Critiques of classification parity

Link **https://arxiv.org/abs/1808.00023**

Recap

- COMPAS: risk scoring system predicting recidivism
  - Y: observed reoffend, X: 20-dim feature based on questionnaire

- ProPublica: COMPAS has different false positive rates P(predicted high risk | not reoffend), and FNR across Blacks and Whites

- Northepoint: but COMPAS has similar predictive value P(reoffend | predicted high risk)

- Chouldechova: impossible to satisfy these simultaneously

General view

	Predicted low risk	**Predicted high risk**
Did not re-offend	True Negative	False Positive
Re-offended	False Negative	True Positive

Perspectives matter

- Viewpoints vary substantially between stakeholders
- **Defendant**: what is the probability I'll be wrongly labeled high-risk?

	Predicted low risk	Predicted high risk
Did not re-offend	True Negative	False Positive
Re-offended	False Negative	True Positive

FPR = FP / (FP + TN)

Perspectives matter

- Viewpoints vary substantially between stakeholders
- **Decision-maker**: of those I've predicted high-risk, what fraction will re-offend?

	Predicted low risk	**Predicted high risk**
Did not re-offend	True Negative	False Positive
Re-offended	False Negative	True Positive

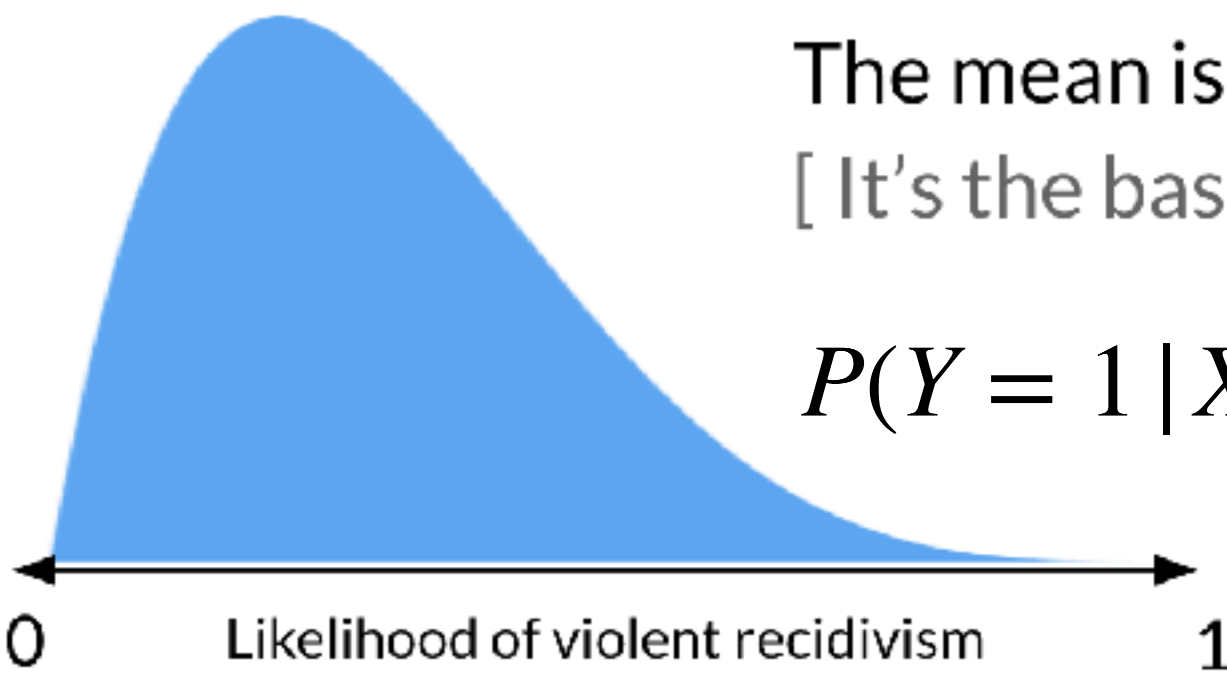
Predictive value (→ Predicted high risk column)

Classification parity

- Equalize FPR, FNR, PV across pre-defined demographic groups
- More generally, we can equalize any measure of performance

<table>
<tr><td></td><td></td><td colspan="2">True condition</td><td></td><td colspan="2"></td></tr>
<tr><td></td><td>Total population</td><td>Condition positive</td><td>Condition negative</td><td>Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$</td><td colspan="2">Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$</td></tr>
<tr><td rowspan="2">Predicted condition</td><td>Predicted condition positive</td><td>True positive</td><td>False positive, Type I error</td><td>Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$</td><td colspan="2">False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$</td></tr>
<tr><td>Predicted condition negative</td><td>False negative, Type II error</td><td>True negative</td><td>False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$</td><td colspan="2">Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$</td></tr>
<tr><td></td><td></td><td>True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$</td><td>False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$</td><td>Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$</td><td rowspan="2">Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$</td><td rowspan="2">$F_1$ score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$</td></tr>
<tr><td></td><td></td><td>False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$</td><td>Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$</td><td>Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$</td></tr>
</table>

Risk distributions



The mean is fixed for all choices of X

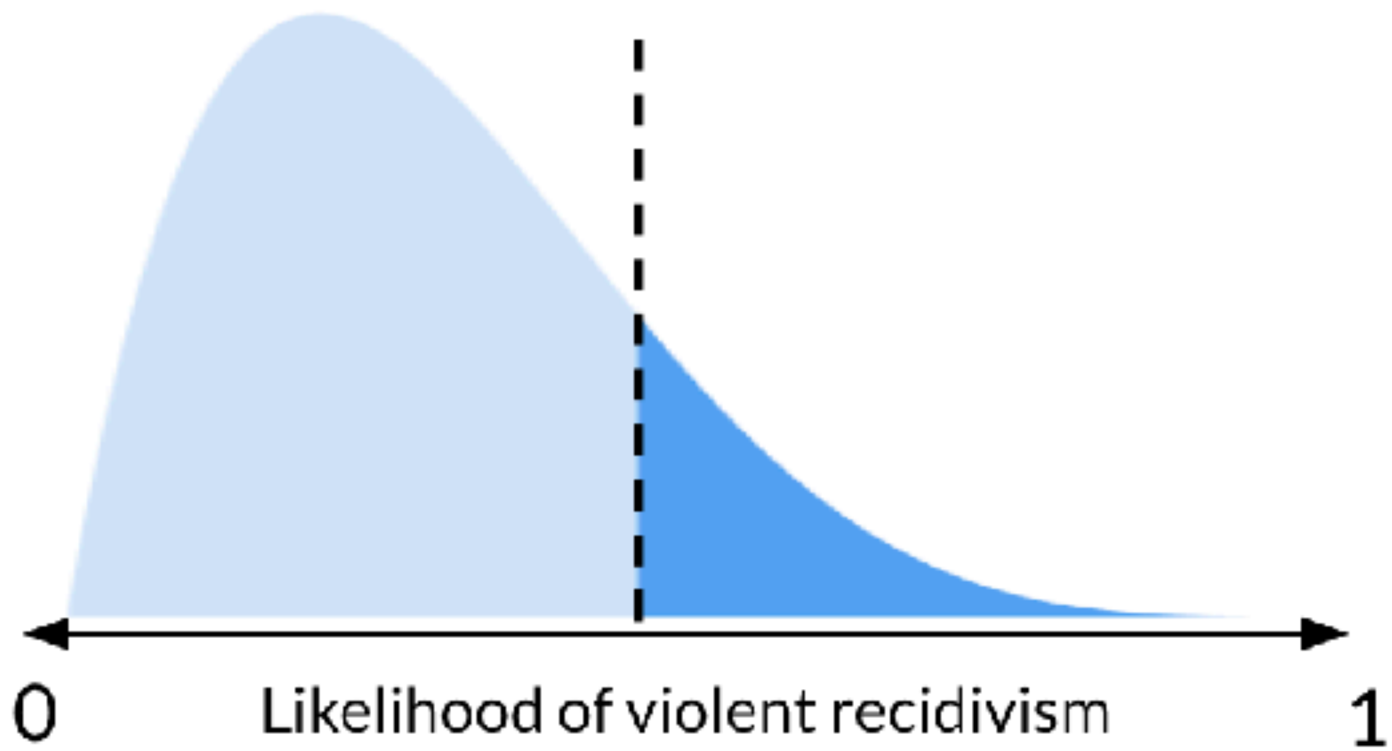
[It's the base rate of recidivism.]

$$P(Y = 1 \mid X)$$

The shape can change based on our choice of X

Slides by Sharad Goel

Applying a threshold



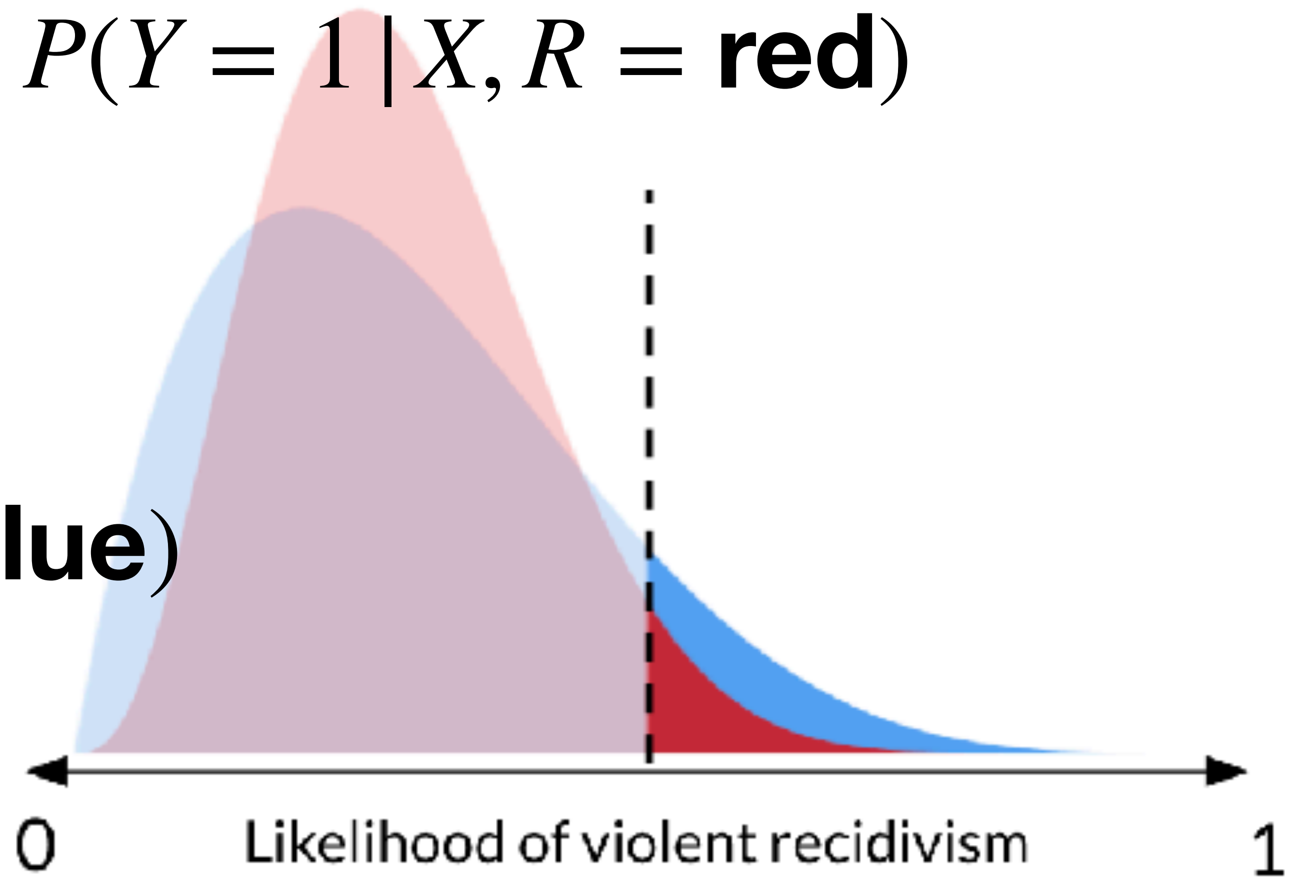
- Threshold rule maximizes social welfare, if errors are equally costly across individuals

Slides by Sharad Goel

Fairness of a single threshold

$P(Y = 1 \mid X, R = \textbf{red})$

$P(Y = 1 \mid X, R = \textbf{blue})$



Equally risky people are treated equally, regardless of group membership. No taste-based discrimination. Inline with legal norms. This is what is done in practice.

