z}_k)$$

# Stochastic gradient algorithm

$$\underset{\theta}{\text{minimize}} \ \mathbb{E}_{P_0}[\ell_\lambda(\theta; Z)] = \mathbb{E}_{P_0}\left[\sup_{\Delta}\left\{\ell(\theta; Z + \Delta) - \frac{\lambda}{2}\|\Delta\|_2^2\right\}\right]$$

**Repeat:**

1. Draw $Z_k \overset{\text{iid}}{\sim} P$
2. Compute (approximate) maximizer

$$\widehat{Z}_k \approx \underset{z}{\text{argmax}}\left\{\ell(\theta; z) - \frac{\lambda}{2}\|z - Z_k\|_2^2\right\}$$

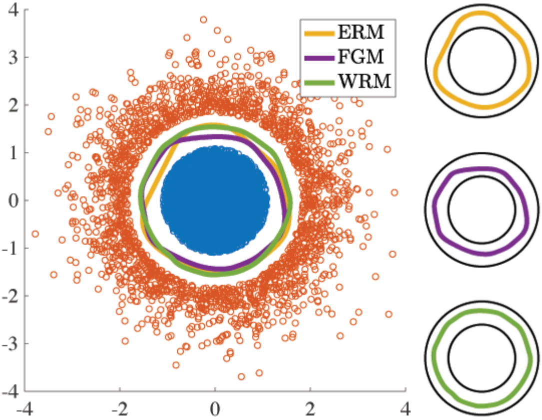
3. For a stepsize $\alpha_k$, update

$$\theta_{k+1} := \theta_k - \alpha_k \nabla_\theta \ell(\theta_k; \widehat{Z}_k)$$

**Theorem(ish):** This converges with all the typical convergence properties

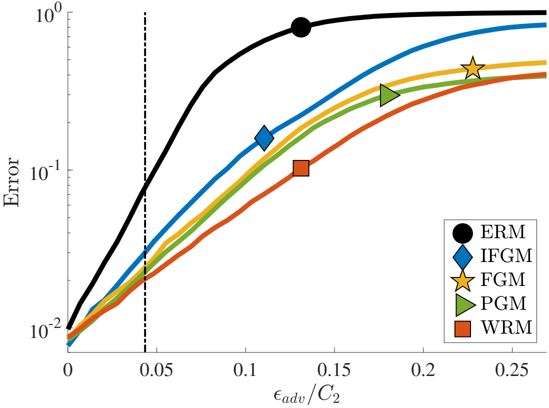
# Simple Visualization

$$y = \operatorname{sign}(\|x\|_2 - \sqrt{2})$$



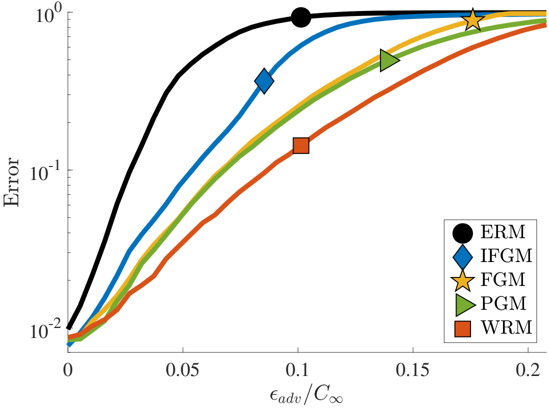
# Experimental results: adversarial classification

- MNIST dataset with 3 convolutional layers, fully connected softmax top layer

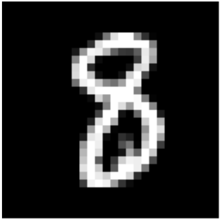


# Experimental results: adversarial classification

- MNIST dataset with 3 convolutional layers, fully connected softmax top layer

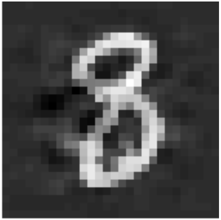


# Reading tea leaves

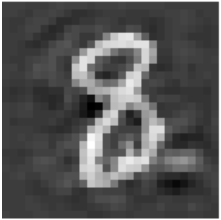


Original

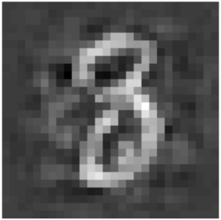
ERM



FGM



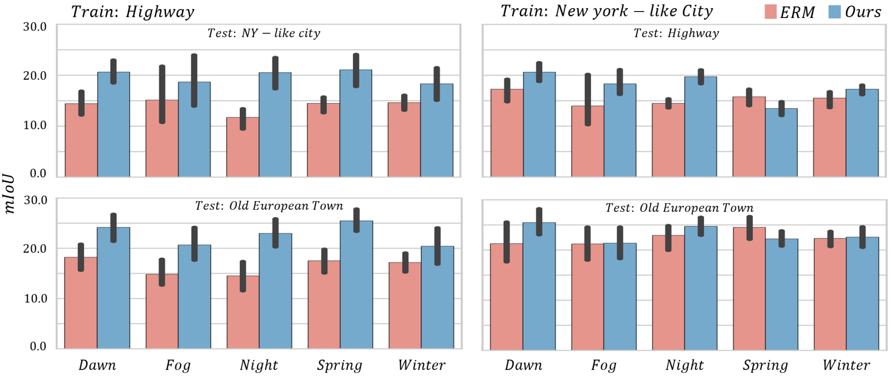
IFGM



PGM

WRM

# Generate examples for new domains




[Volpi*, N.*, Sener, Duchi, Murino, Savarese 18]

