Reliable Machine Learning via Distributional Robustness

Hongseok Namkoong

namkoong@gsb.columbia.edu

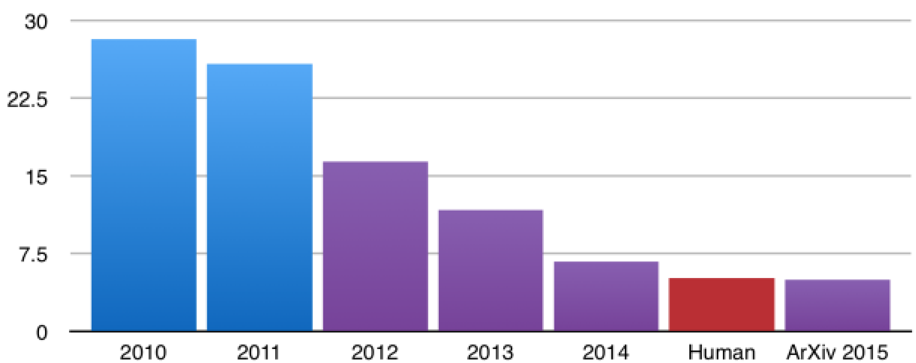
Columbia University

Based on joint works with
John Duchi, Peter Glynn, Tatsu Hashimoto,
Percy Liang, and Megha Srivastava

Progress in machine learning?

Human-level average performance

Image recognition [Eckersley+ '17]



Face recognition [Harris+ '15]

TECH • GOOGLE

Google: Our new system for recognizing faces is the best one ever

By DERRICK HARRIS March 17, 2015

FORTUNE

Poor performance on underrepresented examples

Amazon scraps secret AI recruiting tool that showed bias against women REUTERS

Facial Recognition Is Accurate, if You're a White Guy

By Steve Lohr

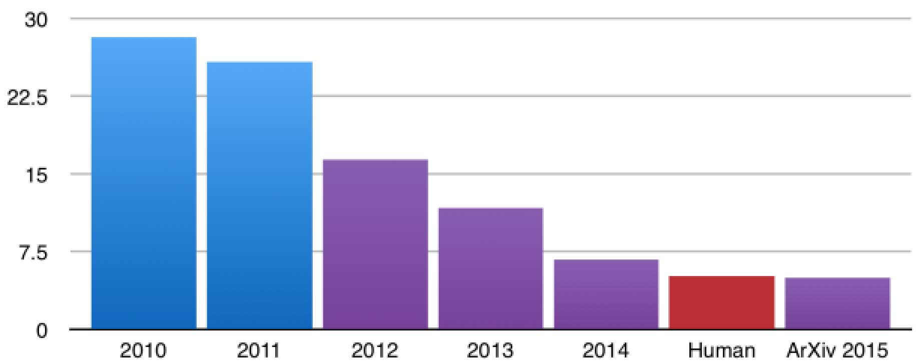
Feb. 9, 2018

The New York Times

Progress in machine learning?

Human-level average performance

Image recognition [Eckersley+ '17]



Face recognition [Harris+ '15]

TECH • GOOGLE

Google: Our new system for recognizing faces is the best one ever

By DERRICK HARRIS March 17, 2015

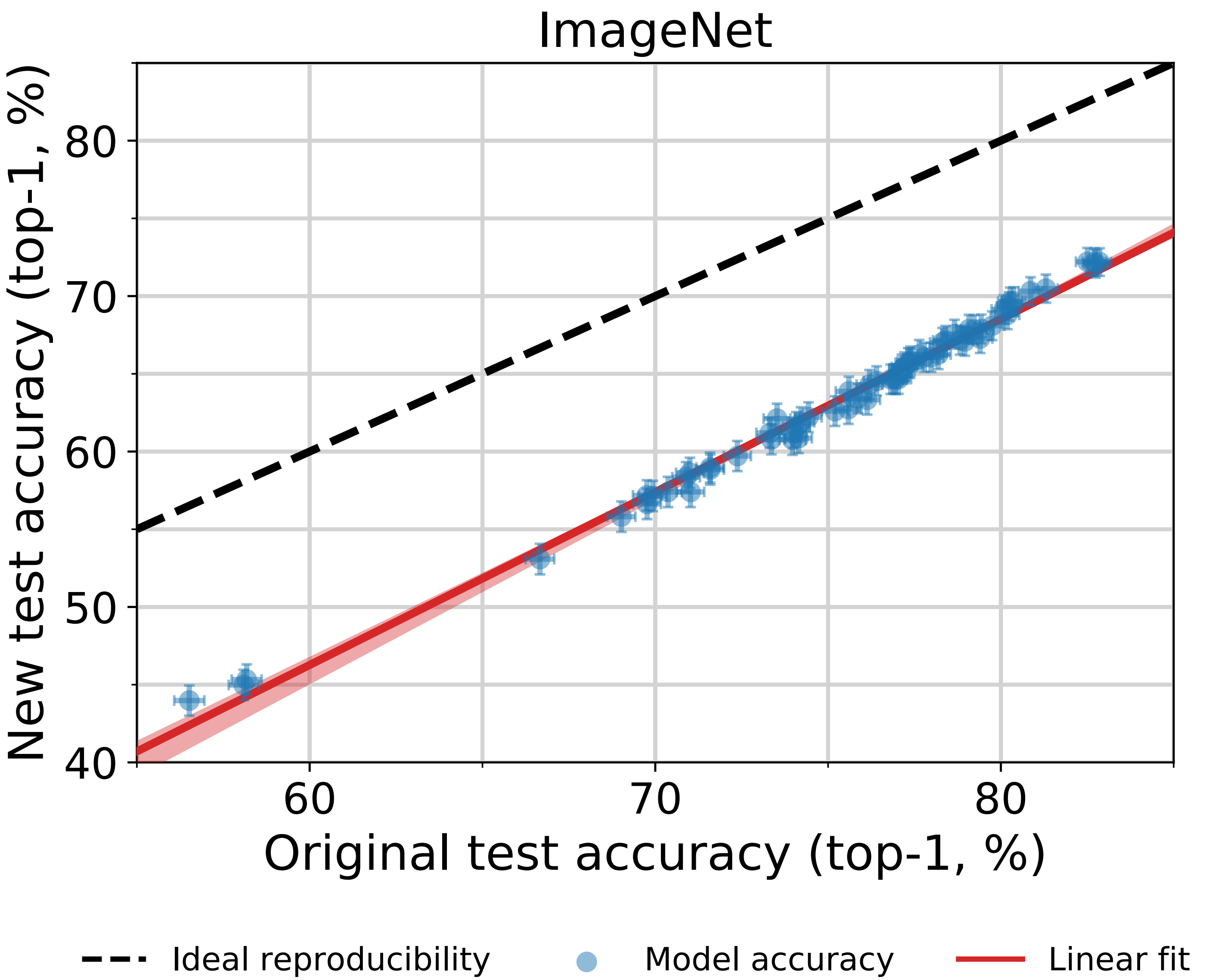
FORTUNE

Poor performance on underrepresented examples

Amazon scraps secret AI recruiting tool that showed bias against women REUTERS

Facial Recognition Is Accurate, if You're a White Guy The New York Times

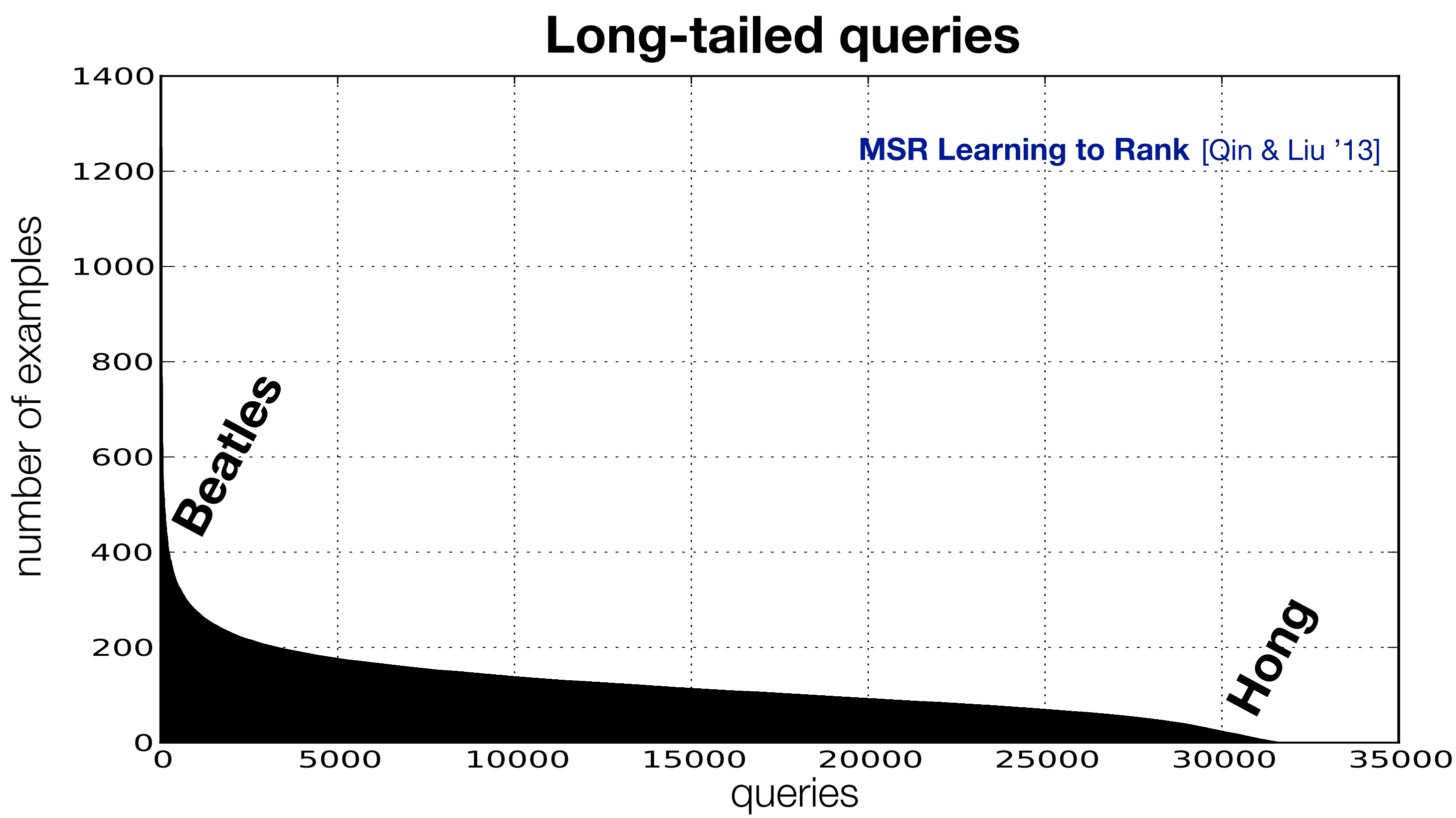
Challenge 0: Robustness



[Does ImageNet classifiers generalize to ImageNet? Recht, Roelofs, Schmidt, Shankar '19]

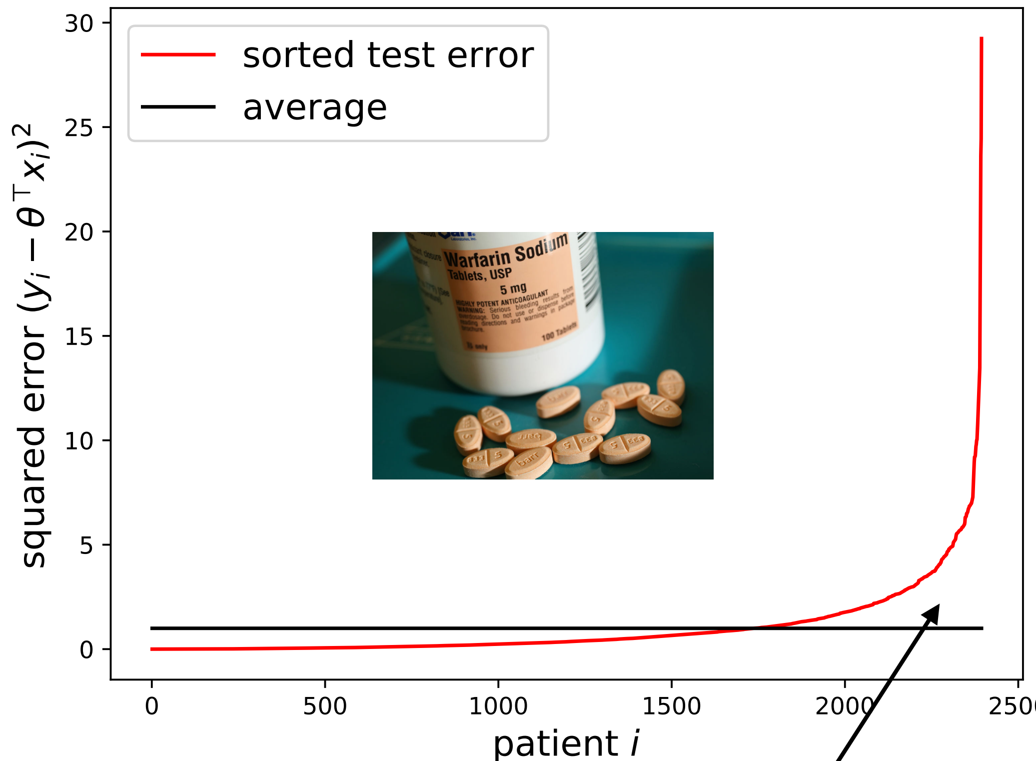
Challenge 1: Long-tails

- Long-tailed data is ubiquitous in modern applications
  - Google (7 yrs ago): constant fraction of queries were new each day
- Tail inputs often determine quality of service