Slides by Sharad Goel

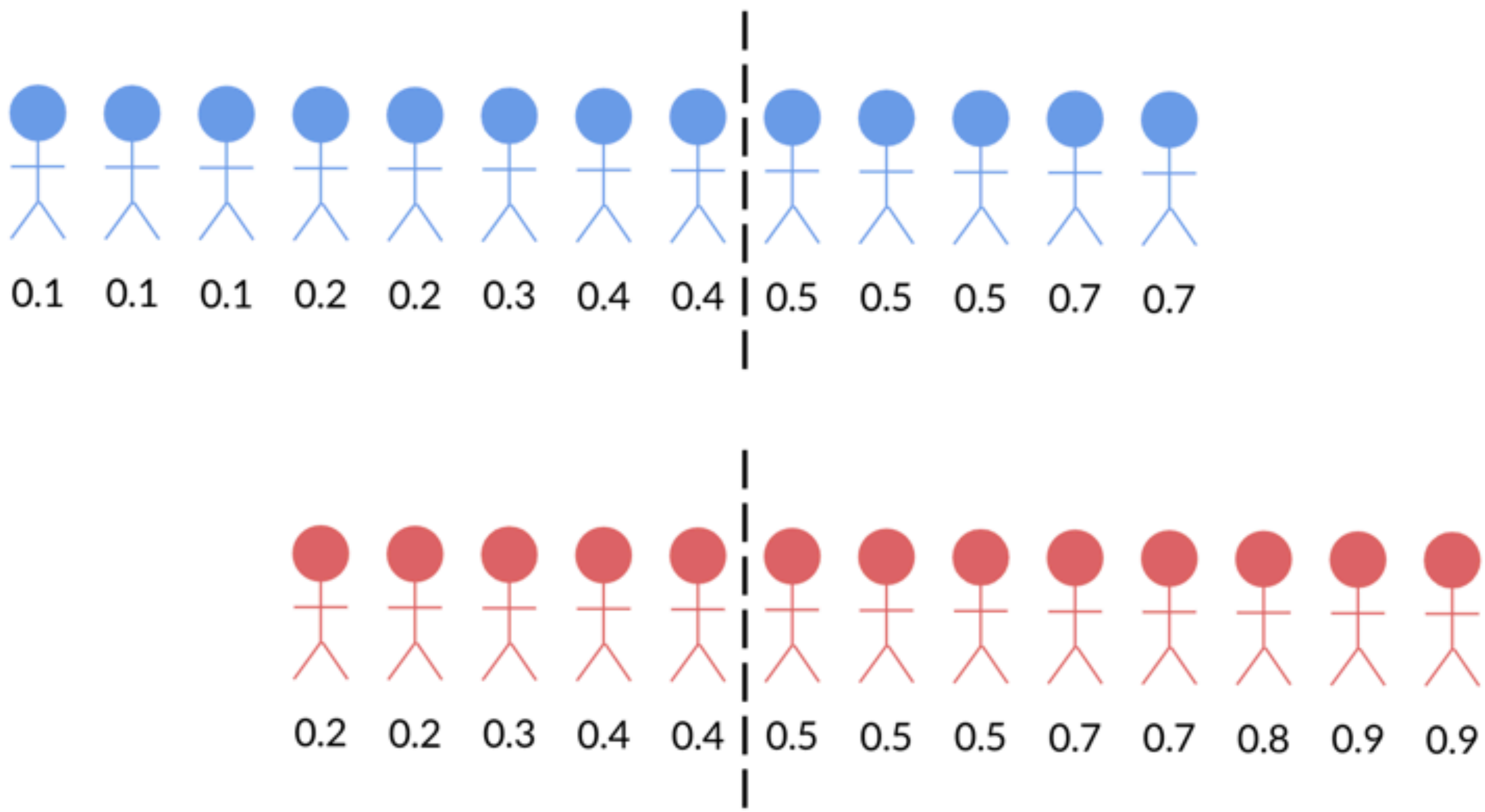
Recap

- FPR = FP / (FP + TN)

- FPR = $\frac{\text{Wouldn't have reoffended \& "predict high risk"}}{\text{Wouldn't have reoffended}}$

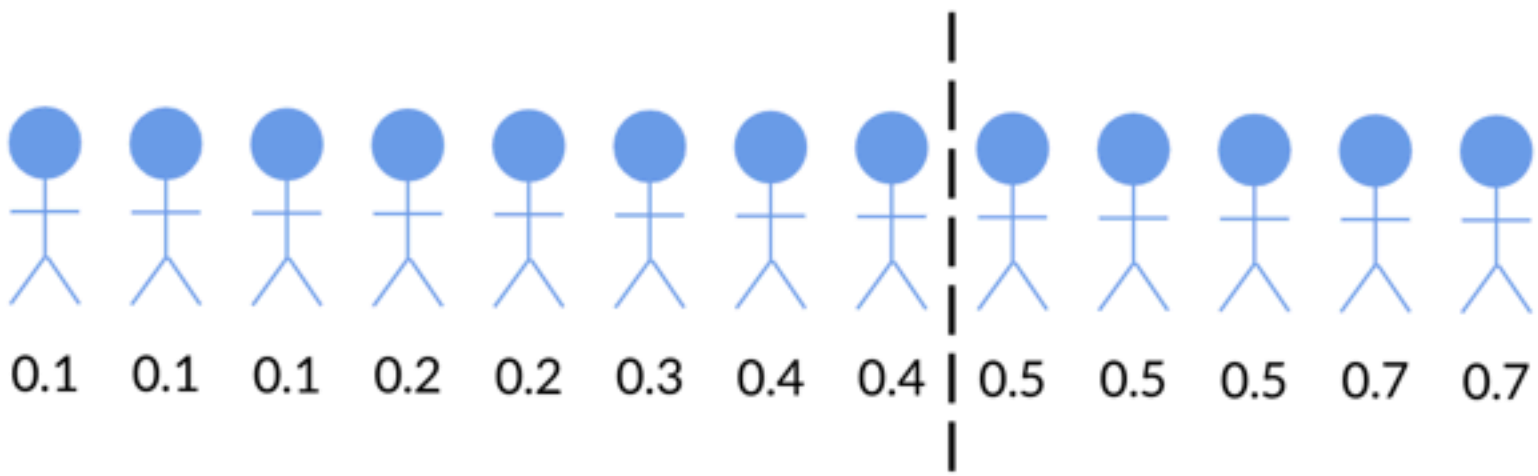
- In Broward County, FL, FPR was 31% for Blacks, and 15% White