# Conclusion

1. Statistical consequences of distributional robustness important
2. Duality provides both certificates and allows efficient methods

# Conclusion

1. Statistical consequences of distributional robustness important
2. Duality provides both certificates and allows efficient methods

**Future work:**

1. More work to do on how to choose robustness sets! ($f$, $c$, $\rho$)
2. When should we use divergence- vs. distance-based?
3. Distributional robustness and temporal shifts
4. Causal connections: correspondence between uncertainty regions vs. interventions and confounding variables
5. Principled view on adversarial training
6. Risk-averse decision-making (reinforcement learning)

# Appendix

# Empirical likelihood main

The *empirical likelihood confidence region* is

# Empirical likelihood main

The *empirical likelihood confidence region* is

$$E_n(\rho) := \left\{ \sum_{i=1}^{n} p_i Z_i : D_{\chi^2}\left(p \| \mathbf{1}/n\right) \le \frac{\rho}{n} \right\}.$$

[Owen 90, Baggerly 98, Newey and Smith 01, Imbens 02]

# Empirical likelihood main

The *empirical likelihood confidence region* is

$$\begin{aligned}
E_n(\rho) &:= \left\{ \sum_{i=1}^n p_i Z_i : D_{\chi^2}\left(p \| \mathbf{1}/n\right) \le \frac{\rho}{n} \right\} \\
&= \left\{ \sum_{i=1}^n p_i Z_i : \frac{1}{n}\sum_{i=1}^n (np_i - 1)^2 \le \frac{\rho}{n}, p^\top \mathbf{1} = 1, p \ge 0 \right\} \\
&= \frac{1}{n}\sum_{i=1}^n Z_i + \left\{ \sum_{i=1}^n u_i Z_i : \|u\|_2^2 \le \frac{\rho}{n^2}, u^\top \mathbf{1} = 0, u \ge -\frac{\mathbf{1}}{n} \right\}
\end{aligned}$$

by letting $u_i = p_i - \frac{1}{n}$.

# Empirical likelihood main

The *empirical likelihood confidence region* is

$$\begin{aligned} E_n(\rho) &:= \left\{ \sum_{i=1}^n p_i Z_i : D_{\chi^2}\left(p \| \mathbf{1}/n\right) \le \frac{\rho}{n} \right\} \\ &= \left\{ \sum_{i=1}^n p_i Z_i : \frac{1}{n} \sum_{i=1}^n (np_i - 1)^2 \le \frac{\rho}{n}, p^\top \mathbf{1} = 1, p \ge 0 \right\} \\ &= \frac{1}{n} \sum_{i=1}^n Z_i + \Bigg\{ \underbrace{\sum_{i=1}^n u_i Z_i}_{\text{Ellipse from } \textit{data}} : \|u\|_2^2 \le \frac{\rho}{n^2}, u^\top \mathbf{1} = 0, u \ge -\frac{\mathbf{1}}{n} \Bigg\} \end{aligned}$$

by letting $u_i = p_i - \frac{1}{n}$.

# Robust Optimization ≈ Variance Regularization main

**Proof Sketch** Let $z_i = \ell(\theta; X_i)$, $u_i = p_i - \frac{1}{n}$, and denote by $\bar{z}$ and $s_n^2$ the sample mean and variance respectively.

# Robust Optimization $\approx$ Variance Regularization main

**Proof Sketch** Let $z_i = \ell(\theta; X_i)$, $u_i = p_i - \frac{1}{n}$, and denote by $\bar{z}$ and $s_n^2$ the sample mean and variance respectively.

$$R_n(\theta; \mathcal{P}_{n,\rho}) = \max_p \left\{ \langle p, z \rangle : D_{\chi^2}\left(p \| \mathbf{1}/n\right) \leq \frac{\rho}{n} \right\}$$

# Robust Optimization $\approx$ Variance Regularization main

**Proof Sketch** Let $z_i = \ell(\theta; X_i)$, $u_i = p_i - \frac{1}{n}$, and denote by $\bar{z}$ and $s_n^2$ the sample mean and variance respectively.

$$R_n(\theta; \mathcal{P}_{n,\rho}) = \max_p \left\{ \langle p, z \rangle : \frac{1}{n} \sum_{i=1}^n (np_i - 1)^2 \leq \frac{\rho}{n}, p^\top \mathbf{1} = 1, p \geq 0 \right\}$$

# Robust Optimization $\approx$ Variance Regularization main

**Proof Sketch** Let $z_i = \ell(\theta; X_i)$, $u_i = p_i - \frac{1}{n}$, and denote by $\bar{z}$ and $s_n^2$ the sample mean and variance respectively.