Example: Predicting warfarin dosage

- Warfarin is the most widely used blood-thinner worldwide
- Task: learn to predict therapeutic warfarin dosage
- Personalized treatment recommendation based on regression models [International Warfarin Pharmacogenetics Consortium '09]
  - Worked best out of polynomial regression, kernel methods, neural networks, regression splines, boosting [IWPC '09]



Tail performance is ***orders of magnitude*** **worse than average**

Another use for Warfarin: **rat poison**

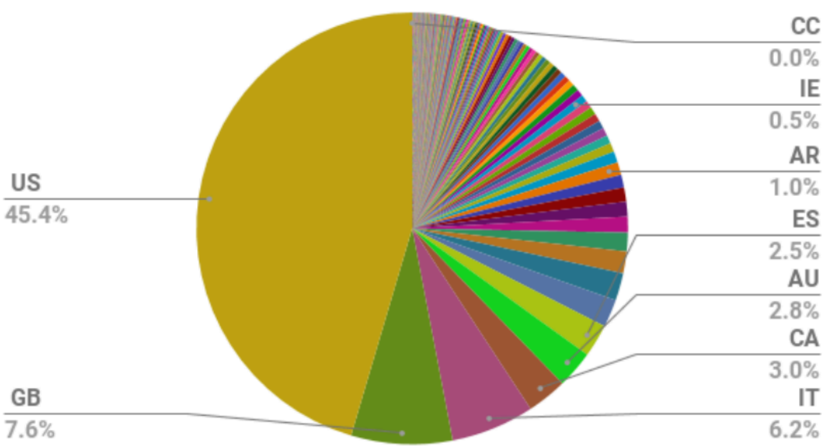
Challenge 2: Lack of diversity in data

- "Clinical trials for new drugs **skew heavily white**"
  - Less than 5% of cancer trial participants were non-white

[Oh et al. '15, Burchard et al. '15, Chen et al., '14, SA Editors '18]

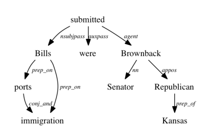
- Majority of image data from **US & Western Europe**

ImageNet: country of origin



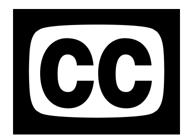
[Shankar et al. '17]

Other examples



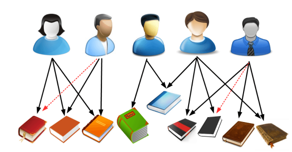
Dependency parsing

[Blodgett+ 16]



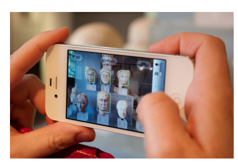
Captioning

[Tatman+ 17]



Recommender systems

[Ekstrand+ 17,18]

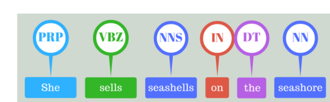


Face recognition

[Grother+ 11]

Language identification

[Blodgett+ 16, Jurgens +17]



Part-of-speech tagging

[Hovy+ 15]

Standard Approach: Average Loss

- Loss/Objective $\ell(\theta; Z)$ where $\theta \in \Theta$ is parameter/decision to be learned, and $Z \sim P_{\text{obs}}$ is random data
- Optimize average performance under P_{obs}

$$\text{minimize}_{\theta \in \Theta} \; \mathbb{E}_{P_{\text{obs}}}[\ell(\theta; Z)]$$

Linear regression $\ell(\theta; X, Y) = (Y - \theta^\top X)^2$

SVM (Classification) $\ell(\theta; X, Y) = (1 - Y\theta^\top X)_+$

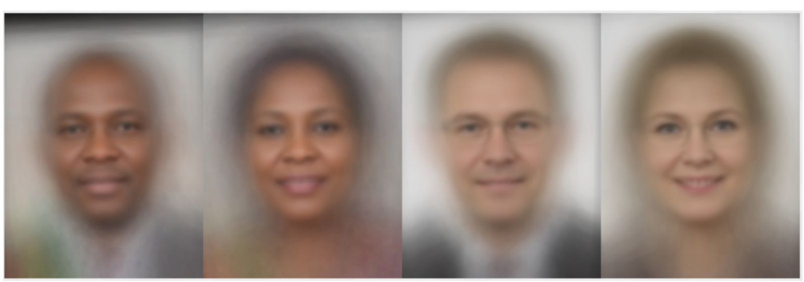
Deep neural networks $\ell(\theta; X, Y) = (Y - \sigma_1(\theta_1 \cdots \sigma_k(\theta_k \cdot X))))^2$

Example: Facial recognition

- Labeled Faces in the Wild, a gold standard dataset for face recognition, is **77.5% male**, and **83.5% White** [Han and Jain ’14]

- Commercial gender classification softwares have **disparate** performance on different subpopulations

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Gendered Shades: Intersectional accuracy disparity [Buolamwini and Gebru ’18]

Distributionally robust optimization

Standard approach: Solve average risk minimization problem

$$\text{minimize}_{\theta \in \Theta} \ \mathbb{E}_{P_{\text{obs}}}[\ell(\theta; Z)]$$

Today: Solve distributionally robust optimization problem

$$\underset{\theta \in \Theta}{\text{minimize}} \ \max_{Q \in \mathcal{P}} \ \mathbb{E}_{Q}[\ell(\theta; Z)]$$

for some carefully chosen set of probabilities $\mathcal{P}$

Idea: Do well almost all the time, instead of on average!

References: [Ben-Tal et al. 13, Bertsimas et al. 16, Blanchet & Murthy 16, Blanchet et al. 16, Gao & Kleywegt 16, Lam & Zhou 17, Lam 18, and many others]

f-divergences

Idea: Instead of using the empirical distribution P_{obs}, look at all distributions "near" it

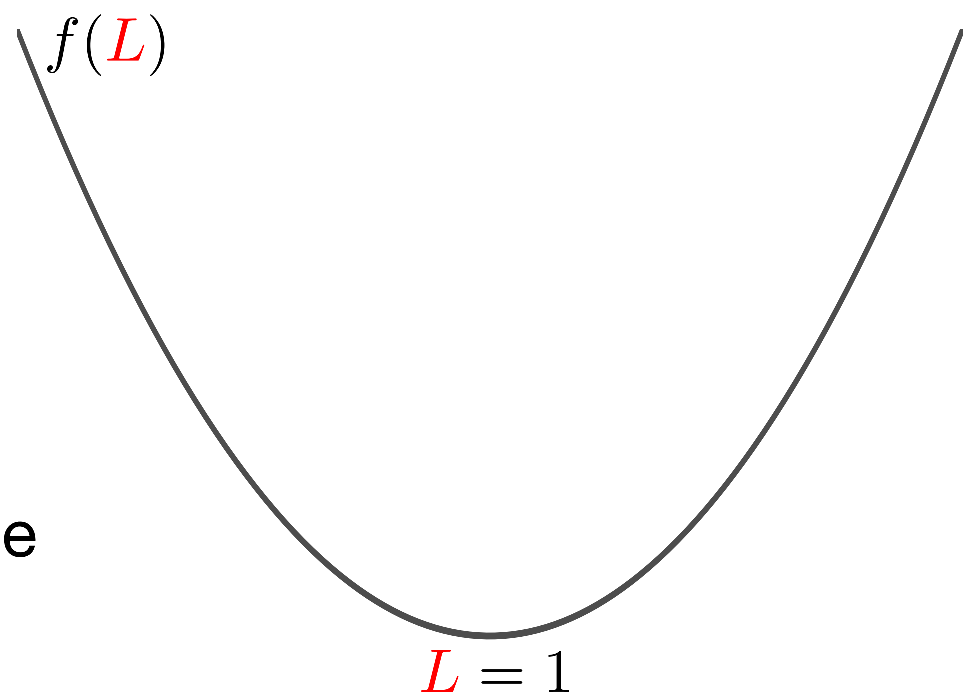
Notion of distance:

f-divergence: If $L = \frac{dQ}{dP}$ is "near 1", then Q and P are near

For a convex function $f : \mathbb{R}_+ \to \mathbb{R}_+$ with $f(1) = 0$,

$$D_f\left(Q \| P\right) := \mathbb{E}_P\left[f\left(\frac{dQ}{dP}\right)\right]$$

As curvature of f decreases, the divergence becomes smaller!



Distributionally robust optimization

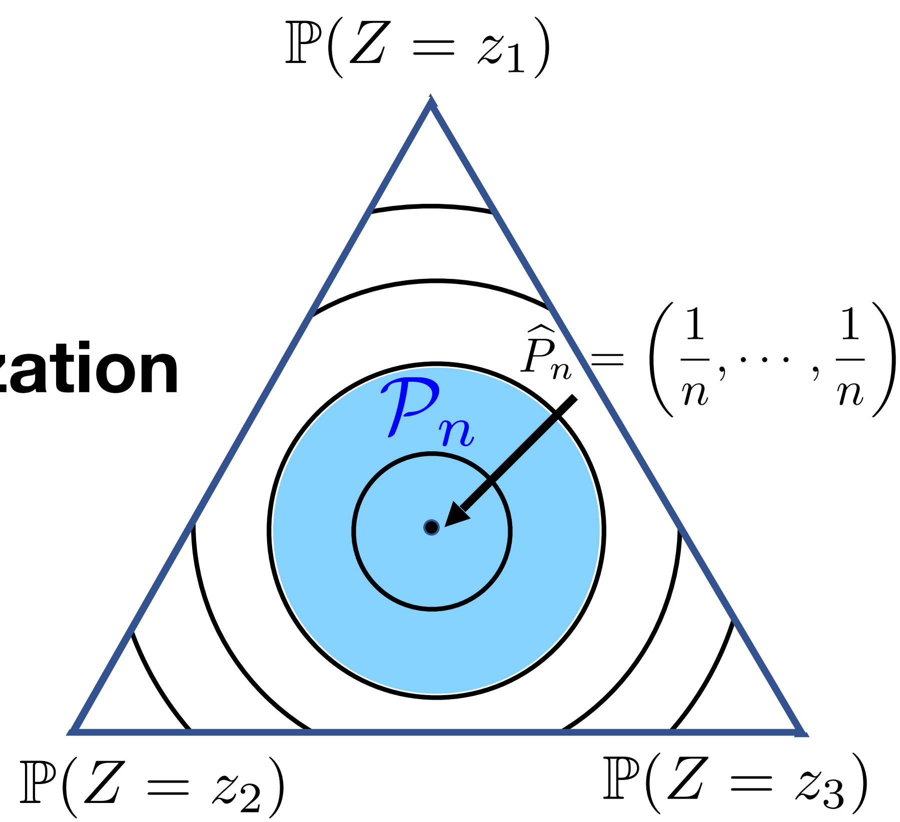
Idea: Instead of using the empirical distribution $\widehat{P}_n$, look at all distributions "near" it

Worst-case region

$$\mathcal{P}_n(\rho) := \left\{ Q : D_f \left(Q \| \widehat{P}_n \right) \leq \rho \right\}$$

Distributionally Robust Optimization

$$\underset{\theta \in \Theta}{\text{minimize}} \quad \max_{q \in \mathcal{P}_n(\rho)} \sum_{i=1}^{n} q_i \ell(\theta; Z_i)$$