Calculating false positive rates



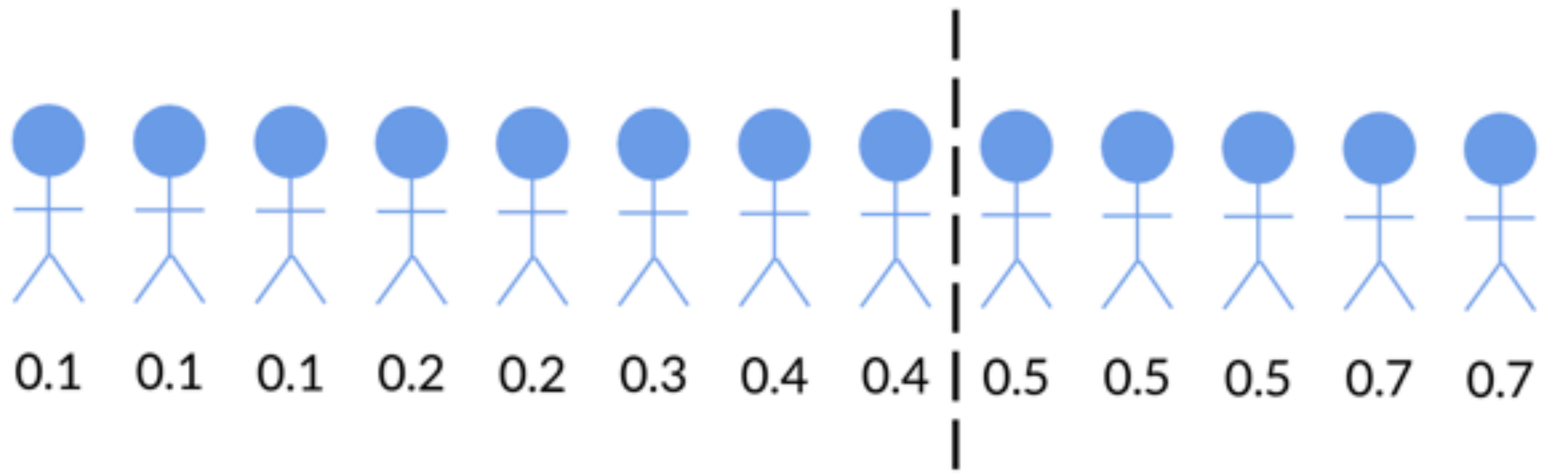
Slides by Sharad Goel

Calculating false positive rates



Slides by Sharad Goel

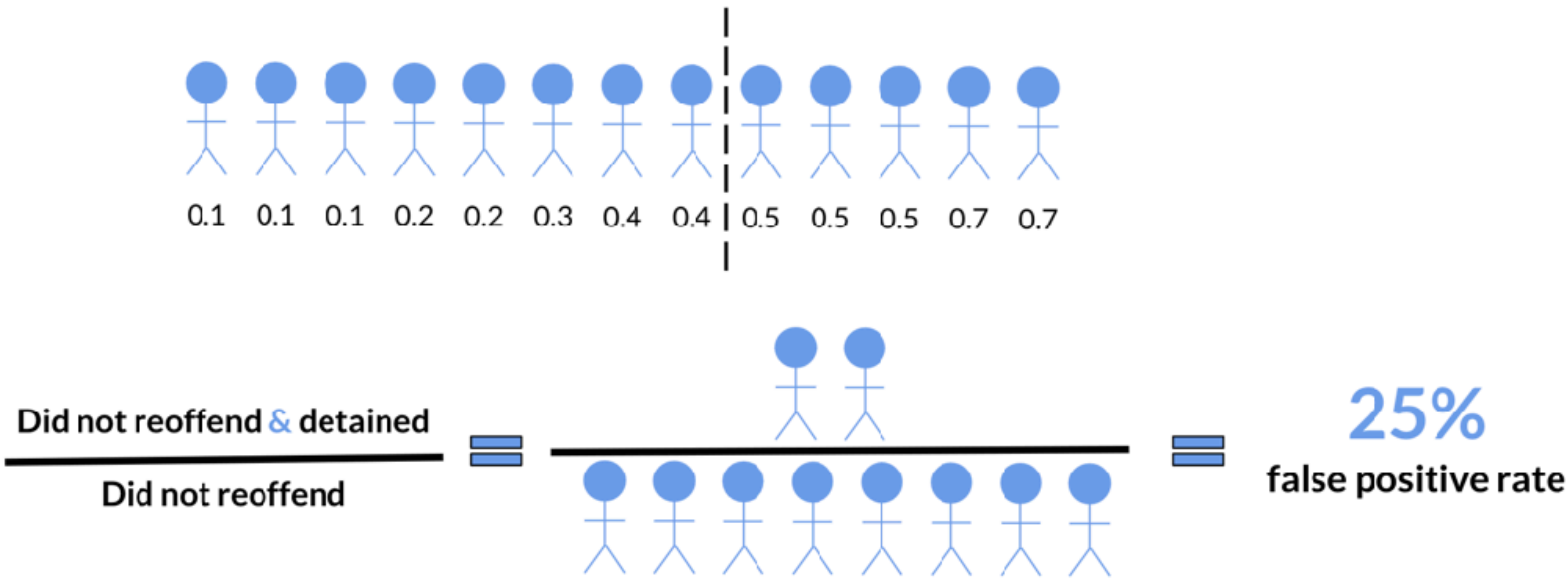
Calculating false positive rates



$$\frac{\text{Did not reoffend \& detained}}{\text{Did not reoffend}} =$$

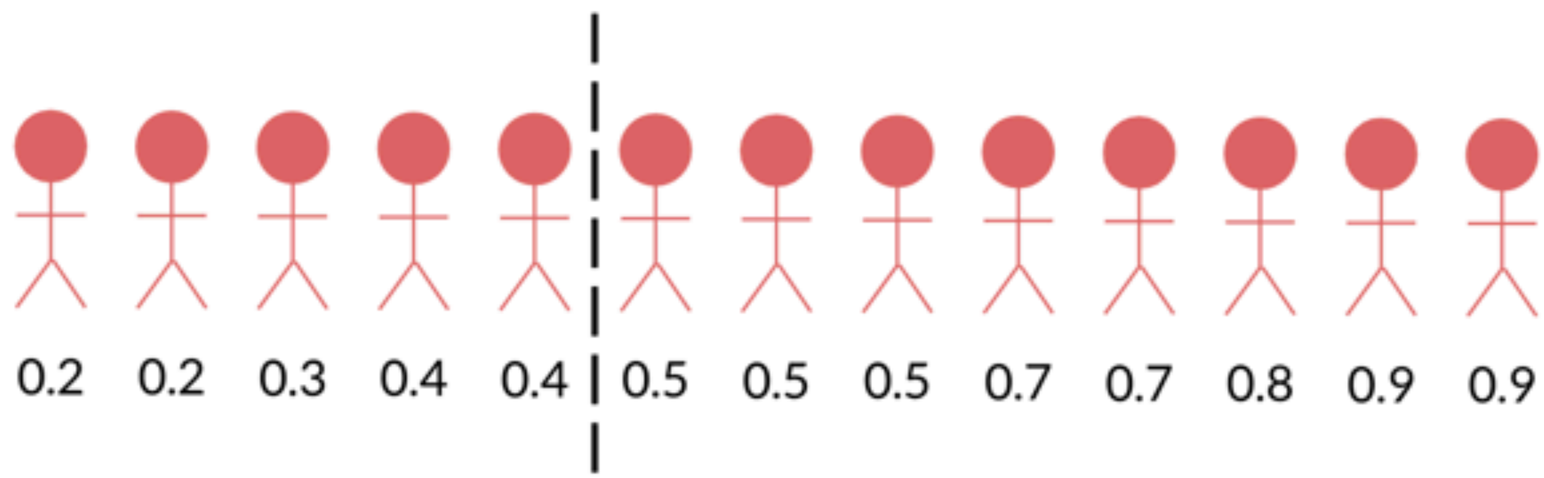
Slides by Sharad Goel

Calculating false positive rates



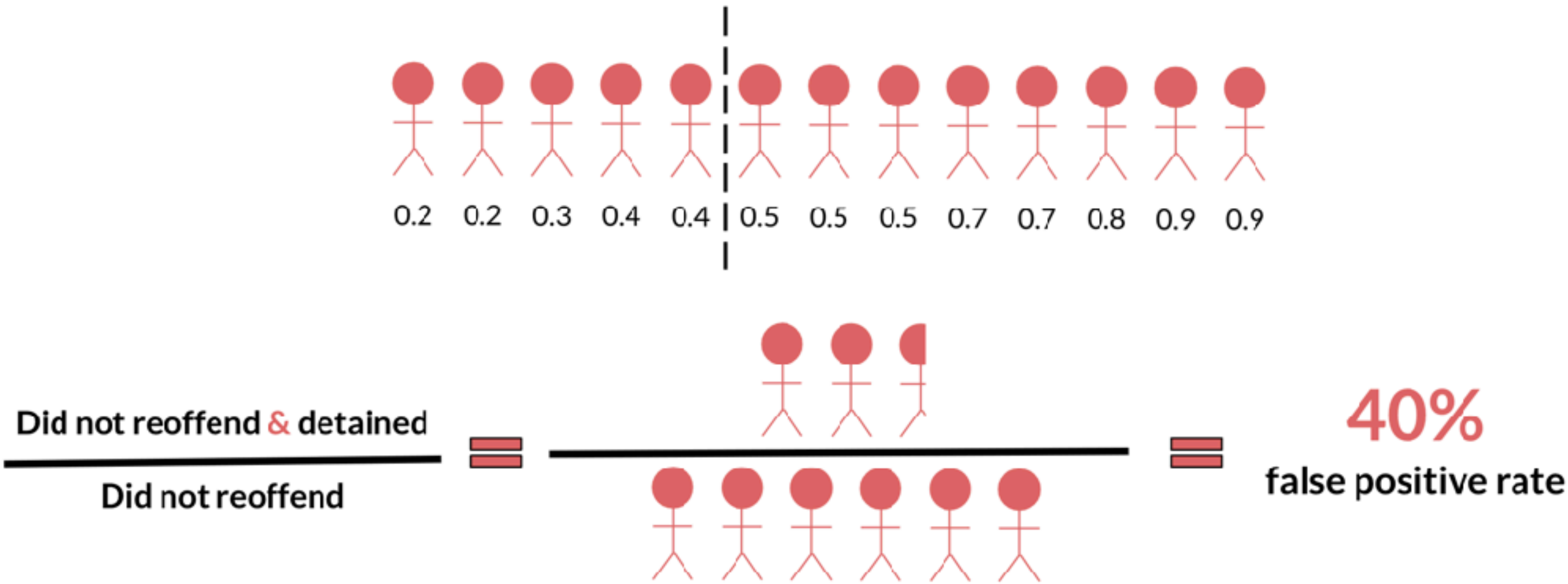
Slides by Sharad Goel

Calculating false positive rates



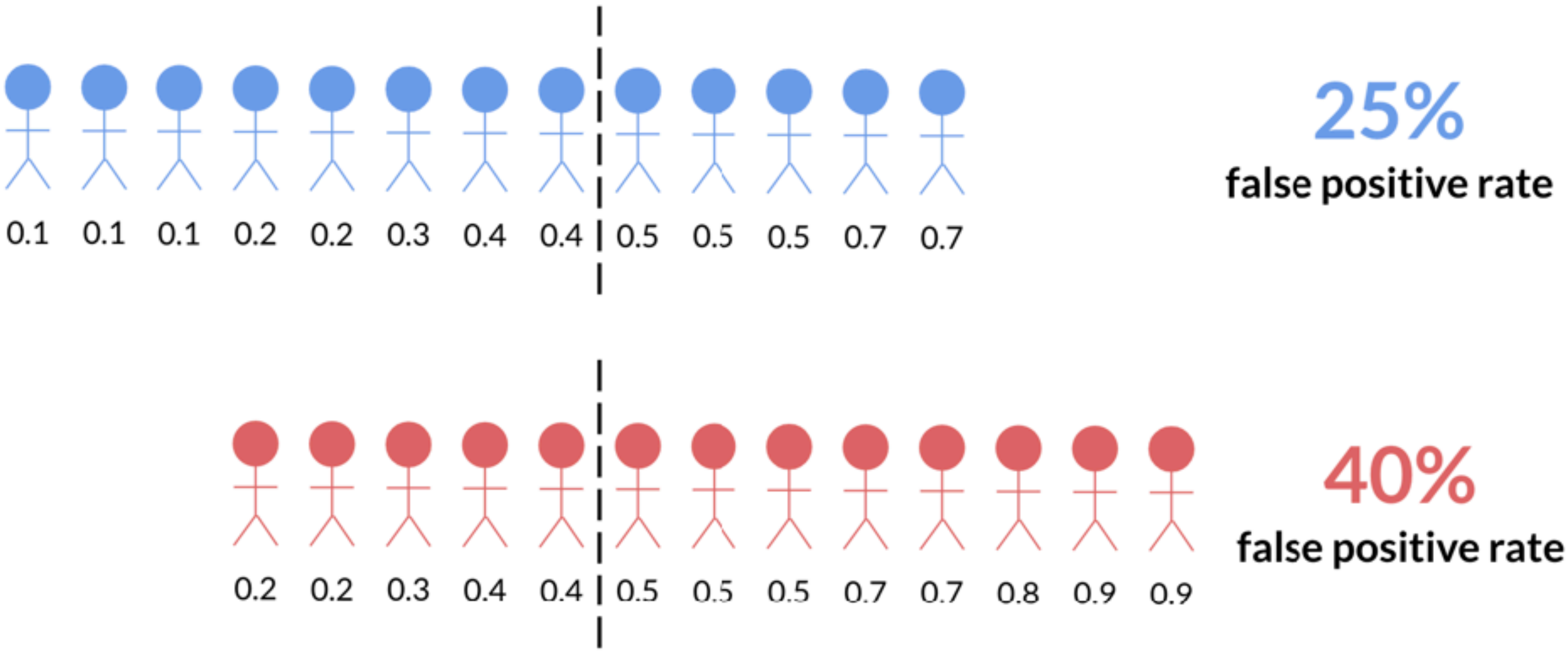
Slides by Sharad Goel

Calculating false positive rates



Slides by Sharad Goel

Calculating false positive rates

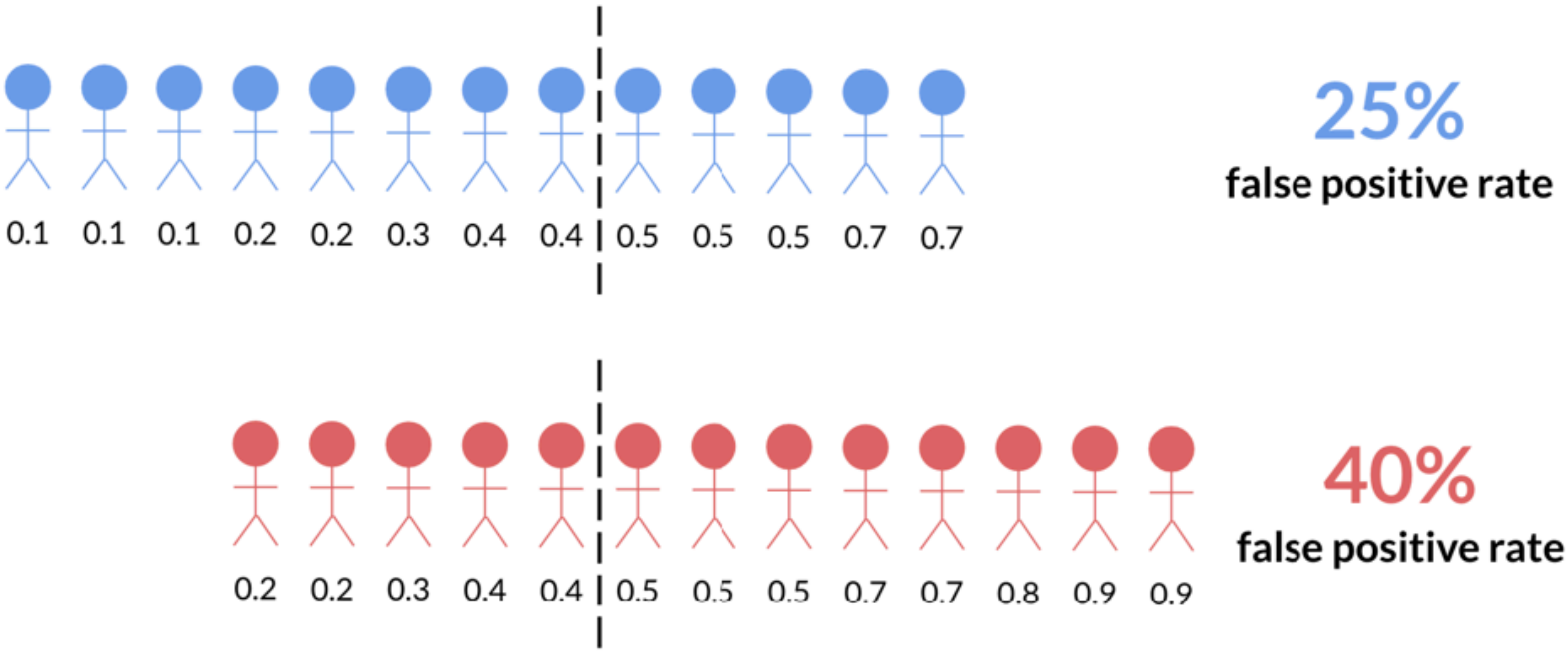


Slides by Sharad Goel

Inframarginality

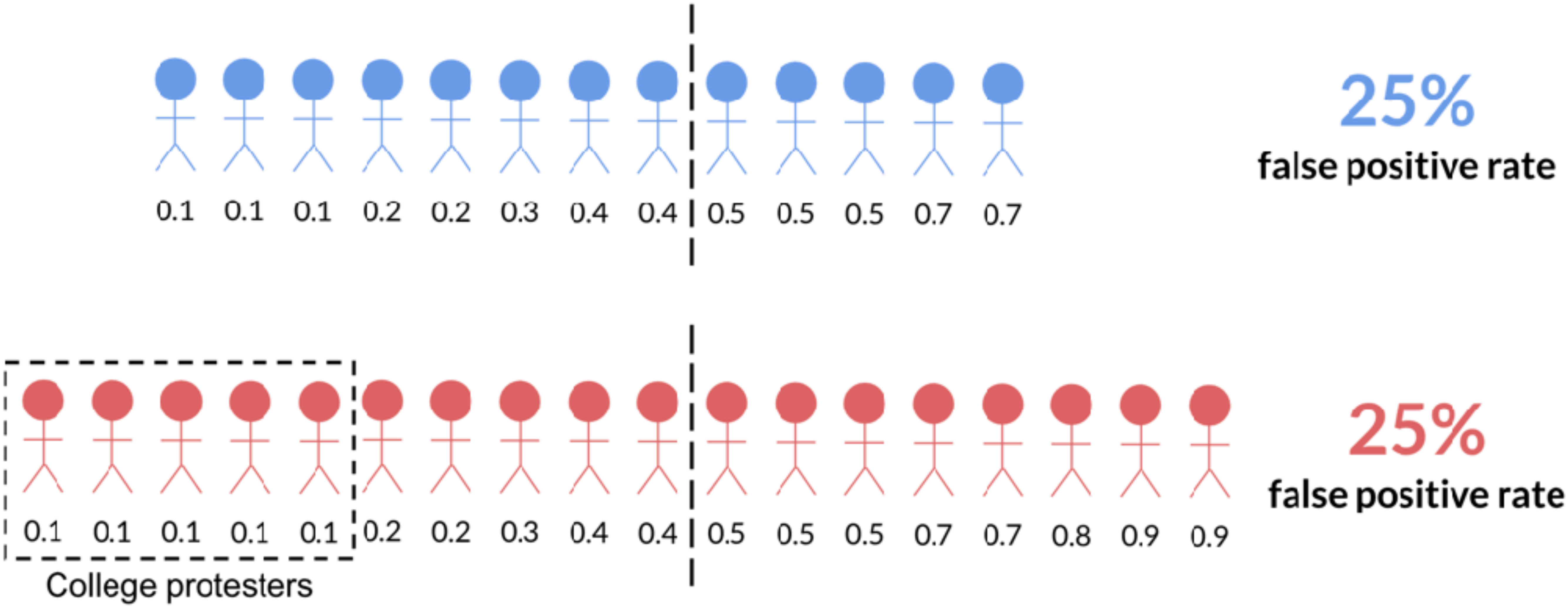
- Infra-marginal: below from the margins
  - This means a metric depends on things away from the threshold
- FPR is a infra-marginal statistic
  - It depends on the entire risk distribution, not just the threshold
  - In general, metrics from confusion matrix suffer similar issues
- This leads to misleading fairness notions when risk distributions differ across groups