$$
\begin{aligned}
R_n(\theta; \mathcal{P}_{n,\rho}) &= \max_p \left\{ \langle p, z \rangle : \frac{1}{n}\sum_{i=1}^n (np_i - 1)^2 \le \frac{\rho}{n}, p^\top \mathbf{1} = 1, p \ge 0 \right\} \\
&= \bar{z} + \max_u \left\{ \langle u, z - \bar{z} \rangle : \|u\|_2^2 \le \frac{\rho}{n^2}, u^\top \mathbf{1} = 0, u \ge -\frac{\mathbf{1}}{n} \right\}
\end{aligned}
$$

# Robust Optimization $\approx$ Variance Regularization main

**Proof Sketch** Let $z_i = \ell(\theta; X_i)$, $u_i = p_i - \frac{1}{n}$, and denote by $\bar{z}$ and $s_n^2$ the sample mean and variance respectively.

$$
\begin{aligned}
R_n(\theta; \mathcal{P}_{n,\rho}) &= \max_p \left\{ \langle p, z \rangle : \frac{1}{n} \sum_{i=1}^n (np_i - 1)^2 \le \frac{\rho}{n}, p^\top \mathbf{1} = 1, p \ge 0 \right\} \\
&= \bar{z} + \max_u \left\{ \langle u, z - \bar{z} \rangle : \|u\|_2^2 \le \frac{\rho}{n^2}, u^\top \mathbf{1} = 0, u \ge -\frac{\mathbf{1}}{n} \right\} \\
&\le \bar{z} + \frac{\sqrt{2\rho}}{n} \|z - \bar{z}\|_2 = \bar{z} + \sqrt{\frac{2\rho}{n} s_n^2} \quad \text{by Cauchy-Schwarz}
\end{aligned}
$$

# Robust Optimization $\approx$ Variance Regularization main

**Proof Sketch** Let $z_i = \ell(\theta; X_i)$, $u_i = p_i - \frac{1}{n}$, and denote by $\bar{z}$ and $s_n^2$ the sample mean and variance respectively.

$$
\begin{aligned}
R_n(\theta; \mathcal{P}_{n,\rho}) &= \max p \left\{ \langle p, z \rangle : \frac{1}{n} \sum_{i=1}^{n} (np_i - 1)^2 \le \frac{\rho}{n}, p^\top \mathbf{1} = 1, p \ge 0 \right\} \\
&= \bar{z} + \max_u \left\{ \langle u, z - \bar{z} \rangle : \|u\|_2^2 \le \frac{\rho}{n^2}, u^\top \mathbf{1} = 0, u \ge -\frac{\mathbf{1}}{n} \right\} \\
&\le \bar{z} + \frac{\sqrt{2\rho}}{n} \|z - \bar{z}\|_2 = \bar{z} + \sqrt{\frac{2\rho}{n} s_n^2} \quad \text{by Cauchy-Schwartz}
\end{aligned}
$$

Last inequality is tight if for all $i$

$$
u_i = \frac{1}{n} \sqrt{\frac{2\rho}{n s_n^2}} (z_i - \bar{z}) \ge -\frac{1}{n}
$$

# Extensions and issues main

**Issue:** What if $\theta^\star \in \mathbb{R}^d$ is not unique?

# Extensions and issues main

**Issue:** What if $\theta^\star \in \mathbb{R}^d$ is not unique?
Let $S = \mathrm{argmin}_{\theta\in\Theta} R(\theta)$ and

$$\boldsymbol{r}^\star = \min_{\theta^\star \in S} \max_{\theta\in S} \|\theta - \theta^\star\|_2$$

Then [Duchi, Glynn & N. 16]

$$\mathbb{P}\left(\inf_{\theta\in\Theta} R(\theta) \le R_n(\widehat{\theta}^{\mathrm{rob}}, \mathcal{P}_{n,\rho})\right)$$
$$\ge \mathbb{P}\left(N(0,1) + \sqrt{\rho} \ge \boldsymbol{r}^\star \sqrt{\rho \mathrm{Var}(\ell(x^\star;\xi))(d+1)}\right) + O(n^{-\frac{1}{2}}).$$

# Extensions and issues main

**Issue:** What if $\theta^\star \in \mathbb{R}^d$ is not unique?
Let $S = \mathrm{argmin}_{\theta \in \Theta} R(\theta)$ and

$$\boldsymbol{r}^\star = \min_{\theta^\star \in S} \max_{\theta \in S} \|\theta - \theta^\star\|_2$$

Then [Duchi, Glynn & N. 16]

$$\mathbb{P}\left(\inf_{\theta \in \Theta} R(\theta) \le R_n(\widehat{\theta}^{\mathrm{rob}}, \mathcal{P}_{n,\rho})\right) \\ \ge \mathbb{P}\left(N(0,1) + \sqrt{\rho} \ge \boldsymbol{r}^\star \sqrt{\rho \mathrm{Var}(\ell(x^\star; \xi))(d+1)}\right) + O(n^{-\frac{1}{2}}).$$

- If $\boldsymbol{r}^\star$ large, then lose confidence, if $\boldsymbol{r}^\star$ small, good shape