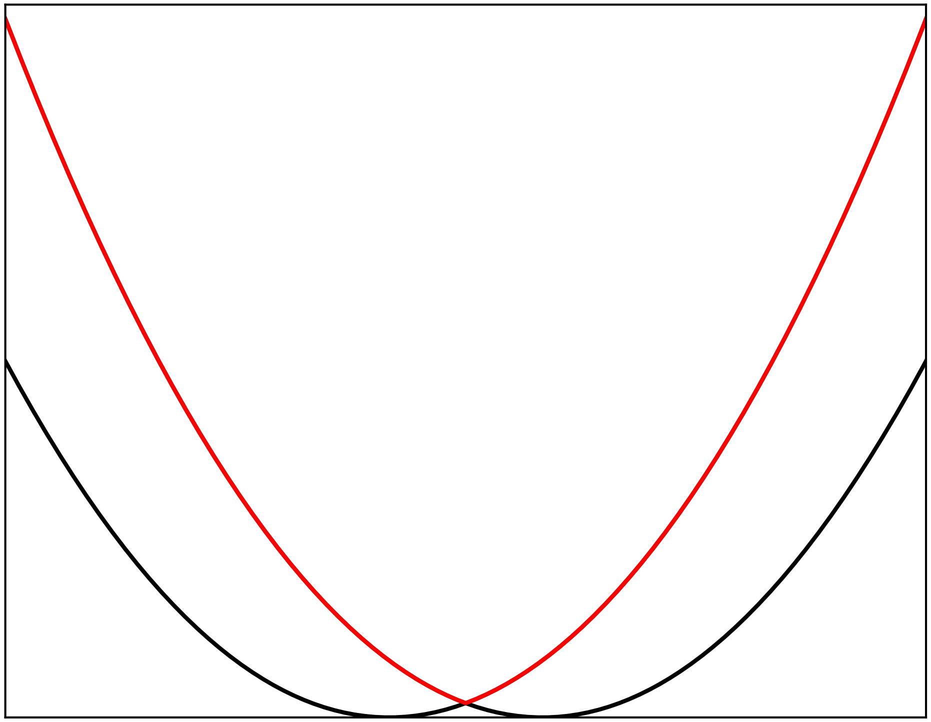
[Ben-Tal et al. 13, Bertsimas et al. 16, Lam & Zhou 17, Lam 18]

Optimization

$$\widehat{\theta}_n^{\mathrm{rob}} = \operatorname*{argmin}_{\theta \in \Theta} \max_{Q \in \mathcal{P}_n(\rho)} \sum_{i=1}^{n} q_i \ell(\theta; Z_i)$$

Nice properties

- Convex if loss is convex
- Conic forms [Ben-Tal et al. 13]
- Gradient descent [N. & Duchi 18]
- SGD on the dual



Outline

- Understanding DRO
  - DRO = worst-case subpopulation performance
- Trade-off: robustness vs. convergence
- Experiments
- Extensions: covariate shift

First idea: pre-defined groups

Given pre-defined demographic groups $g \in \mathcal{G}$,

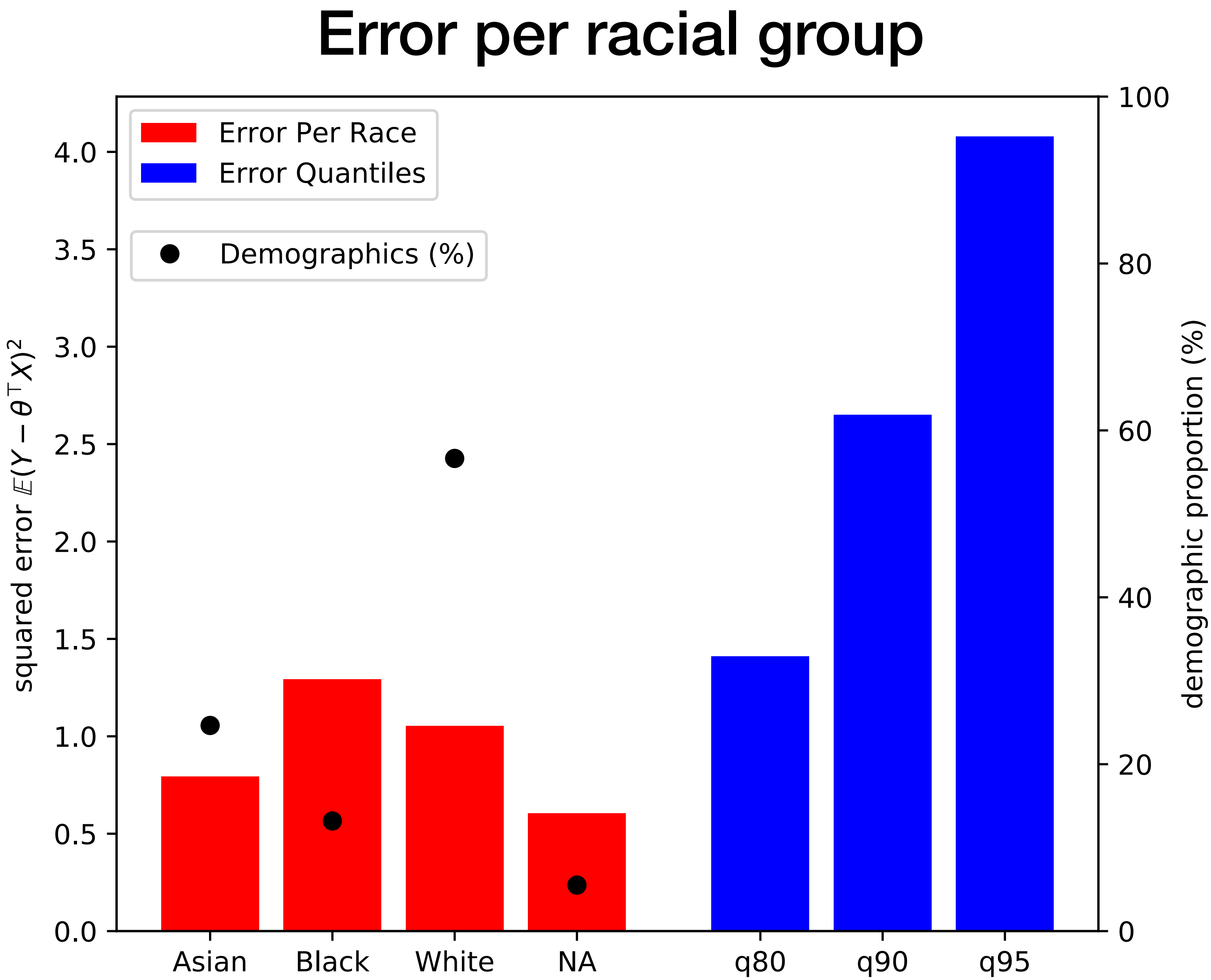
- Separate model for **each** group $\mathbb{E}_{P_g}[\ell(\theta_g; Z)]$
- One model for **worst-off** group $\max_{g \in \mathcal{G}} \mathbb{E}_{P_g}[\ell(\theta; Z)]$ [Meinshausen & Buhlmann '15]

See also [Kearns et al. '18, Kim et al. '19]

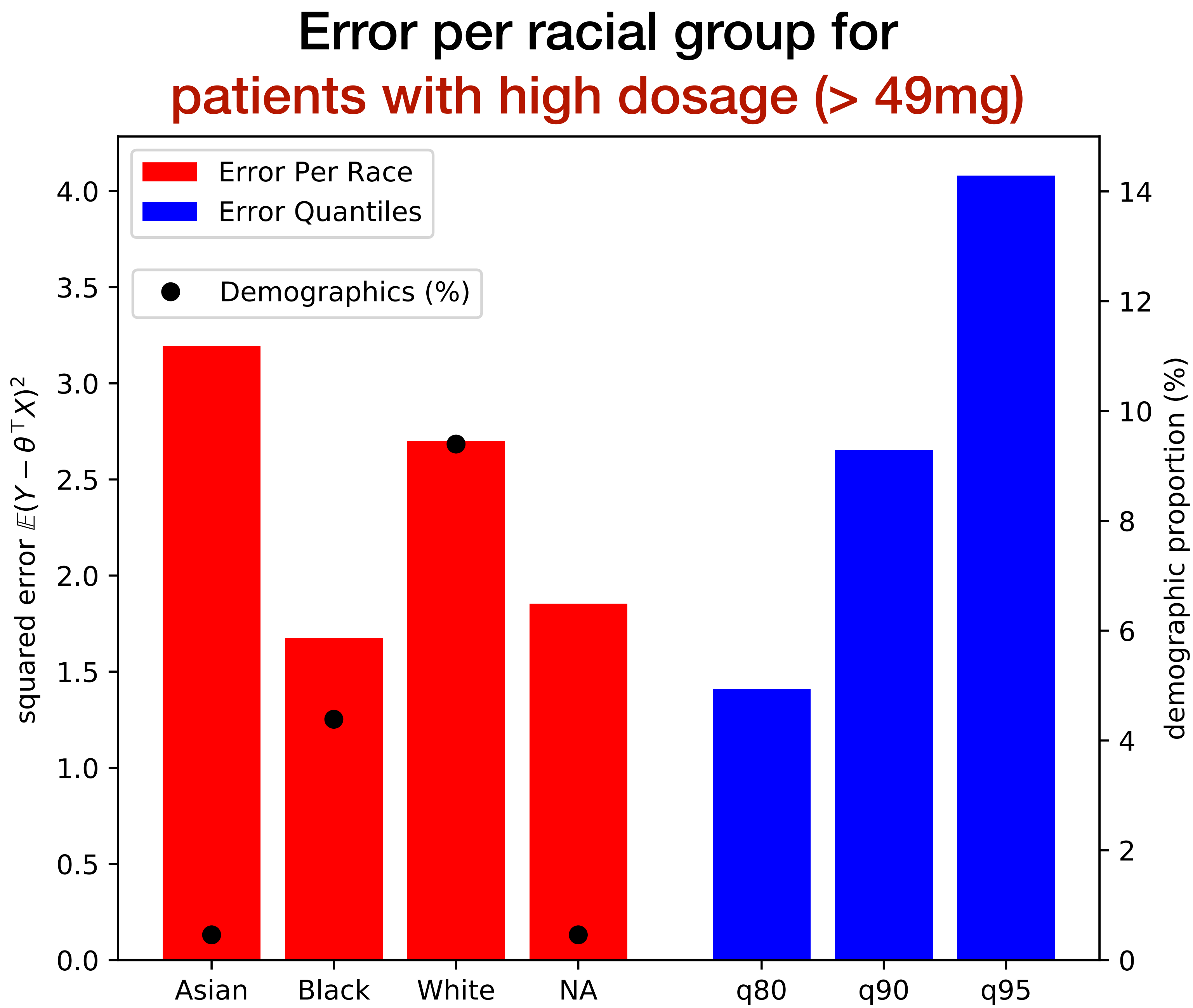
Problems

- In some applications, demographic information is **unavailable** (e.g. speech recognition), or **illegal** to use (e.g. insurance)
- Protected groups are **hard to define** a priori
  - variables often comprise continuous spectrum
  - performance determined in an **intersectional** fashion
- Accounting for intersections gives **exponentially many subgroups**
  - computational & statistical difficulties