Calculating false positive rates



Slides by Sharad Goel

The problem with false positive rates



Limitations

- Argument so far based on when Y and X are fixed
  - In a world where the legal, political, economic systems work against marginalized communities, data will embody inequities and biases
  - Both label and features biased
- Based on $P(Y = 1 \mid X)$ being known
  - Estimating this uniformly over features X is notoriously difficult
  - Model selection nontrivial

Limitations

- Distribution shift
  - All discussion so far based on data from Broward County, FL
  - Demographics (X), Y | X changes over time and space
- Agents' behavior may change in response to introduction of the system; introduces dynamics through time and space
- Externalities

More broadly

- Should this system exist at all?
- Is detaining people at higher risk of recidivating the right intervention?
- Structural shifts in the socioeconomic, legal, political system
  - When / how can prediction models help? As opposed to replicating the patterns in the world
  - different recidivism rates is a result of historical social and economic discrimination

Worst-case subpopulations, tail-performance, and distributional robustness

Links

https://www2.isye.gatech.edu/people/faculty/Alex_Shapiro/SPbook.pdf

https://arxiv.org/abs/1810.08750

https://arxiv.org/abs/2007.13982

https://arxiv.org/abs/1806.08010

Rest of the lecture

- So far, we focused on binary classification problems with pre-defined demographic groups
- What about generic loss minimization problems?
- Goal: guarantee good performance (low loss) uniformly over demographic subgroups
  - Quantify what “uniformly” means
  - Quantify what “demographic subgroups” mean
- Previous caveats apply about (very) limited scope

Standard Approach: Average Loss

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

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

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

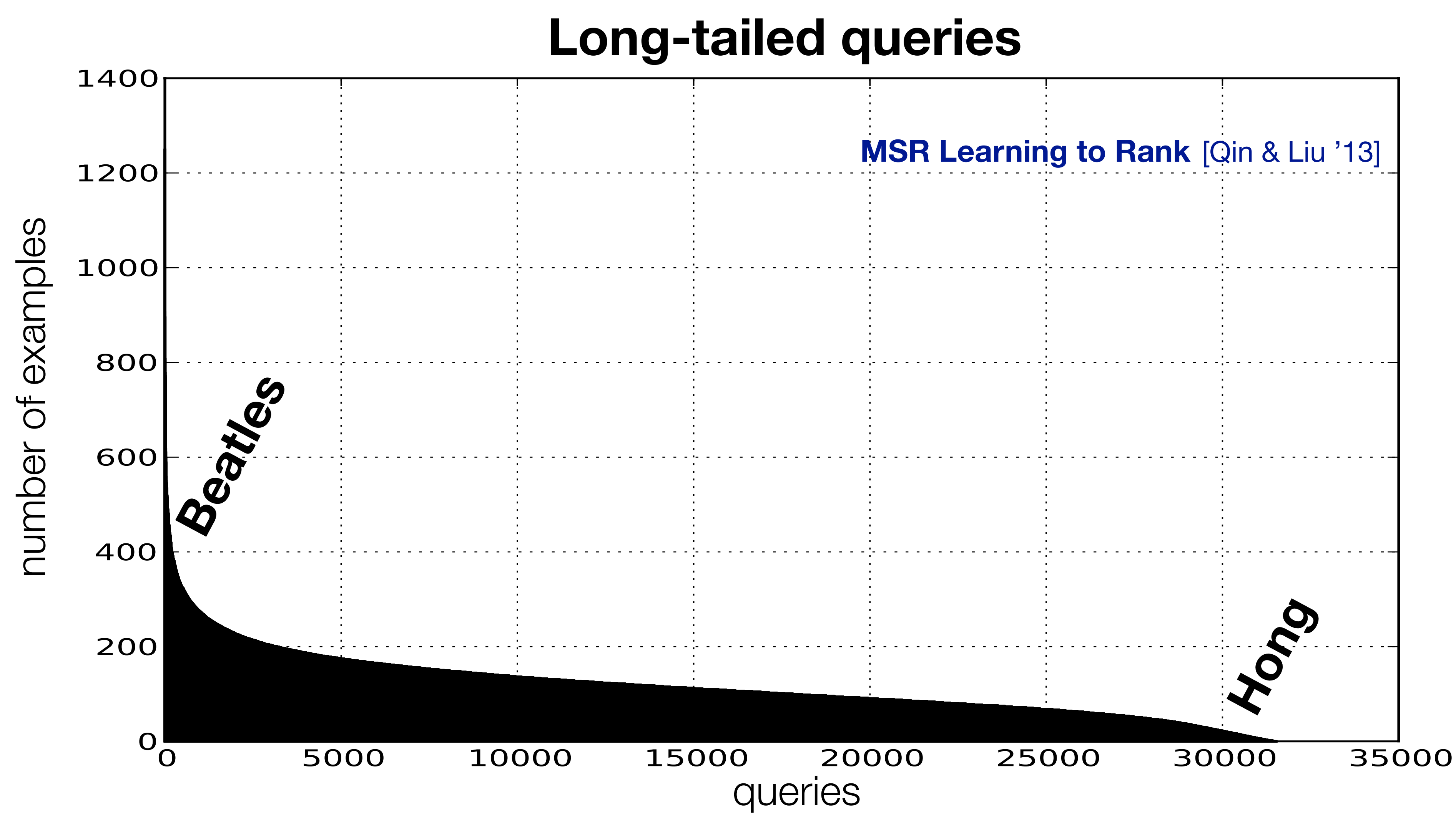
SVM (Classification) $\ell(\theta; X, Y) = (1 - Y\theta^\top X)_+$

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

More examples: newsvendor, portfolio, scheduling…

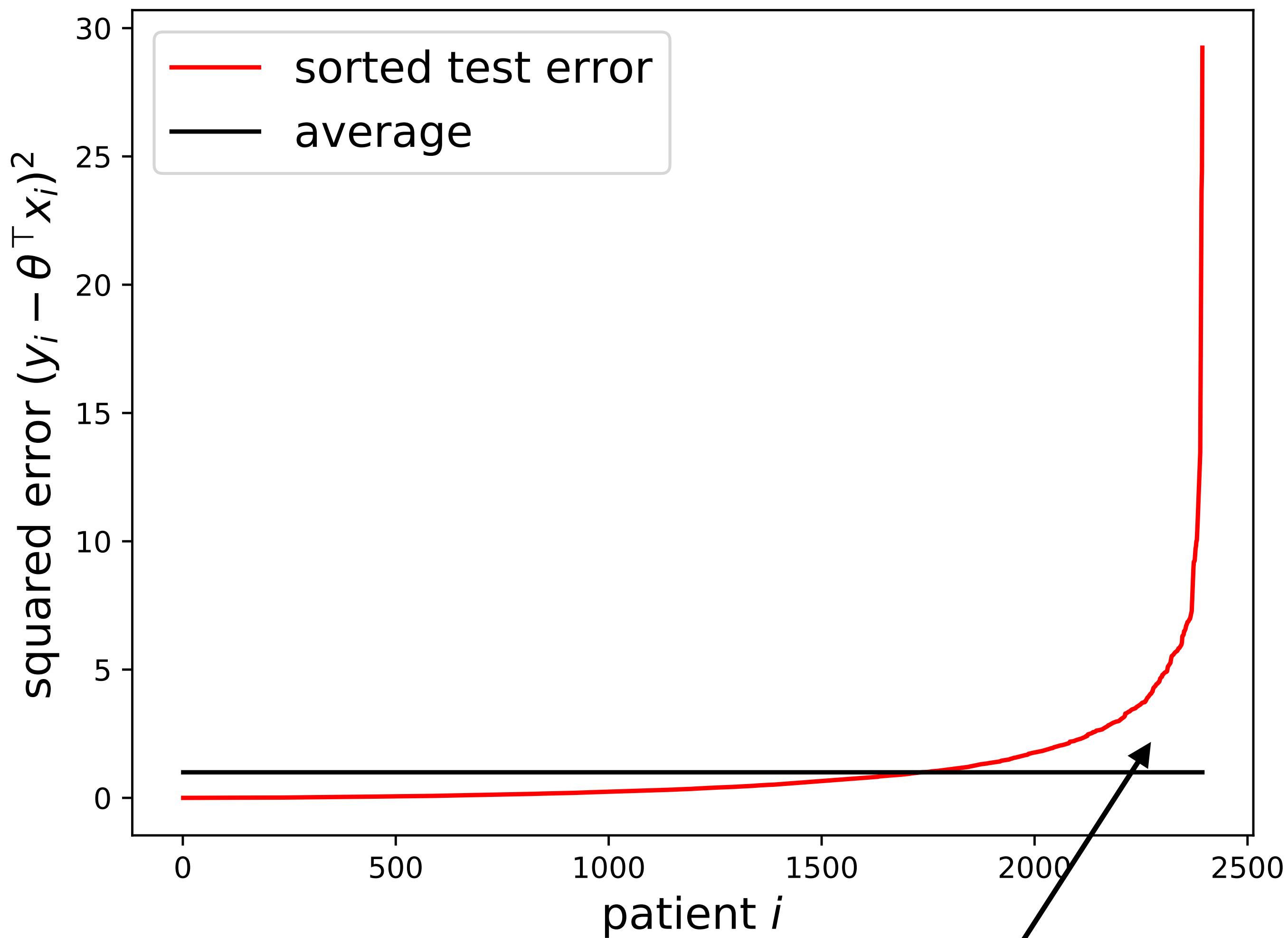
Challenge 1: Long-tails

- Long-tailed data is ubiquitous in modern applications
  - At Google, a constant fraction of queries are 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**

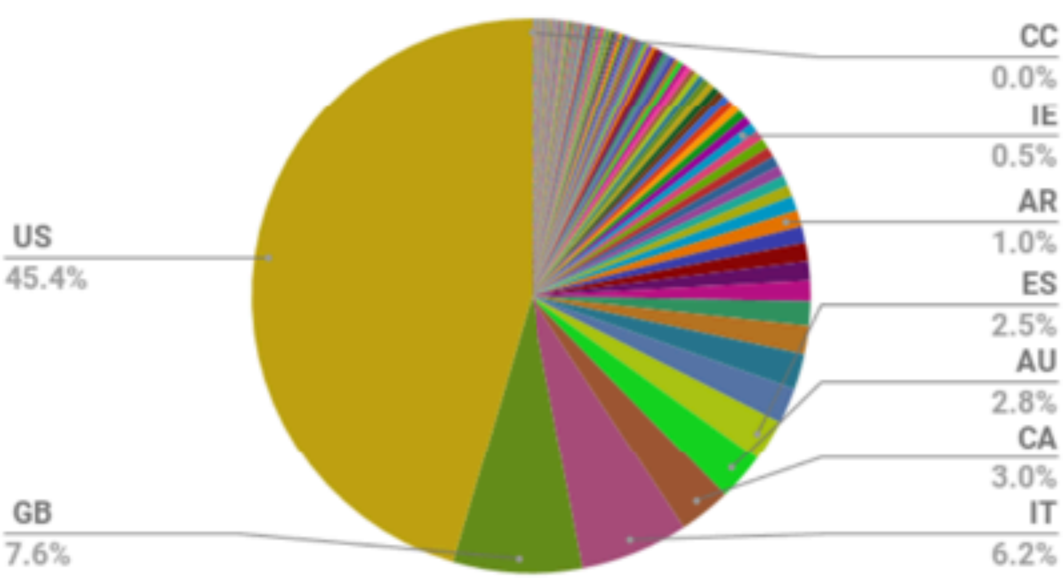
Another use for Warfarin: **rat poison**