Example: Predicting warfarin dosage



Example: Predicting warfarin dosage



Protecting against large shifts

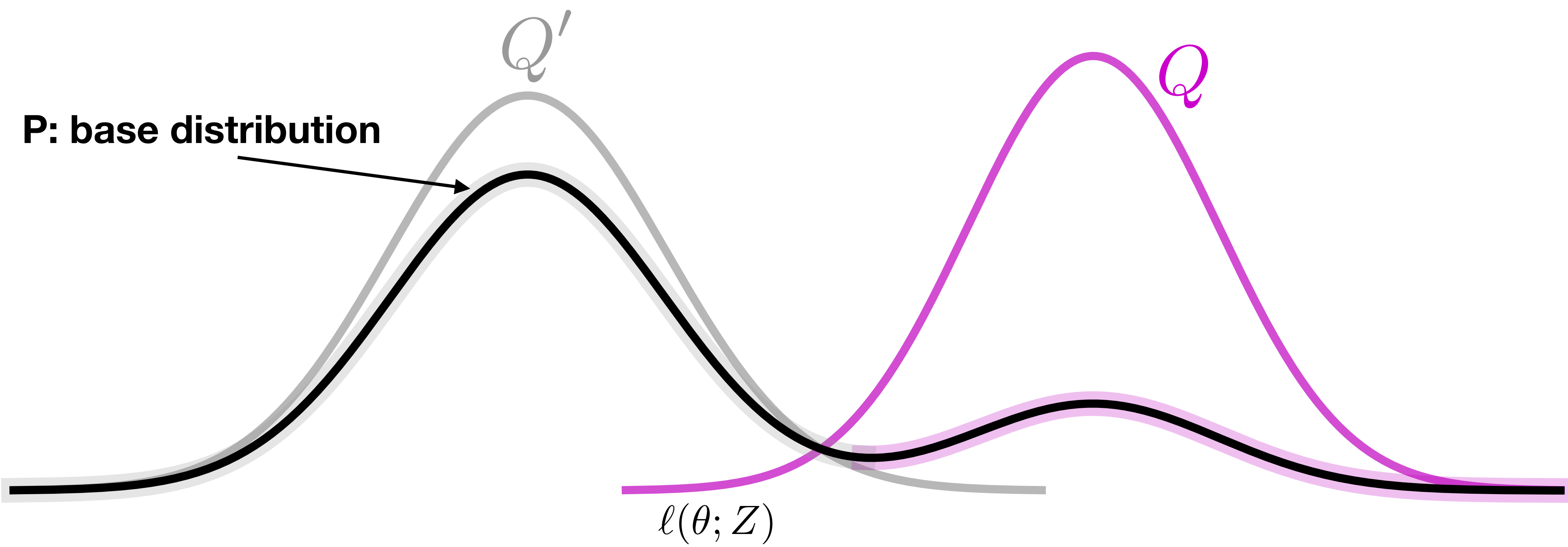
Automatically find worst-off subpopulations, and optimize performance on them

$$\underset{\theta \in \Theta}{\text{minimize}} \quad \max_{Q : D_f(Q \| P_{\text{obs}}) \le \rho} \mathbb{E}_Q[\ell(\theta; Z)]$$

Subpopulations

- Q is a **subpopulation** of P if it's a mixture component

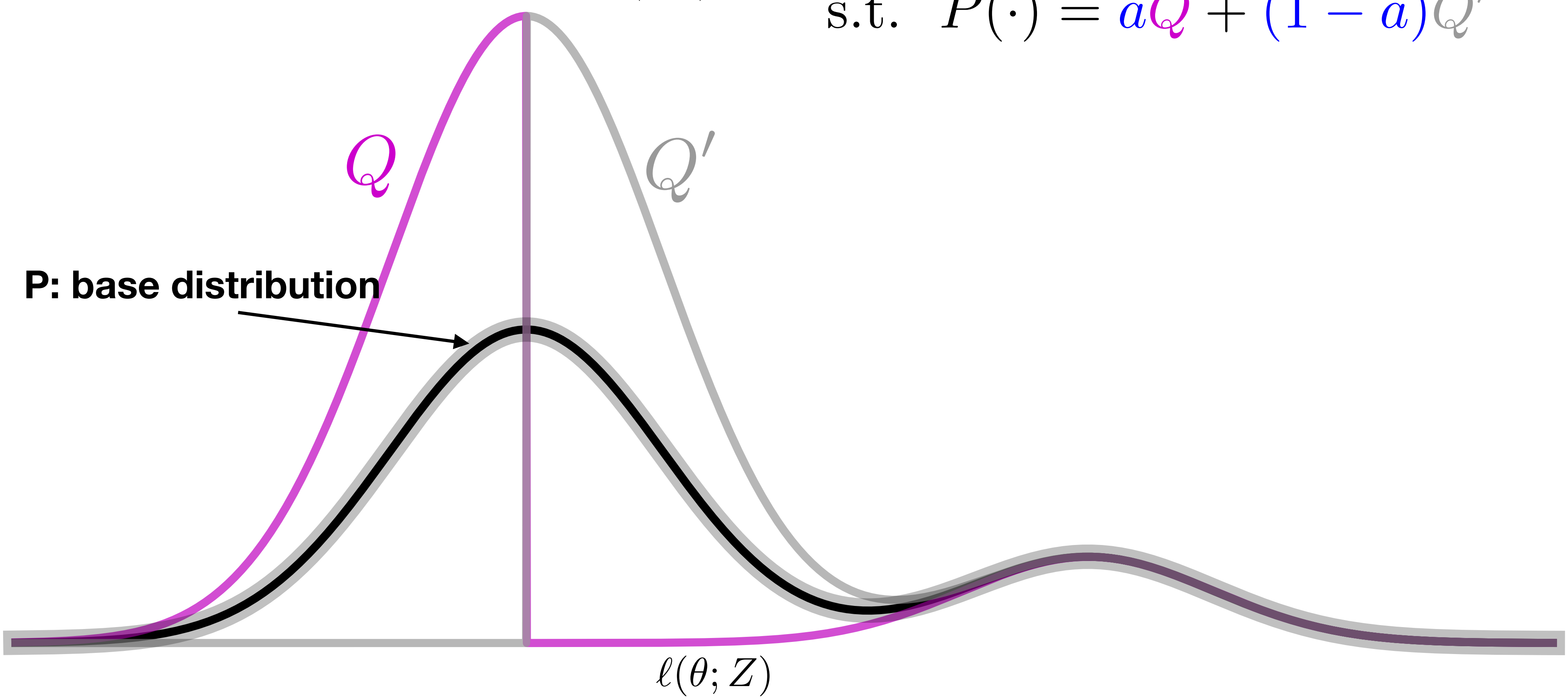
Q is a **subpopulation** $\iff$ $\exists \text{proportion } a \in (0,1], \text{ prob. } Q'$ s.t. $P(\cdot) = aQ + (1-a)Q'$



Subpopulations

- Q is a **subpopulation** of P if it's a mixture component

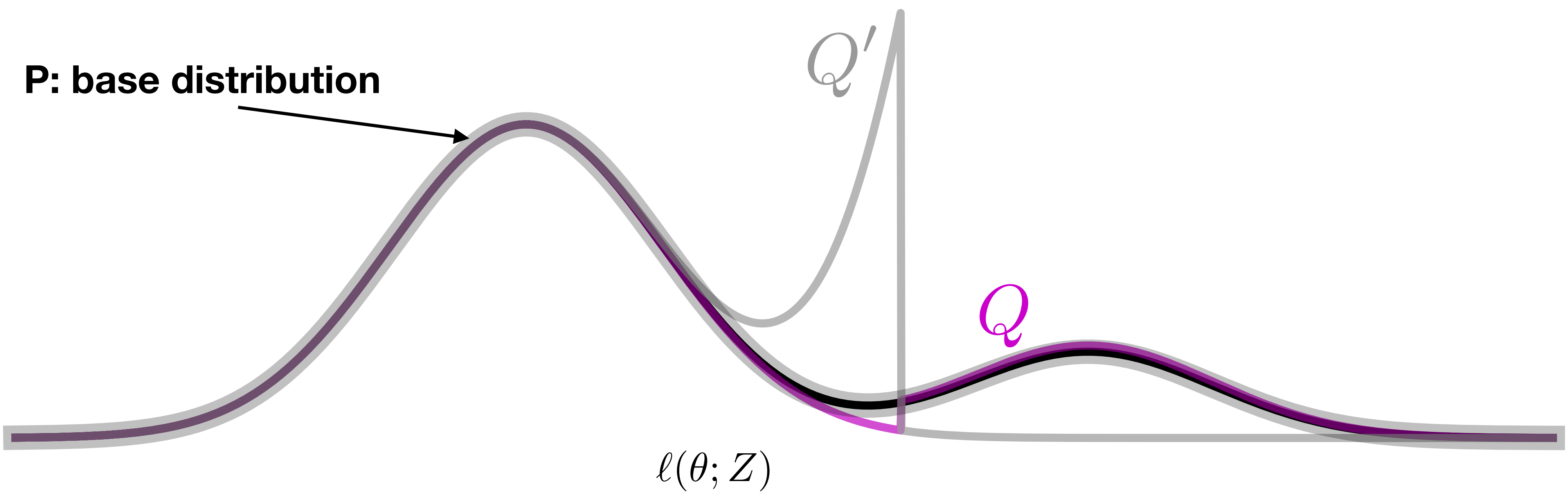
Q is a **subpopulation** $\Longleftrightarrow$ $\exists$proportion $a \in (0, 1]$, prob. Q' s.t. $P(\cdot) = aQ + (1-a)Q'$



Subpopulations

- Q is a **subpopulation** of P if it's a mixture component

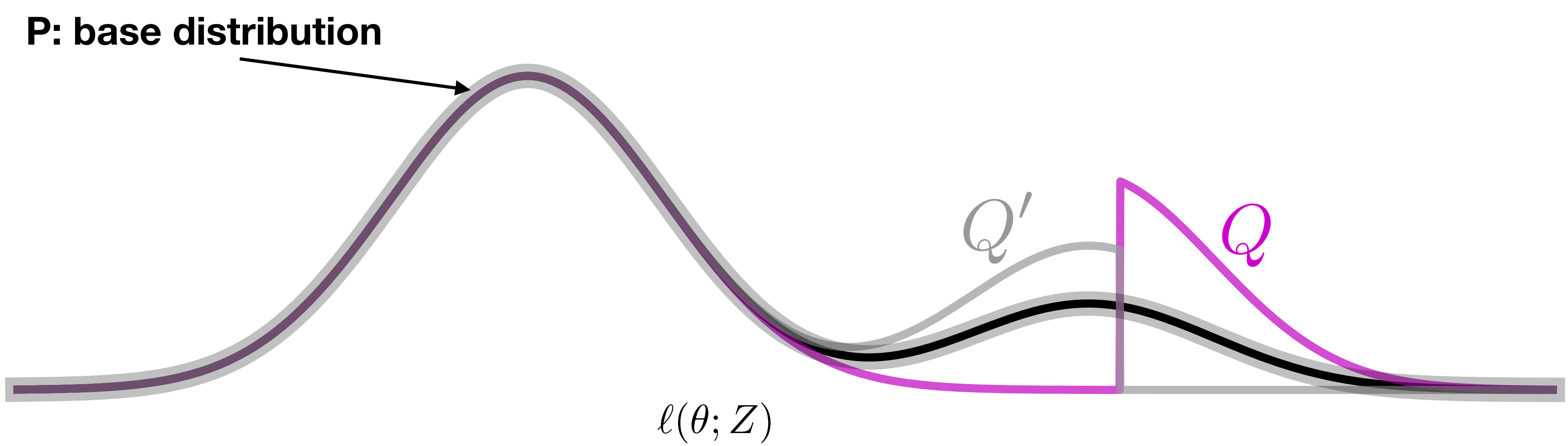
Q is a **subpopulation** $\Longleftrightarrow$ $\exists \text{proportion } a \in (0,1], \text{ prob. } Q'$ s.t. $P(\cdot) = aQ + (1-a)Q'$



Subpopulations

- Q is a **subpopulation** of P if it's a mixture component

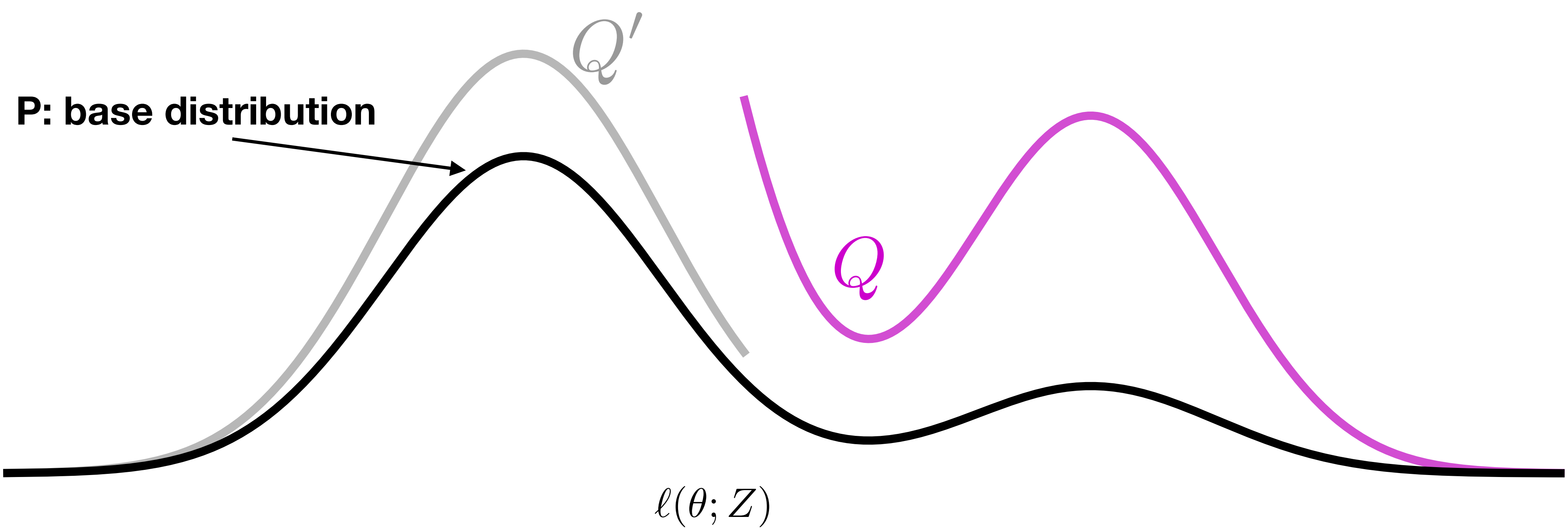
Q is a **subpopulation** $\Longleftrightarrow$ $\exists$proportion $a \in (0, 1]$, prob. Q' s.t. $P(\cdot) = aQ + (1-a)Q'$