Challenge 2: Lack of Diversity in Data

- "Clinical trials for new drugs **skew heavily white**" [Oh et al. '15, Burchard et al. '15, SA Editors '18]
  - From 1993-2013, **98.1%** of all studies on respiratory diseases did not report inclusion of **minority subjects** [Burchard et al. '18]
  - Racial minorities more likely to suffer from respiratory diseases

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

ImageNet: country of origin



[Shankar et al. '17]

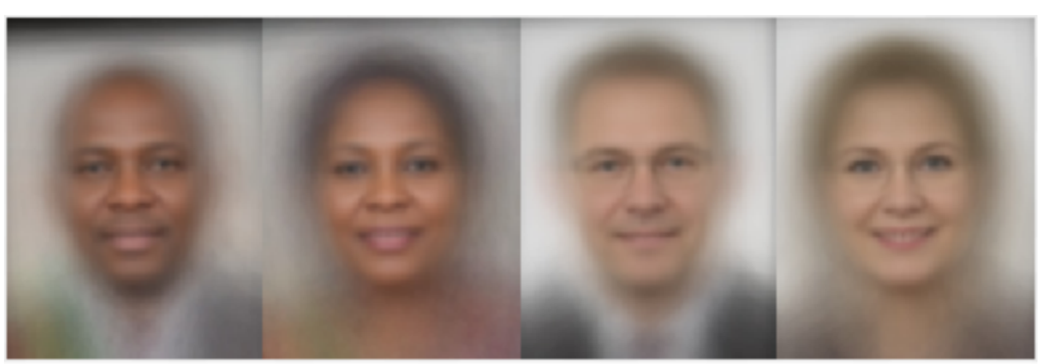
Other examples:

- Language identification [Blodgett et al. '16, Jurgens et al. '17]
- Part of speech tagging [Hovy & Sgaard '15]
- Video captioning [Tatman '17]
- Recommenders [Ekstrand et al. '17, '18]

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]

First Idea: Pre-defined groups

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

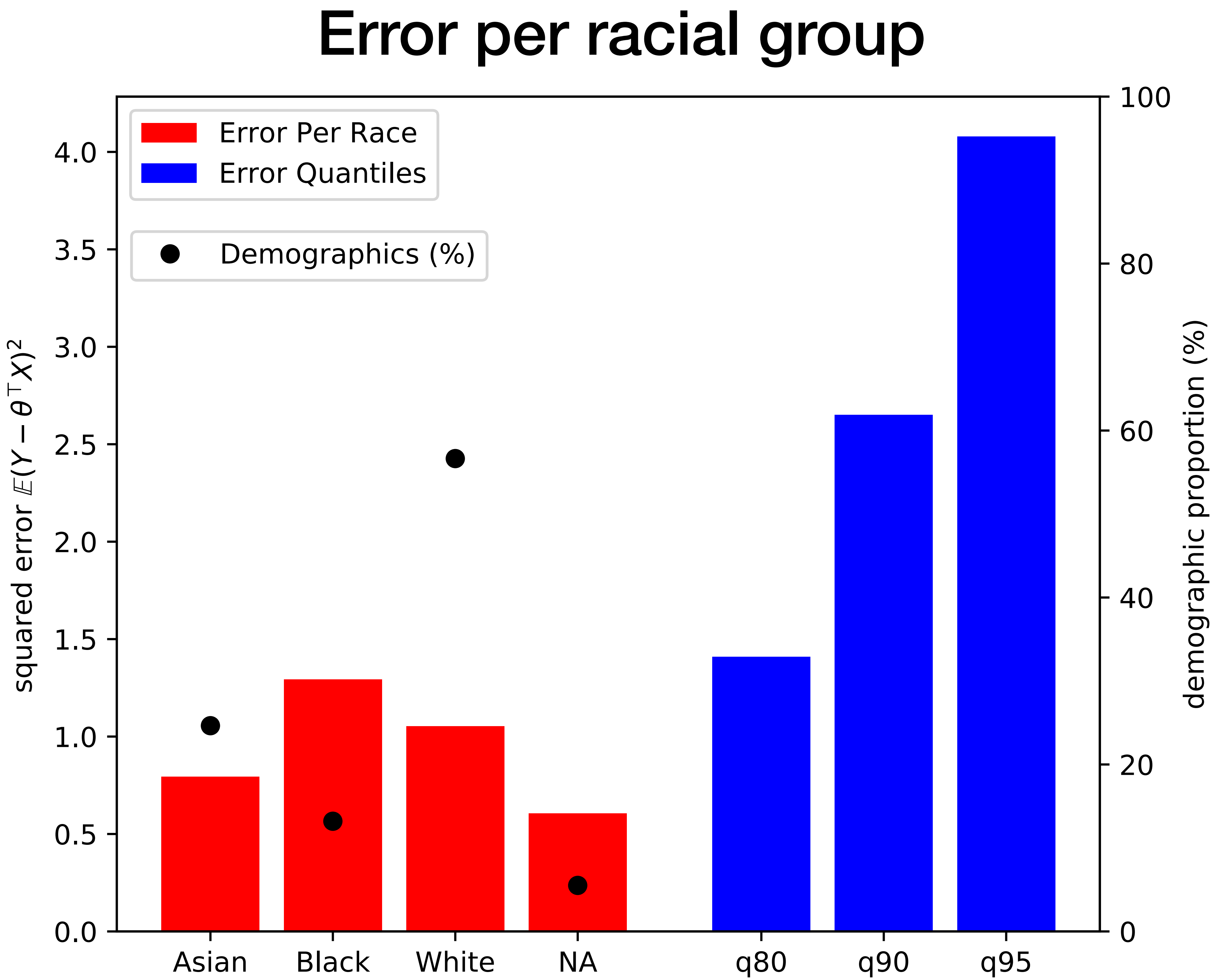
- Separate model for **each** group $\mathbb{E}_{P_g}[\ell(\theta_g; Z)]$
- One model for **worst-off** group $\max_{g \in \mathcal{G}} \mathbb{E}_{P_g}[\ell(\theta; Z)]$ [Meinshausen & Buhlmann ’15]

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

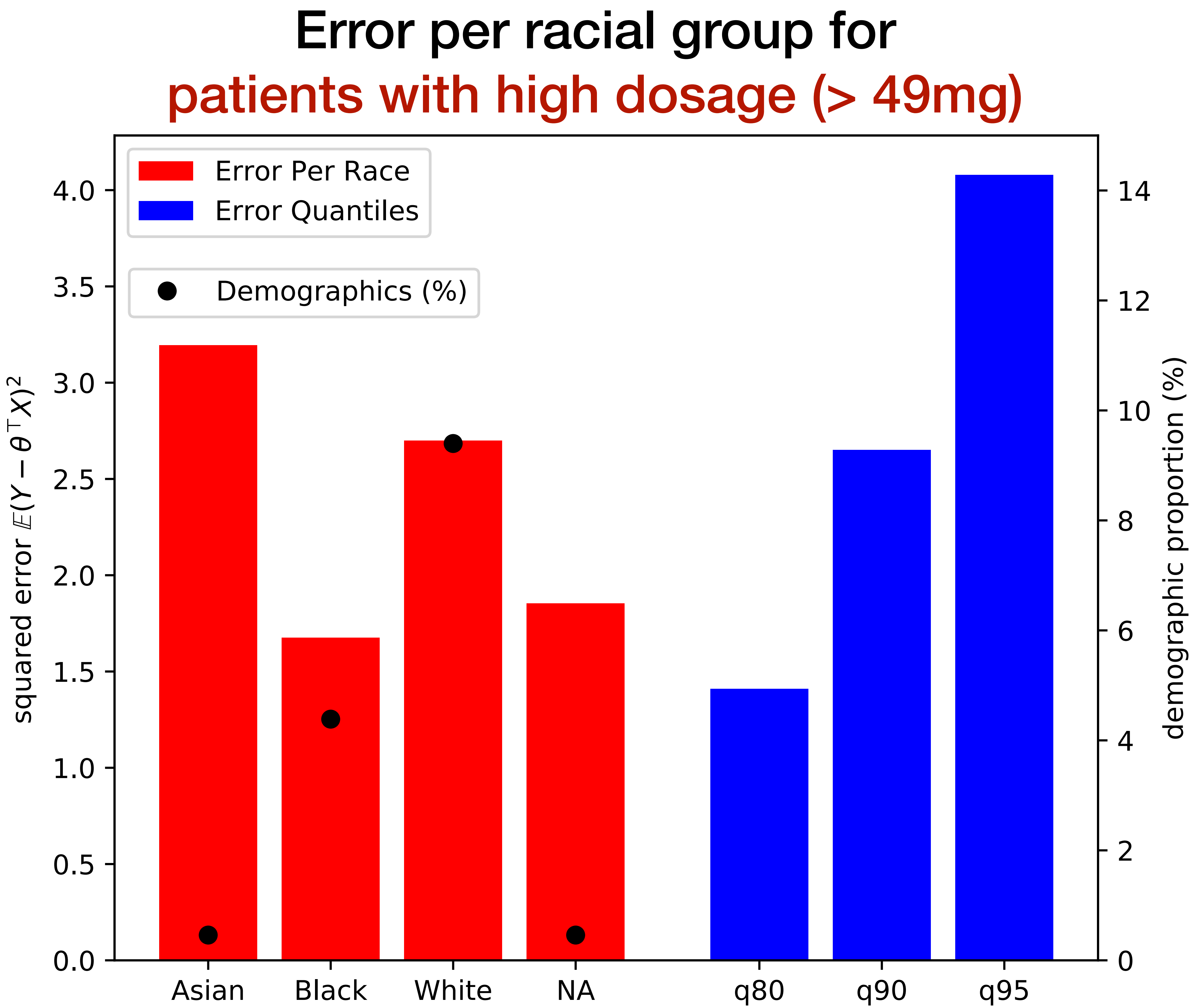
Problems

- 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 (e.g. skin color)
  - performance determined in an **intersectional** fashion
- Accounting for intersections gives **exponentially many subgroups**
  - computational & statistical difficulties

Example: Predicting Warfarin Dosage



Example: Predicting Warfarin Dosage



Preview

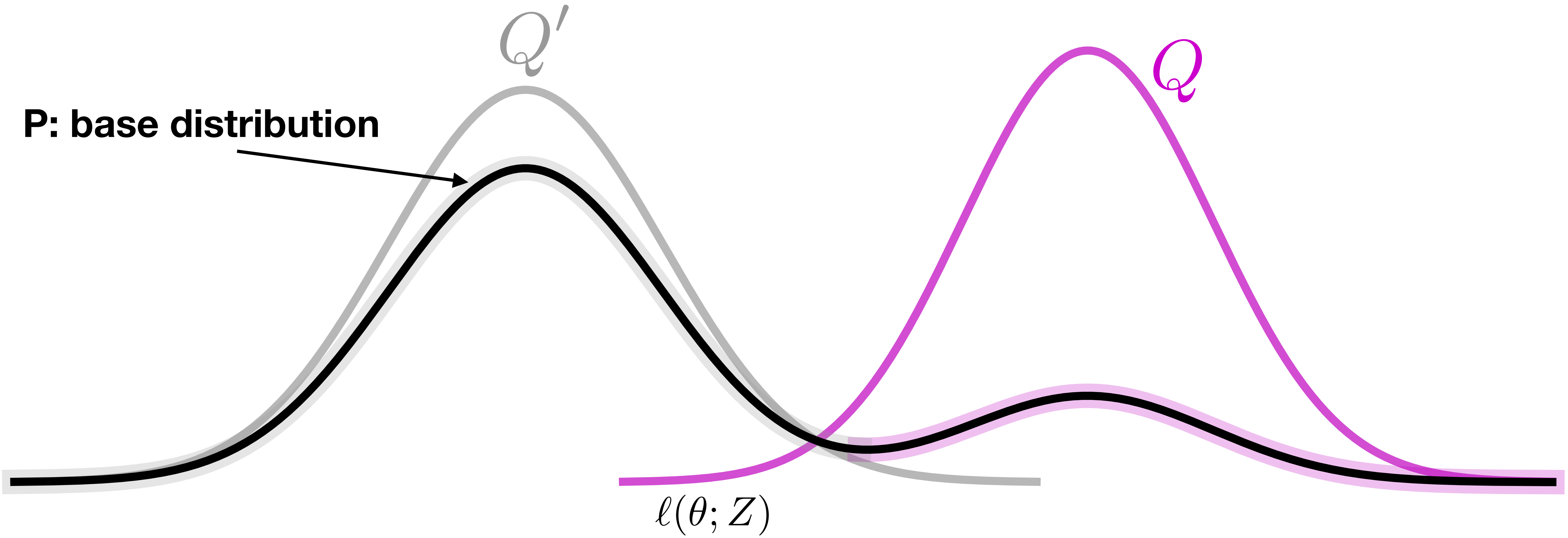
Automatically find **worst-off subpopulations**, and **optimize** performance on them

- Guarantee **uniform performance** across subpopulations
- **Computationally** efficient
- Characterize **statistical price** of subpopulation performance

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. $P(\cdot) = aQ + (1-a)Q'$

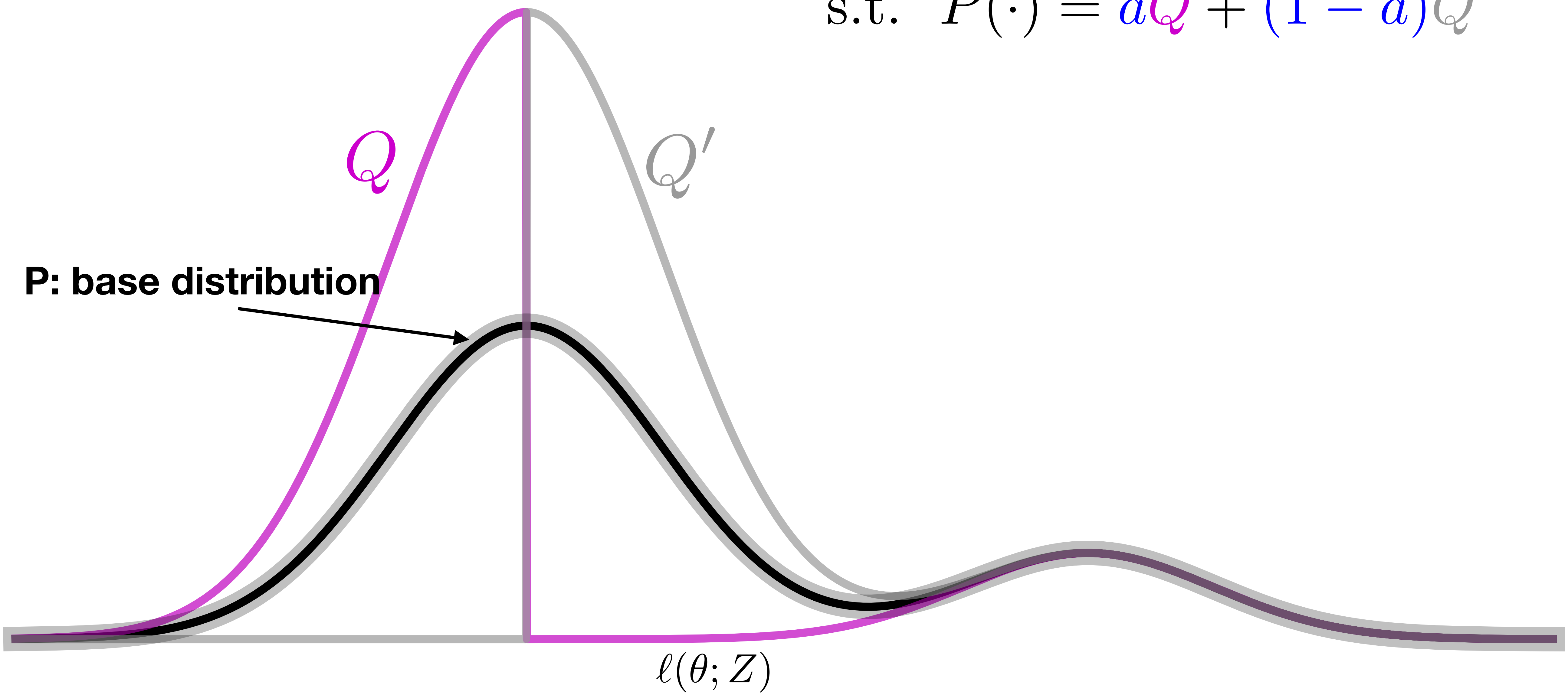


Subpopulations

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

Q is a **subpopulation** $\longleftrightarrow$ $\exists$proportion $a \in (0,1]$, prob. Q'

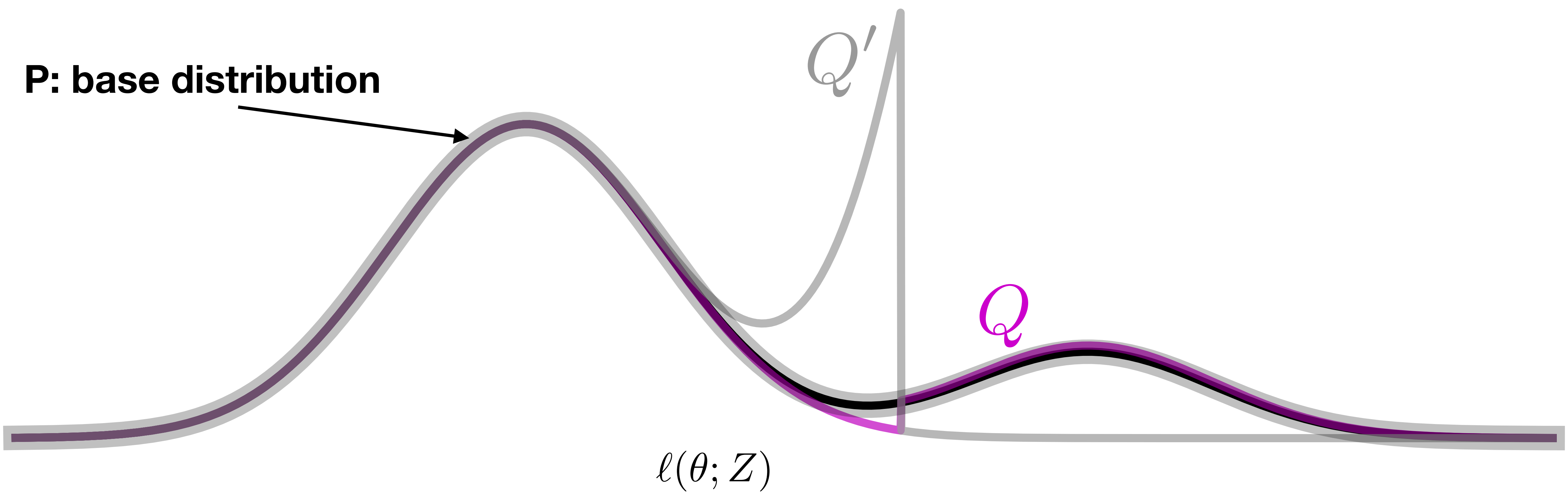
s.t. $P(\cdot) = aQ + (1-a)Q'$



Subpopulations

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

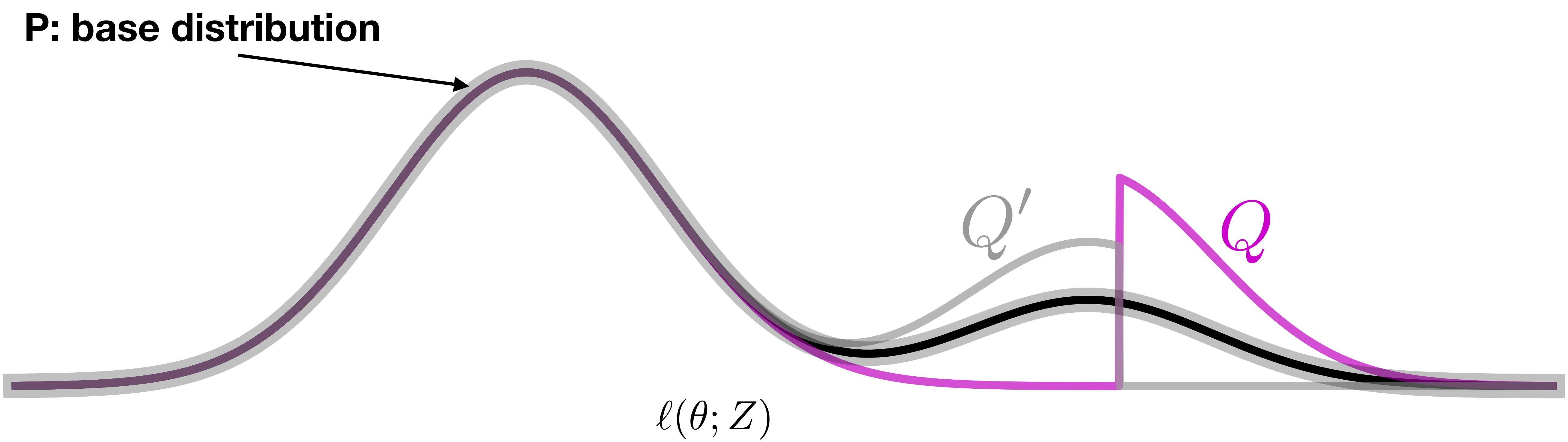
Q is a **subpopulation** $\Longleftrightarrow$ $\exists \text{proportion } a \in (0,1], \text{ prob. } Q'$ $\text{s.t. } P(\cdot) = aQ + (1-a)Q'$



Subpopulations

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

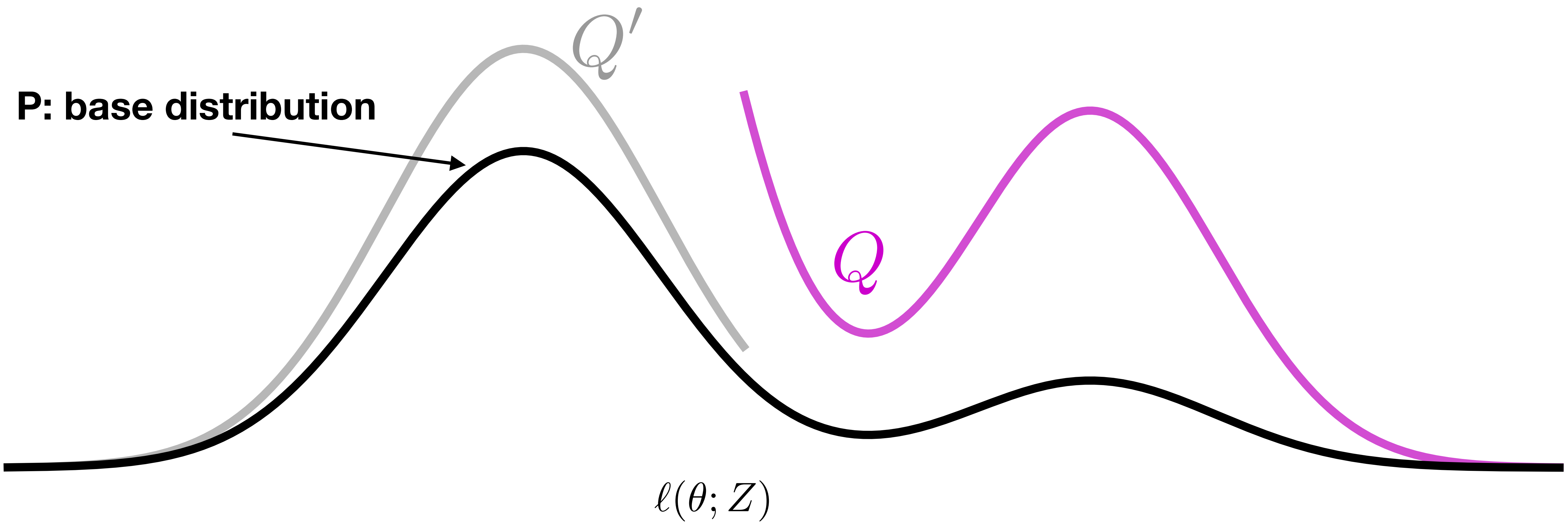
Q is a **subpopulation** $\longleftrightarrow$ $\exists \text{proportion } a \in (0,1], \text{ prob. } Q'$ s.t. $P(\cdot) = aQ + (1-a)Q'$