Subpopulations

- Q is a **subpopulation** of P if it's a mixture component

Q is a **subpopulation** $\Longleftrightarrow$ $\exists \text{proportion } a \in (0,1], \text{ prob. } Q'$ s.t. $P(\cdot) = aQ + (1-a)Q'$



Subpopulations

- Q is a **subpopulation** of P if it's a mixture component

Q is a **subpopulation** ⟷ $\exists \text{proportion } a \in (0,1], \text{ prob. } Q'$ s.t. $P(\cdot) = aQ + (1-a)Q'$

Notation

$$Q \succeq \alpha \iff \left\{ Q : \begin{array}{l} \exists \text{probability } Q', \text{ and } a \geq \alpha \\ \text{s.t. } P = aQ + (1-a)Q' \end{array} \right\}$$

subpopulation with **proportion** larger than $\alpha \in (0,1]$

Random minority proportions

- Worst-case loss over **subpopulations** larger than $\alpha \in (0, 1]$

$$\sup_{Q \succeq \alpha} \mathbb{E}_Q[\ell(\theta; Z)]$$

- Let $A \sim P_A$ be a random minority proportion
- Take another worst-case over $P_A \in \mathcal{P}_A$

worst-case over **subpopulation** larger than $A \in (0, 1]$

$$\sup_{P_A \in \mathcal{P}_A} \mathbb{E}_{A \sim P_A} \left[\sup_{Q \succeq A} \mathbb{E}_Q[\ell(\theta; Z)] \right]$$

worst-case over **probability P_A on minority proportion A**

DRO = worst-case subpopulations

Let $\mathcal{P}$ be a convex set of probability distributions.

Lemma: There is $\mathcal{P}_A$, a set of probabilities on $(0, 1]$ s.t.

$$\sup_{Q \in \mathcal{P}} \mathbb{E}_Q[\ell(\theta; Z)] = \sup_{P_A \in \mathcal{P}_A} \mathbb{E}_{A \sim P_A} \left[\sup_{Q \succeq A} \mathbb{E}_Q[\ell(\theta; Z)] \right]$$

See [Kusuoka 01, Pflug and Romisch 07]

DRO optimizes worst-case subpopulation loss!

Back to f-divergences

$$f_k(t) = (k(k-1))^{-1}(t^k - 1) \text{ for } k \in (1, \infty)$$

Lemma: f-div DRO optimizes worst-case subpopulation

$$\sup_{Q : D_{f_k}(Q \| P_{\text{obs}}) \le \rho} \mathbb{E}_Q[\ell(\theta; Z)] = \sup_{P_A \in \mathcal{P}_{A,k,\rho}} \mathbb{E}_{A \sim P_A}\left[\sup_{Q \succeq A} \mathbb{E}_Q[\ell(\theta; Z)]\right]$$

where $\alpha_k(\rho)^{-1} := (1 + k(k-1)\rho)^{1/k}$, and

$$\mathcal{P}_{A,k,\rho} := \left\{ \text{Set of random minority proportions lower bounded by } \alpha_k(\rho) \right\}$$

See also [Dentcheva 10]

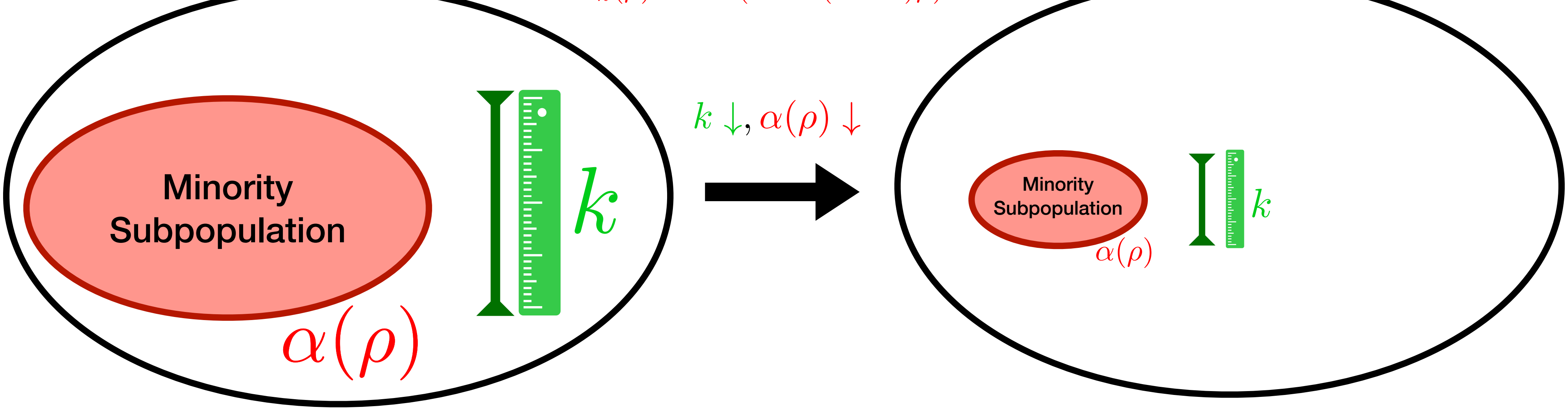
Back to f-divergences

$$f_k(t) = (k(k-1))^{-1}(t^k - 1) \text{ for } k \in (1, \infty)$$

$$\operatorname*{minimize}_{\theta\in\Theta}\left\{\sup_{Q:D_{f_k}(Q\|P_{\mathrm{obs}})\le\rho}\mathbb{E}_Q[\ell(\theta;Z)] = \sup_{P_A\in\mathcal{P}_{A,k,\rho}}\mathbb{E}_{A\sim P_A}\left[\sup_{Q\succeq A}\mathbb{E}_Q[\ell(\theta;Z)]\right]\right\}$$

Less robust | **More robust**

$$\alpha_k(\rho)^{-1} := (1 + k(k-1)\rho)^{1/k}$$



- Heuristically, tune k and $\alpha(\rho)$ on some preliminary subpopulation

A principle: minimax

1. We choose procedure $\widehat{\theta}$, nature chooses $P_{\rm obs}$
2. Receive data Z_i i.i.d. from $P_{\rm obs}$, $\widehat{\theta}$ makes decision

Define $\mathcal{R}_{k,\rho}(\theta; P) := \sup_{Q : D_{f_k}(Q \| P) \le \rho} \mathbb{E}_Q[\ell(\theta; Z)]$

Minimax (excess) risk [Wald 39, von Neumann 28]:

$$\min_{\widehat{\theta}} \max_{P_{\rm obs} \in \mathcal{D}_{\rm obs}} \left\{ \mathbb{E}_{P_{\rm obs}}[\mathcal{R}_{k,\rho}(\widehat{\theta}(Z_1^n); P_{\rm obs})] - \min_{\theta \in \Theta} \mathcal{R}_{k,\rho}(\theta; P_{\rm obs}) \right\}$$

Worst case over distributions $\mathcal{D}_{\rm obs}$

Best case over procedures $\widehat{\theta} : \mathcal{Z}^n \to \Theta$

Main result

Theorem (Duchi & Namkoong '20)

$$\min_{\widehat{\theta}} \max_{P_{\rm obs} \in \mathcal{D}_{\rm obs}} \left\{ \mathbb{E}_{P_{\rm obs}}[\mathcal{R}_{k,\rho}(\widehat{\theta}(Z_1^n); P_{\rm obs})] - \min_{\theta \in \Theta} \mathcal{R}_{k,\rho}(\theta; P_{\rm obs}) \right\} \approx n^{-\frac{1}{k_* \vee 2}}$$

where $k_* = k/(k-1)$.

$k \in [2, \infty)$: parametric

$k \in (1, 2)$: slower

Worst case over distributions $\mathcal{D}_{\rm obs}$

Best case over procedures $\widehat{\theta} : \mathcal{Z}^n \to \Theta$

Two pronged approach

1. Convergence guarantee: find good procedure
2. Lower bound: show no procedure can do better

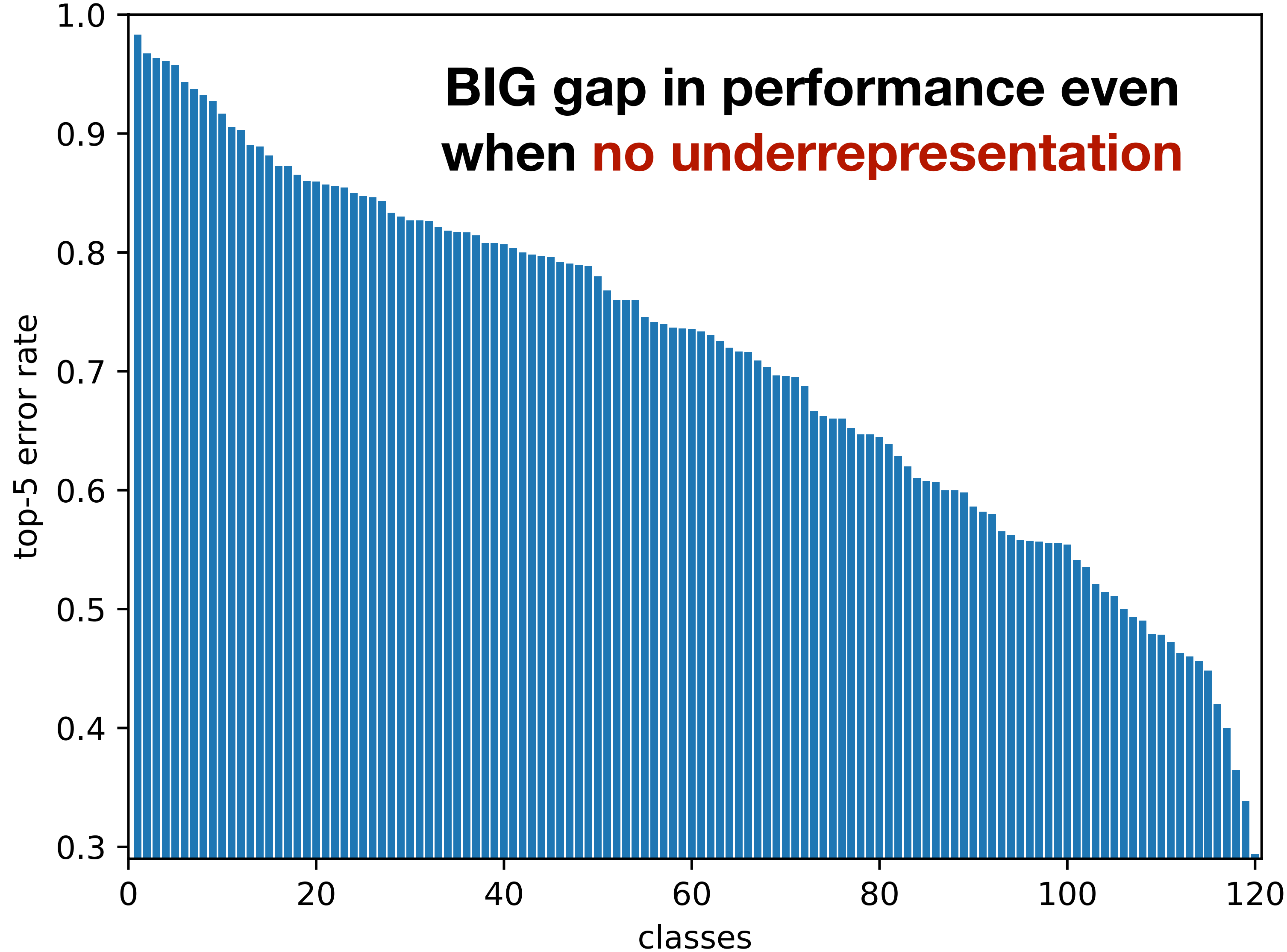
Fine-grained recognition

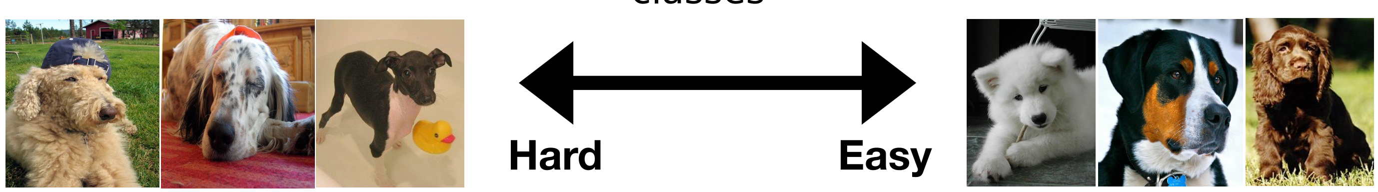
- Task: classify image of dog to breed (120 classes)
- Kernel features