Subpopulations

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

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



Subpopulations

- $\mathcal{Q}$ is a **subpopulation** of P if it's a mixture component

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

Notation

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

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

Subpopulations

Notation

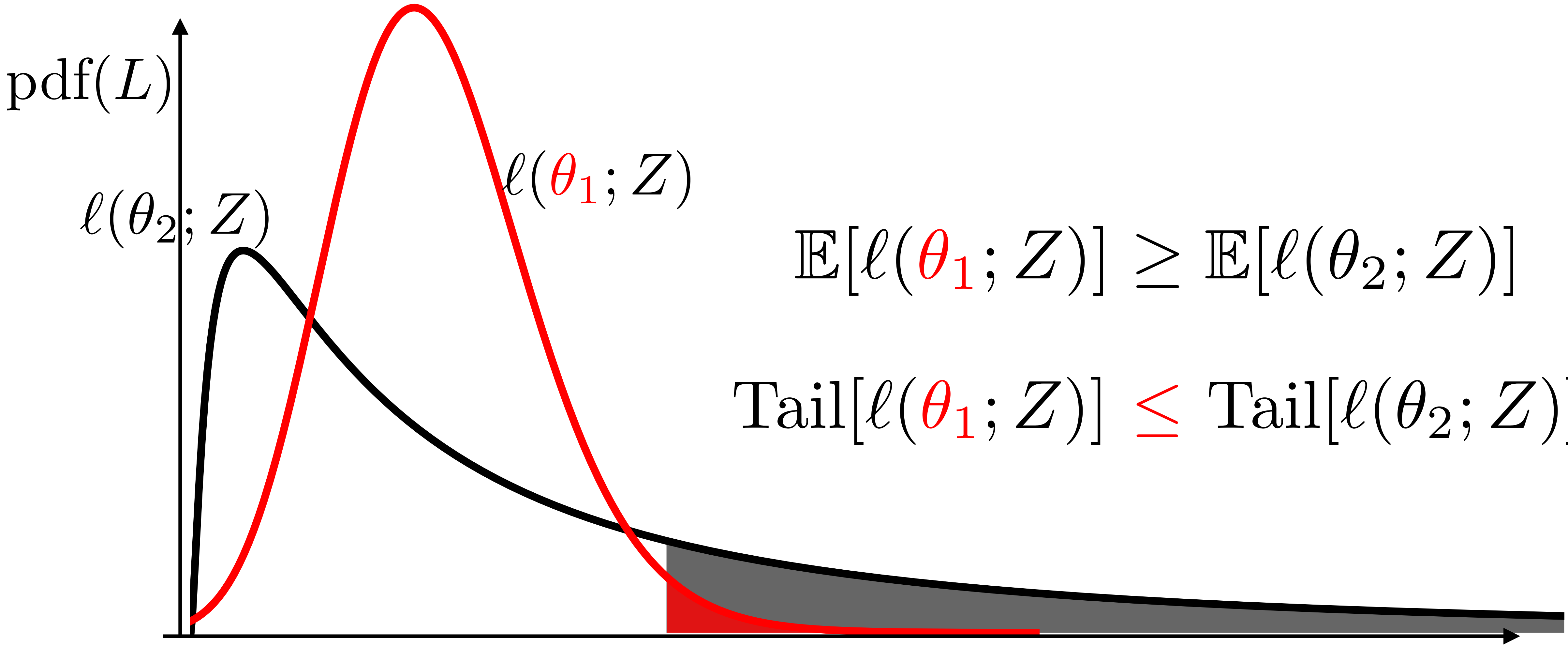
$$Q \succeq \alpha \iff \left\{ Q : \begin{array}{l} \exists \text{probability } Q', \text{ and } a \geq \alpha \\ \text{s.t. } P = aQ + (1-a)Q' \end{array} \right\}$$

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

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

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

Risk Aversion



$$\mathbb{E}[\ell(\theta_1; Z)] \geq \mathbb{E}[\ell(\theta_2; Z)]$$

$$\mathrm{Tail}[\ell(\theta_1; Z)] \leq \mathrm{Tail}[\ell(\theta_2; Z)]$$

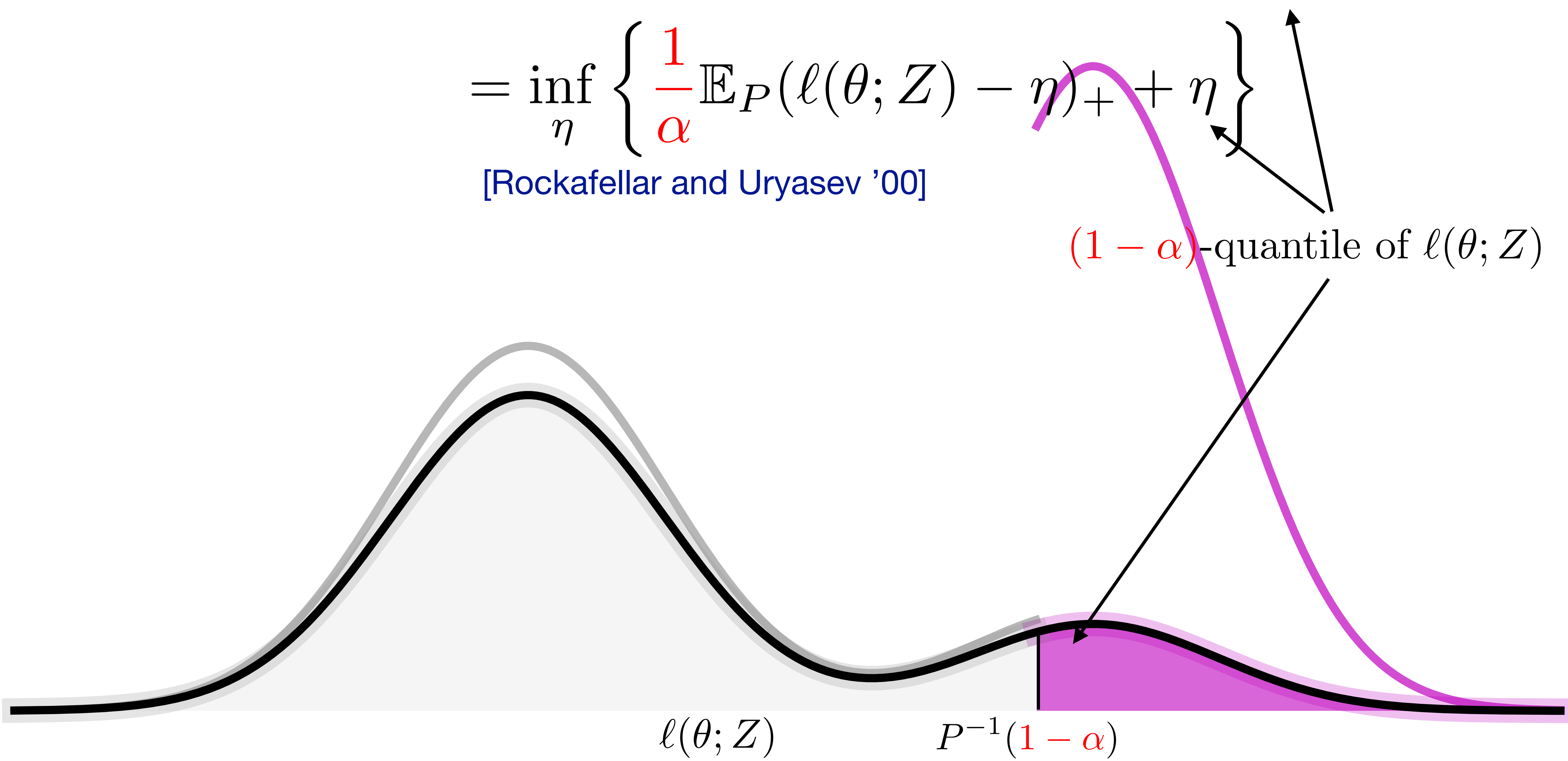
<u>Risk-aversion:</u> prefer θ_1 over θ_2

Conditional Value-at-Risk

- CVaR defines a tail-average after the $(1-\alpha)$-quantile

$$\mathrm{CVaR}_{\alpha}(\theta; P) := \mathbb{E}_P[\ell(\theta; Z) \mid \ell(\theta; Z) \geq P^{-1}(1-\alpha)]$$
$$= \inf_{\eta} \left\{ \frac{1}{\alpha} \mathbb{E}_P (\ell(\theta; Z) - \eta)_+ + \eta \right\}$$

[Rockafellar and Uryasev '00]



CVaR & Worst-case Subpopulations

$$\mathrm{CVaR}_{\alpha}(\theta; P) := \mathbb{E}_P[\ell(\theta; Z) \mid \ell(\theta; Z) \geq P^{-1}(1-\alpha)]$$

$$= \inf_{\eta} \left\{ \frac{1}{\alpha} \mathbb{E}_P(\ell(\theta; Z) - \eta)_+ + \eta \right\}$$

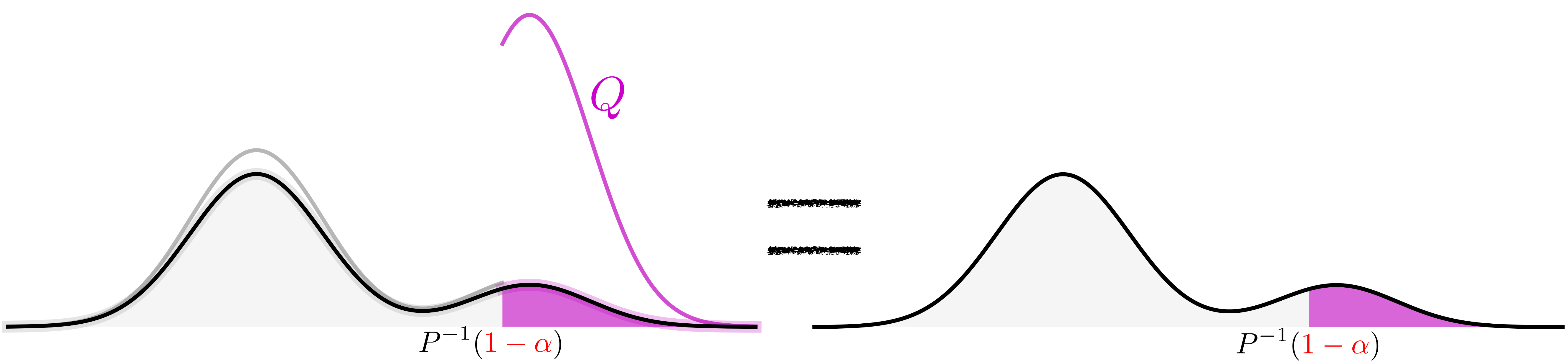
$(1-\alpha)$-quantile

Lemma: worst-case subpopulation loss = CVaR

$$\sup_{Q} \left\{ \mathbb{E}_Q[\ell(\theta; Z)] : \begin{array}{l} \exists \text{probability } Q', \text{ and } a \geq \alpha \\ \text{s.t. } P = aQ + (1-a)Q' \end{array} \right\}$$

$$\sup_{Q \succeq \alpha} \mathbb{E}_Q[\ell(\theta; Z)] = \mathrm{CVaR}_{\alpha}(\theta; P)$$

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



Conditional Value-at-Risk

- CVaR defines a tail-average after the $(1-\alpha)$-quantile

$$\begin{aligned}\mathrm{CVaR}_{\alpha}(\theta; P) &:= \mathbb{E}_P[\ell(\theta; Z) \mid \ell(\theta; Z) \geq P^{-1}(1-\alpha)] \\ &= \inf_{\eta} \left\{ \frac{1}{\alpha} \mathbb{E}_P(\ell(\theta; Z) - \eta)_+ + \eta \right\}\end{aligned}$$

- Only count inputs that suffer loss higher than η
- If $\theta \mapsto \ell(\theta; Z)$ is convex, then jointly convex in (θ, η)
- Tail-performance = worst-case subpopulation performance

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**

Coherent Risk Measures [Artzner '99]

Definition A risk measure $\mathcal{R} : L^p(\mathcal{Z}) \to \mathbb{R}$ is **coherent** if

1) Convexity: for $t \in [0,1]$

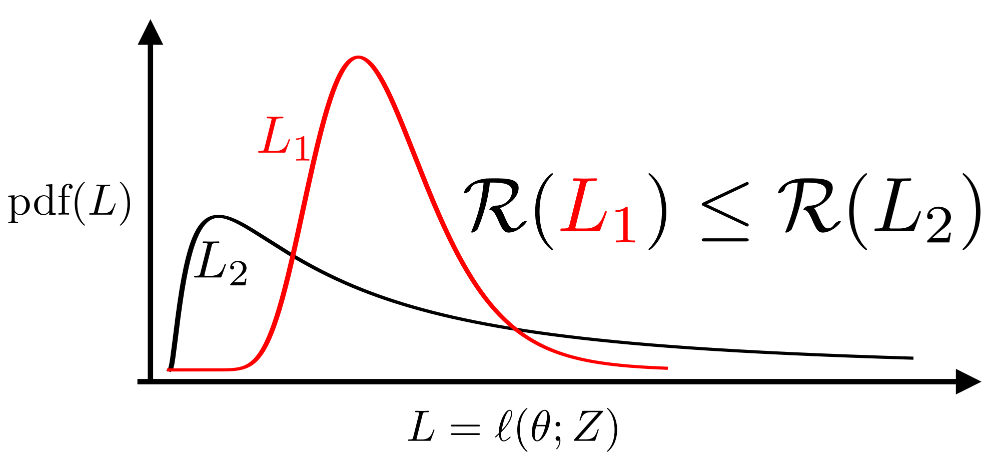
$$\mathcal{R}(tL + (1-t)L') \leq t\mathcal{R}(L) + (1-t)\mathcal{R}(L')$$