Stanford Dogs Dataset [Khosla et al. '11]

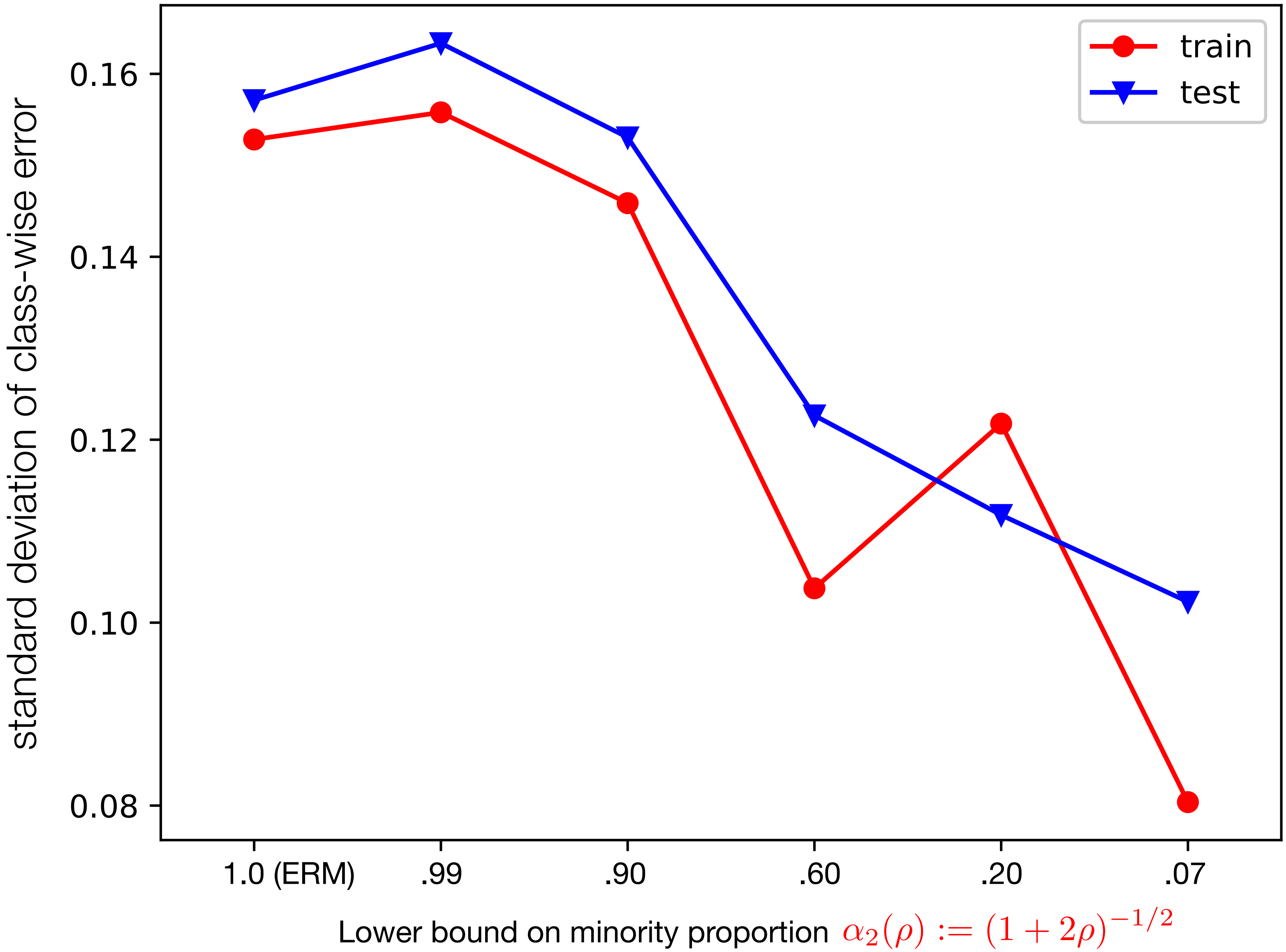
No underrepresentation:
same number of images per class

ERM error rate



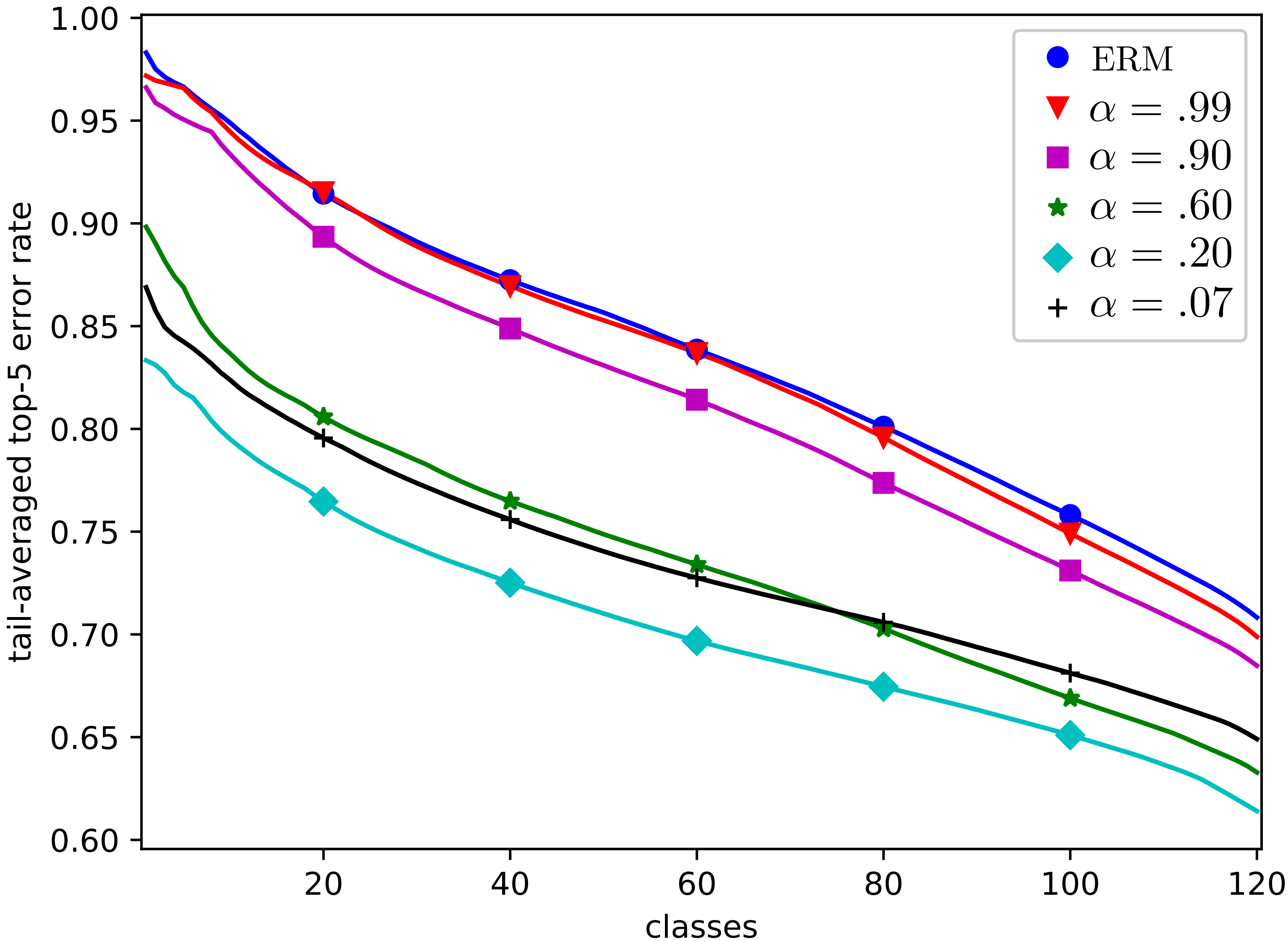


Variation in error over 120 class



$$f_2(t) \approx t^2 - 1$$

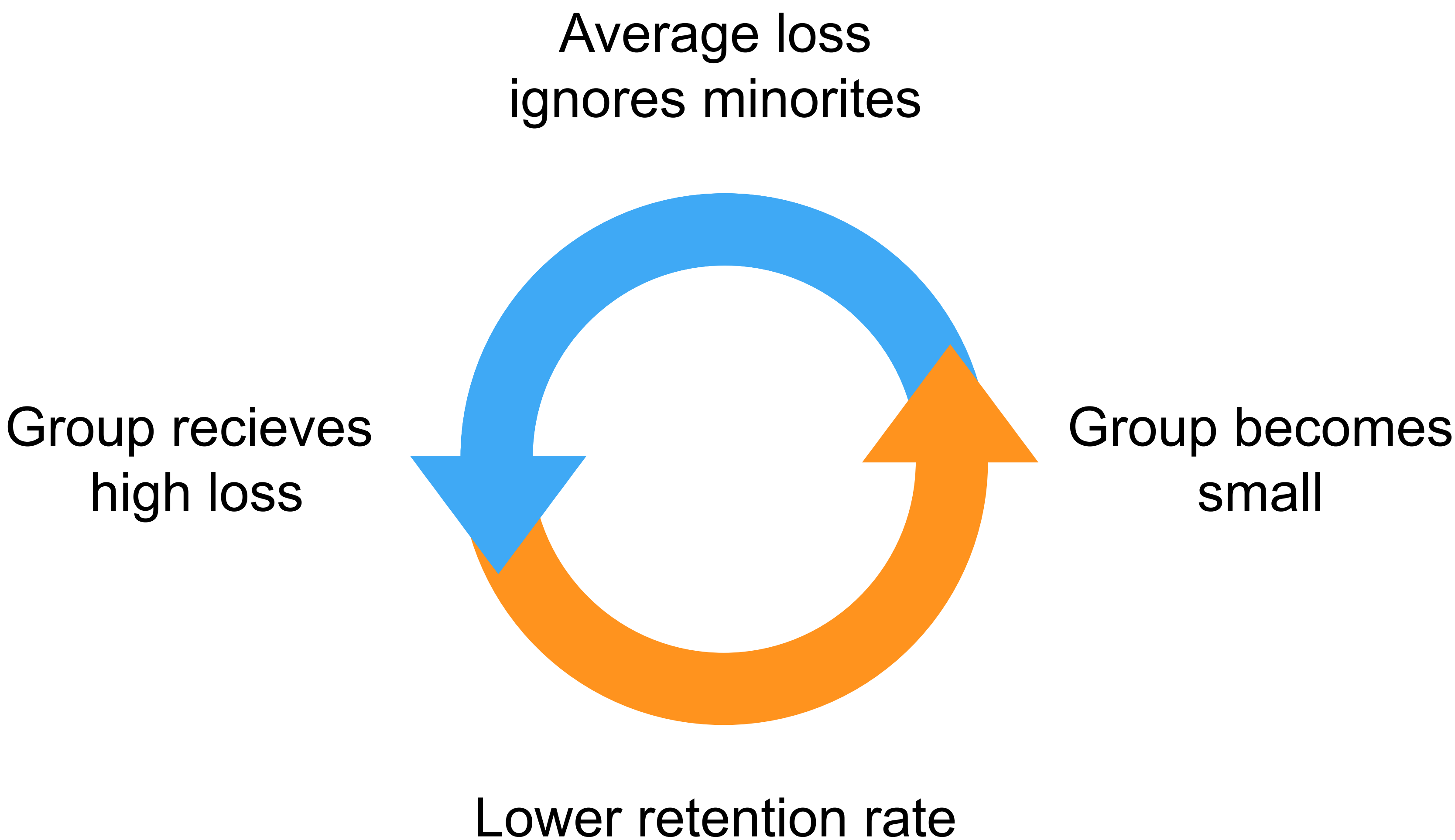
Worst x-classes



Lower bound on minority proportion $\alpha_2(\rho) := (1 + 2\rho)^{-1/2}$

Takeaway: Guarantee uniform performance across dog breeds

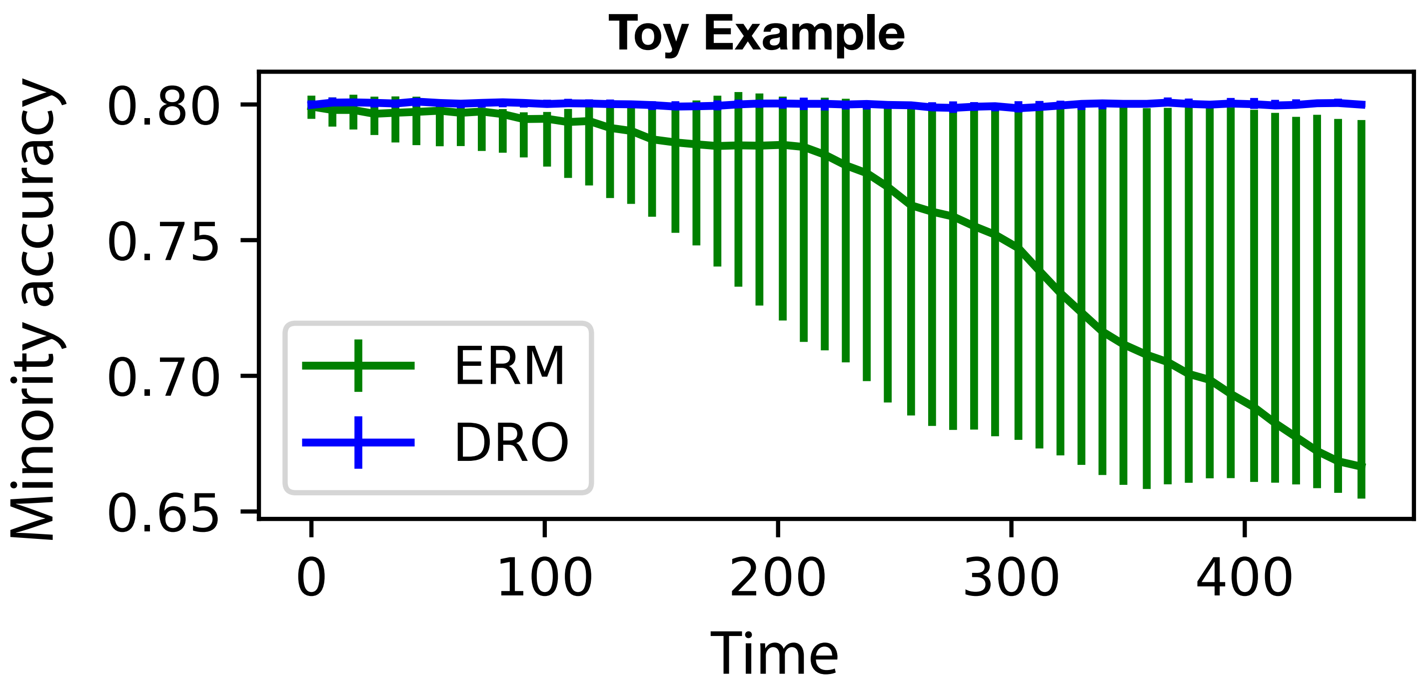
Repeated loss minimization



Problem: Degradation over time

Problem: Degradation over time

Small disparities can amplify to exacerbate subpopulation performance



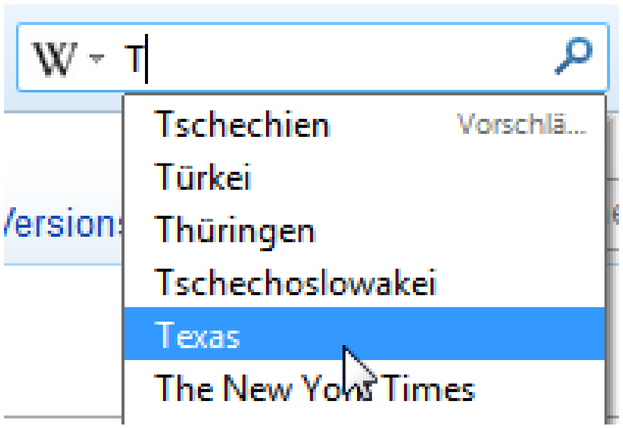
"Theorem" (HSN**L'18)** Under general user retention dynamics,

1) ERM is unstable

2) minimizing $\mathcal{R}_{p,\alpha}(\theta; P^t_{\text{obs}})$ controls latent minority proportions over time

Experiment: Auto-complete

Motivation: Autocomplete system for text



Problem: Atypical text doesn't get surfaced

African American Vernacular (AAVE)

If u wit me den u pose to RESPECT ME

Standard American English (SAE)

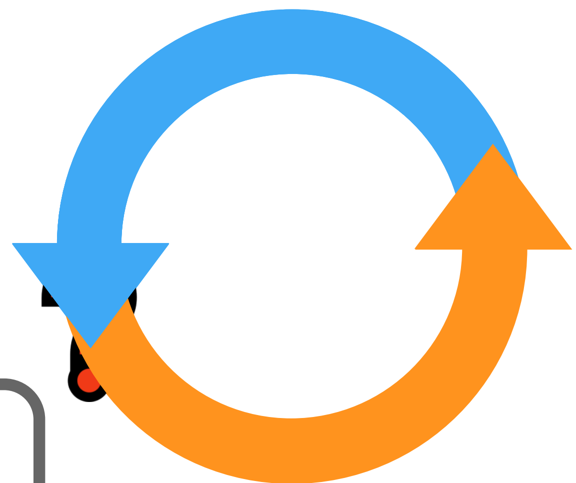
If you are with me then you are supposed to respect me.

Experiment: Auto-complete

Retention feedback loop

Average loss
ignores **AAVE**

AAVE group
receives high loss

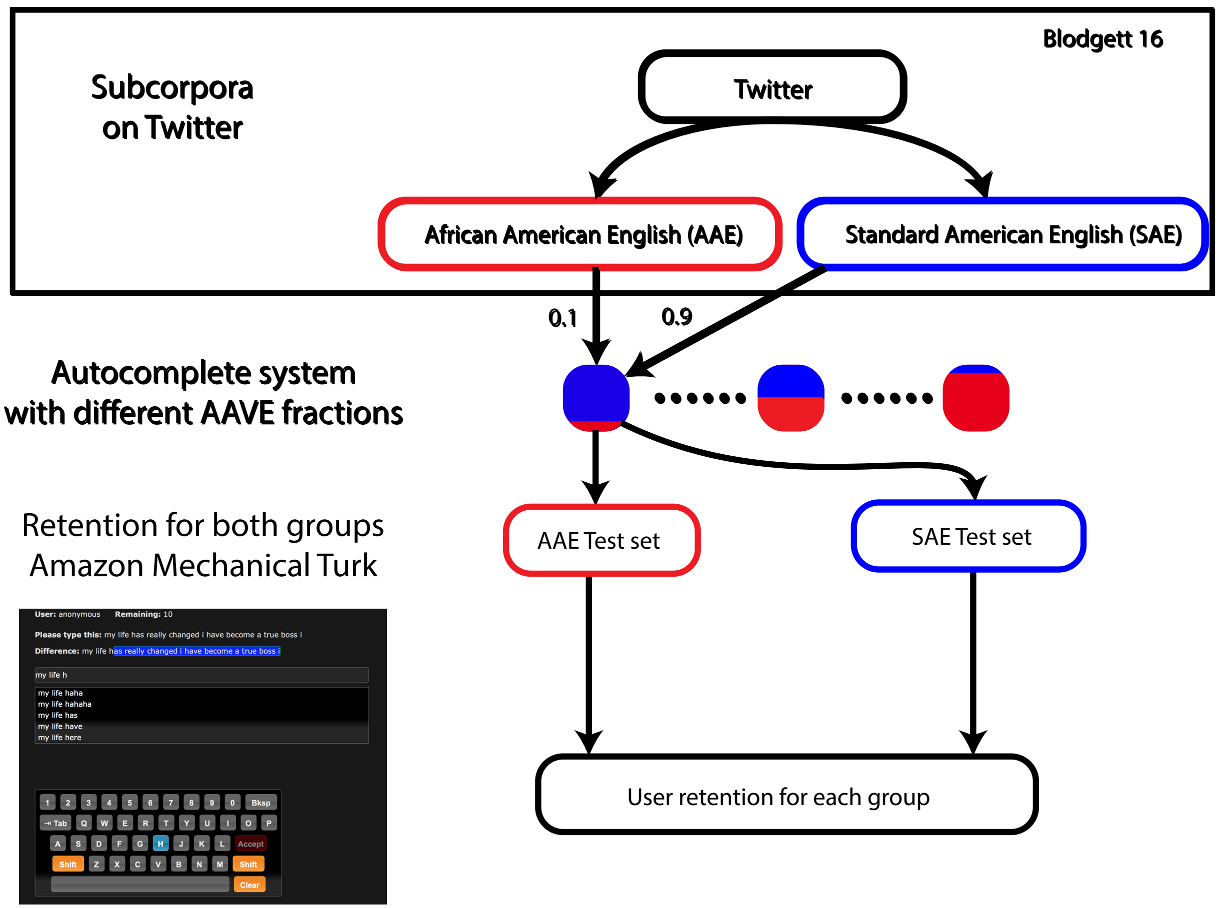


AAVE group
becomes smaller

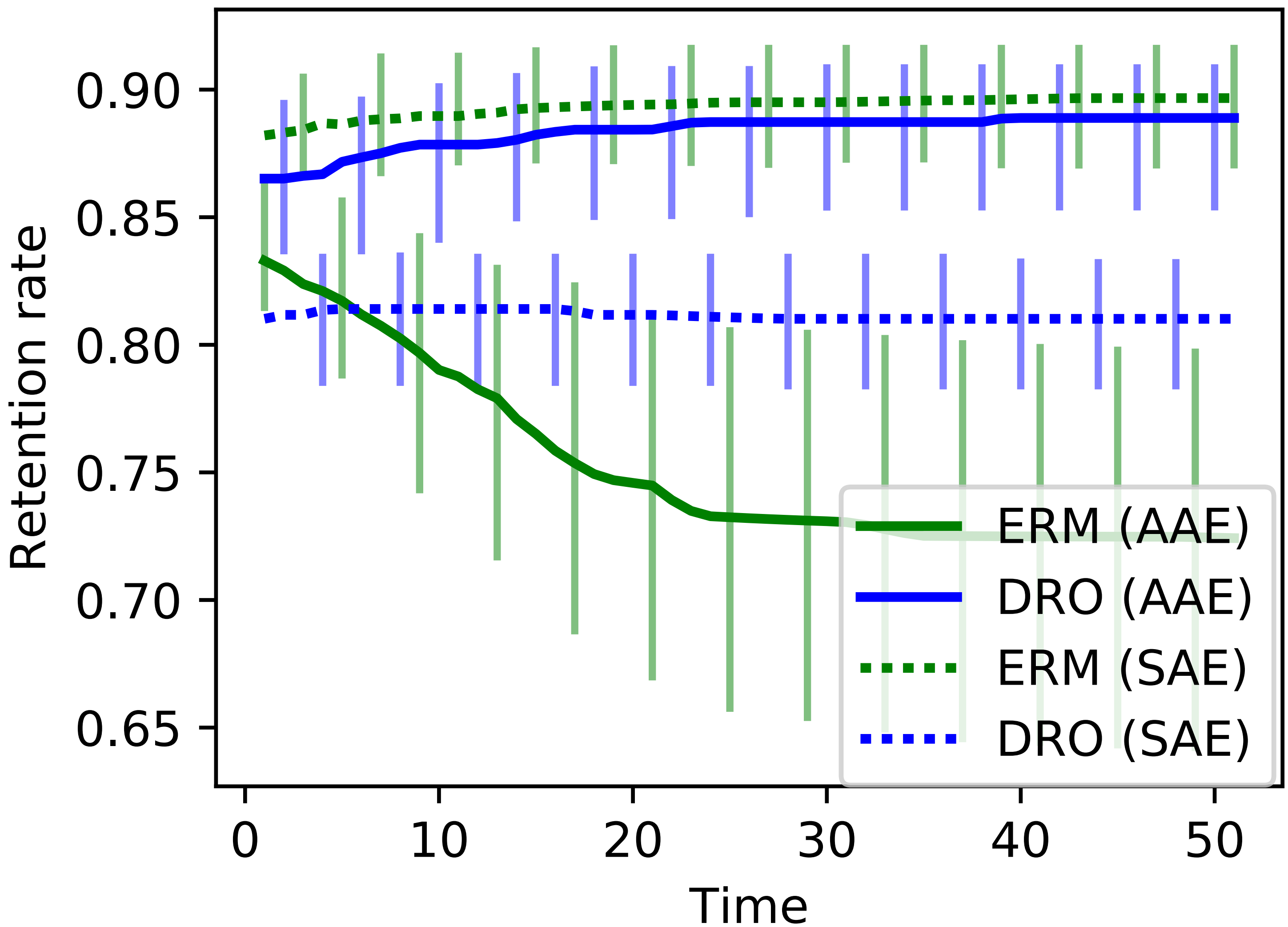
Lower retention rate

How does loss relate
to retention rate?

Experiment: Auto-complete



Mitigating Disparity Amplification



Takeaway: **Control minority proportion → uniform performance over time**

Covariate shift

- Conditional distribution $P_{Y|X}$ **fixed**
- Only consider **subpopulations** of marginal P_X

Notation

$$Q_X \succeq \alpha \quad \Longleftrightarrow \quad \left\{ Q_X : \begin{array}{l} \exists \text{probability } Q'_X, \text{ and } a \geq \alpha \\ \text{s.t. } P_X = aQ_X + (1-a)Q'_X \end{array} \right\}$$

subpopulation over X with **proportion** larger than $\alpha \in (0, 1]$

$$\sup_{Q_X \succeq \alpha} \left\{ \mathbb{E}_{Q_X \times P_{Y|X}}[\ell(\theta; X, Y)] = \mathbb{E}_{Q_X}[\ell_c(\theta; X)] \right\}$$

$$\ell_c(\theta; X) := \mathbb{E}_{P_{Y|X}}[\ell(\theta; X, Y) \mid X]$$

Covariate shift

Standard approach: Solve average risk minimization problem

$$\underset{\theta \in \Theta}{\text{minimize}} \ \mathbb{E}_{P_{\text{obs}}}[\ell(\theta; X, Y)]$$

DRO over covariate shift

$$\underset{\theta \in \Theta}{\text{minimize}} \ \sup_{Q_X \succeq \alpha} \mathbb{E}_{Q_X}[\ell_c(\theta; X)]$$

worst-case loss over **subpopulations in X** larger than $\alpha \in (0, 1]$

Problem: We don't observe $\ell_c(\theta; X) := \mathbb{E}_{P_{Y|X}}[\ell(\theta; X, Y) \mid X]$!

Hard to estimate because of limited replicate labels $Y|X$

Dual representation

Lemma (Duchi, Hashimoto & N '19)

Let $\ell_c(\theta; X) := \mathbb{E}_{P_{Y|X}}[\ell(\theta; X, Y) \mid X]$.

$$\sup_{Q_X \succeq \alpha} \mathbb{E}_{Q_X}[\ell_c(\theta; X)] = \inf_{\eta} \left\{ \frac{1}{\alpha} \mathbb{E}_{P_X} \left(\ell_c(\theta; X) - \eta\right)_+ + \eta \right\}$$

Only care about X with conditional risk worse than η

For any $k, k_* > 1$ such that $1/k + 1/k_* = 1$

$$\begin{aligned} \mathbb{E}_{P_X} \left(\ell_c(\theta; X) - \eta\right)_+ &\leq \left(\mathbb{E}_{P_X} \left(\ell_c(\theta; X) - \eta\right)_+^{k_*}\right)^{1/k_*} \\ &= \sup_{h \geq 0, \mathbb{E}[h(X)^k] \leq 1} \mathbb{E}[h(X)(\ell(\theta; X, Y) - \eta)] \end{aligned}$$

Variational form

Lemma (Duchi, Hashimoto & N '19)

If $x \mapsto \ell_c(\theta; x)$, and $(x, y) \mapsto \ell(\theta; x, y)$ are L-Lipschitz,

$$\left(\mathbb{E}_{P_X}\left(\ell_c(\theta; X) - \eta\right)_+^{k_*}\right)^{1/k_*} = \sup_{h \geq 0, \mathbb{E}[h(X)^k] \leq 1, \mathrm{O}(L)\text{-smooth}} \mathbb{E}[h(X)(\ell(\theta; X, Y) - \eta)]$$

for any $k, k_* > 1$ such that $1/k + 1/k_* = 1$

Estimable bound

$$\sup_{Q_X \succeq \alpha} \mathbb{E}_{Q_X}[\ell_c(\theta; X)] \leq \inf_{\eta} \left\{ \frac{1}{\alpha} \sup_{h \geq 0, \mathbb{E}[h(X)^k] \leq 1, \mathrm{O}(L)\text{-smooth}} \mathbb{E}[h(X)(\ell(\theta; X, Y) - \eta)] + \eta \right\}$$

Replaced $\ell_c(\theta; X) := \mathbb{E}_{P_{Y|X}}[\ell(\theta; X, Y) \mid X]$ with $\ell(\theta; X, Y)$

Estimator

Standard approach: Solve empirical risk minimization problem

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

Worst-case subpopulation approach: Optimize worst-case subpopulation performance

$$\underset{\theta \in \Theta, \eta}{\text{minimize}} \left\{ \frac{1}{\alpha} \sup_{h \geq 0, \frac{1}{n} \sum_{i=1}^{n} h(X_i)^k \leq 1, \text{O}(L)\text{-smooth}} \frac{1}{n} \sum_{i=1}^{n} h(X_i)(\ell(\theta; X_i, Y_i) - \eta)] + \eta \right\}$$

Can efficiently solve using dual version. See paper for details.

Semantic similarity

- Given two word vectors (GloVe), predict their semantic similarity [Agirre et al. '09]
- Per word pair, there are 13 human annotations on similarity in range {0, …, 10}
- Train on 1989 indiv. annotations, test on 246 averaged values

Similarity

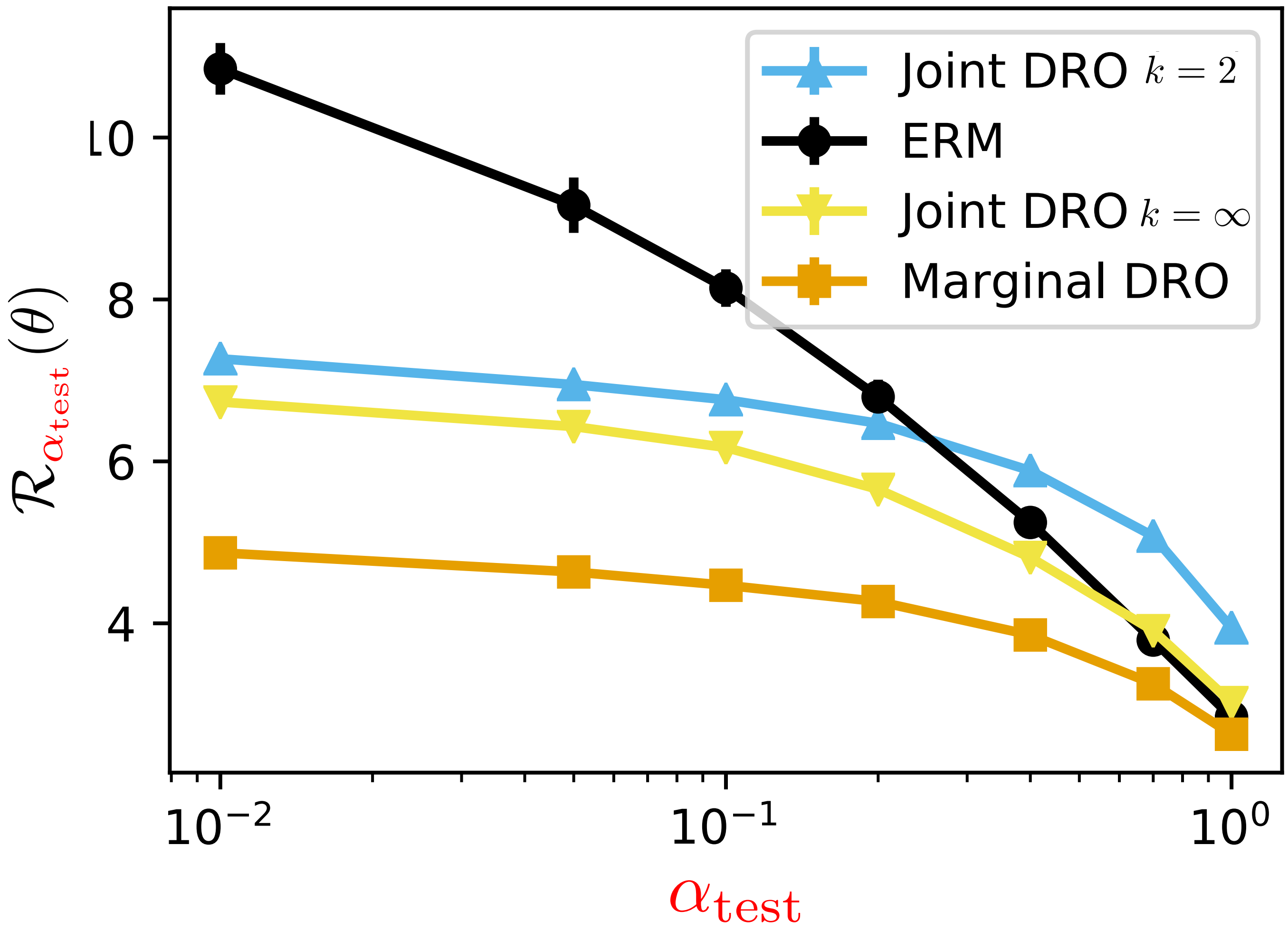
$$\ell(\theta; x^1, x^2, y) = |y - (x^1 - x^2)^\top \theta_1 (x^1 - x^2) - \theta_2|$$

Word 1 Word 2

- Fix train-time $\alpha = .3$, test on varying α_{test}

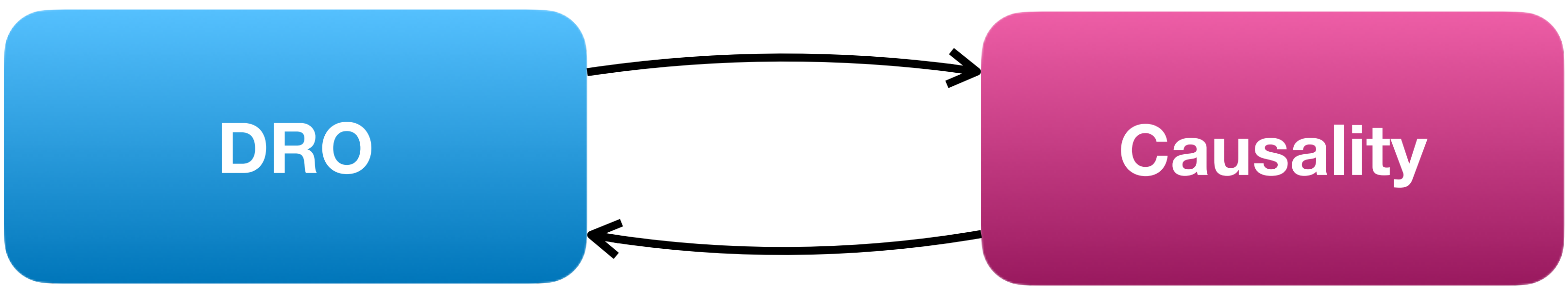
Semantic similarity

$$\mathcal{R}_{\alpha_{\text{test}}}(\theta) := \sup_{Q_X \succeq \alpha_{\text{test}}} \mathbb{E}_{Q_X \times P_{Y|X}}[\ell(\theta; X, Y)]$$



Endnote

- DRO = Worst-case subpopulation performance
- **The** question: choice of worst-case region



Duchi and Namkoong. Learning models with uniform performance via distributionally robust optimization. Forthcoming in Annals of Statistics, 2020.

Duchi, Hashimoto, and Namkoong. Distributionally robust losses against mixture covariate shifts. Under review, 2020.

Hashimoto, Srivastava, Namkoong, A. Sinha, and P. Liang. Fairness without demographics in repeated loss minimization. In International Conference on Machine Learning, 2018.