2) Monotonicity: if $L \leq L'$ a.s., then $\mathcal{R}(L) \leq \mathcal{R}(L')$

3) Translation Equivariance: for $c \in \mathbb{R}$

$$\mathcal{R}(L + c) = \mathcal{R}(L) + c$$

4) Positive Homogeneity: for $t > 0$, $\mathcal{R}(tL) = t\mathcal{R}(L)$



Risk-aversion: prefer L_1

Worst-case subpopulations = coherence

Worst-case over **all subpopulations** Q_0

$$\mathcal{R}_{\mathcal{P}_A}(W) := \sup_{P_A \in \mathcal{P}_A} \mathbb{E}_{A \sim P_A}\left[\sup_{Q \succeq A} \mathbb{E}_Q[W]\right]$$

Worst-case over **probability** P_A **on minority proportion**

Lemma (Kusuoka '01, Pflug & Romisch '07)

Under mild regularity, for any coherent risk measure, there is a **convex** set $\mathcal{P}_A$ of probabilities such that the risk measure is equal to $\mathcal{R}_{\mathcal{P}_A}(\cdot)$

From previous lecture, we have DRO = coherence = worst-case subpopulations

f-divergences DRO

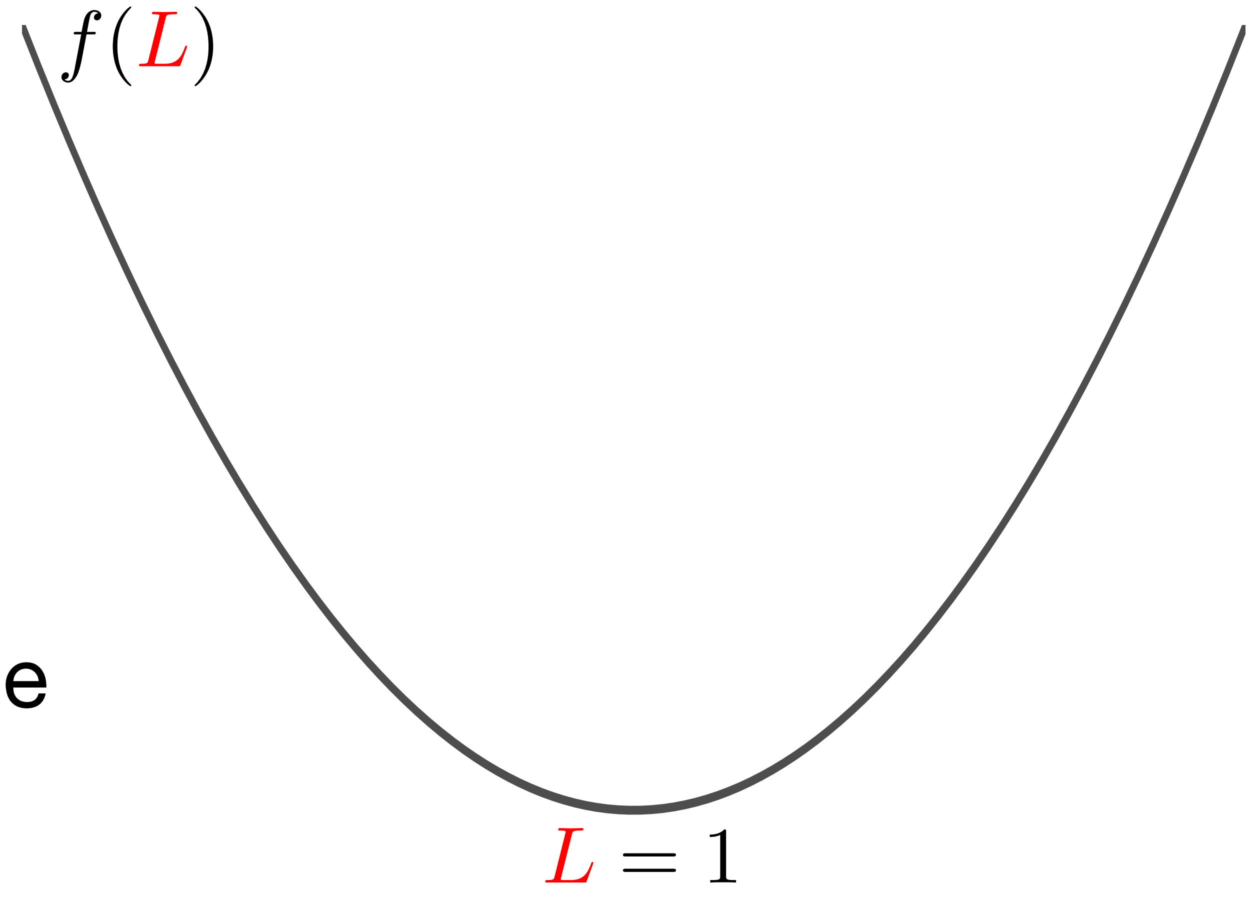
f-divergence: If $L = \frac{dQ}{dP}$ is “near 1”, then Q and P are near

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!



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

f-divergences DRO

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

Lemma: f-div DRO optimizes worst-case subpopulation

$$\sup_{Q : D_{f_k}(Q \| P_{\text{obs}}) \le \rho} \mathbb{E}_Q[\ell(\theta; Z)] = \inf_{\eta} \left\{ \frac{1}{\alpha} \left(\mathbb{E}_P(\ell(\theta; Z) - \eta)_+^{k_*} \right)^{\frac{1}{k_*}} + \eta \right\}$$

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

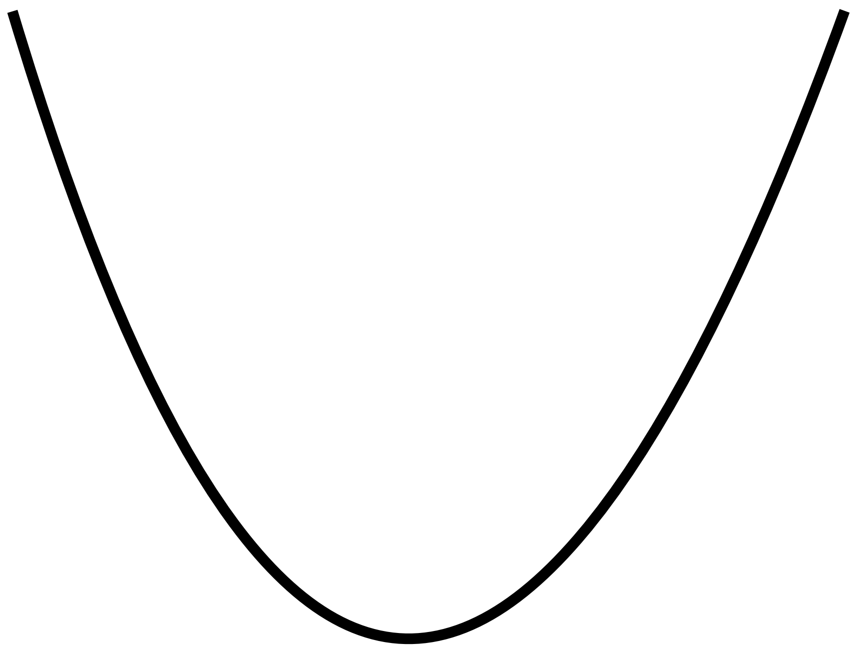
where $\alpha_k(\rho)^{-1} := (1 + k(k-1)\rho)^{1/k}$, and $k_* = k/(k-1)$

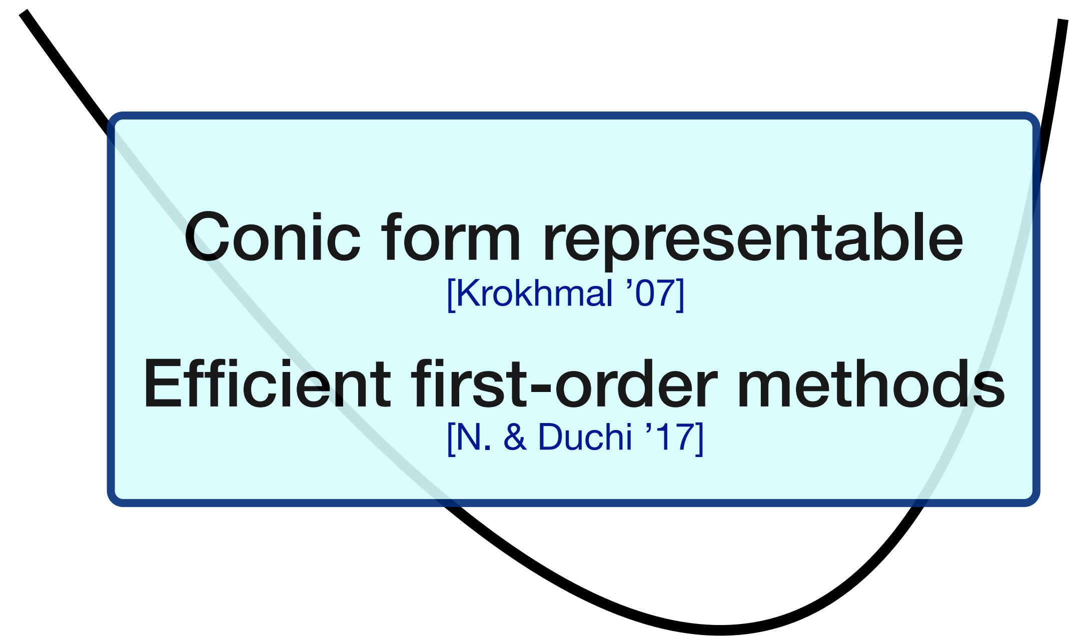
$\mathcal{P}_{A,k,\rho} := \{$ **Set of random minority proportions lower bounded by** $\alpha_k(\rho)$ $\}$

See also [Dentcheva 10]

Convexity

$$\underset{\theta \in \Theta, \eta}{\text{minimize}} \left\{ \frac{1}{\alpha} \left(\mathbb{E}_{P_{\text{obs}}} (\ell(\theta; Z) - \eta)_+^{k_*} \right)^{\frac{1}{k_*}} + \eta \right\}$$





Conic form representable
[Krokhmal '07]

Efficient first-order methods
[N. & Duchi '17]

convex loss

$\theta \mapsto \ell(\theta; Z)$

→

convex worst-case risk

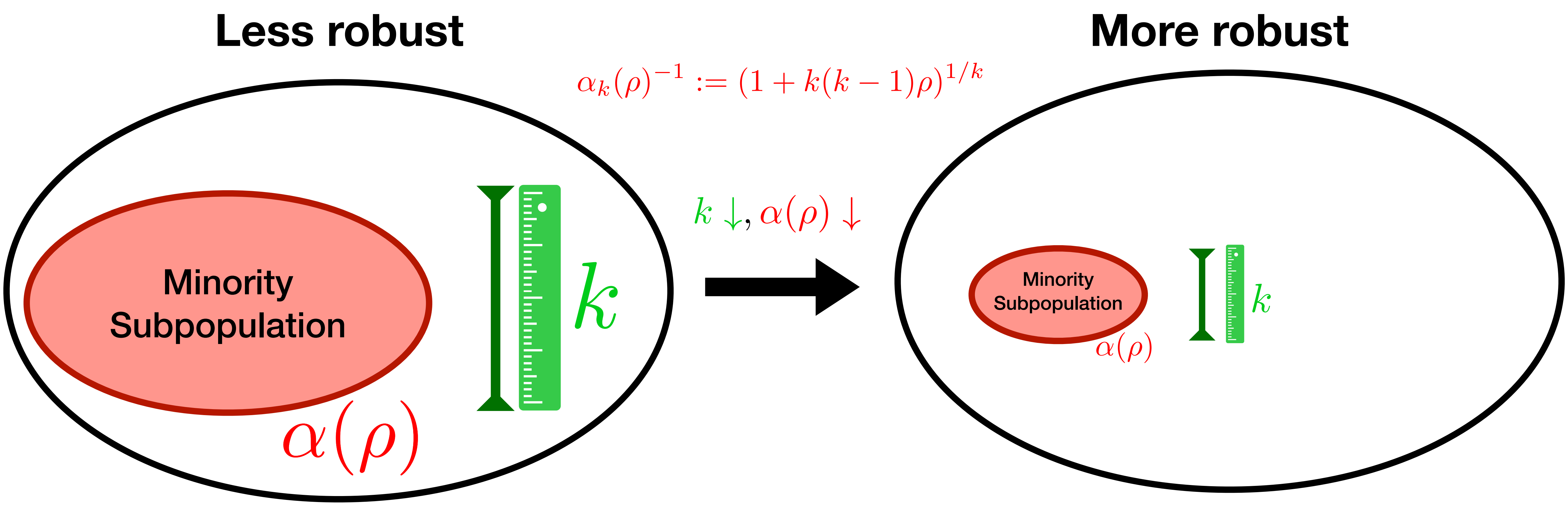
$\theta \mapsto \mathcal{R}_{p,\alpha}(\theta; \widehat{P}_{\text{obs},n})$

Example: $\ell(\theta; X, Y) = \frac{1}{2}(Y - \theta^\top X)^2$

Interpretation

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

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



- 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

Convergence guarantee

Plug-in procedure:

Let $\widehat{P}_n$ be the empirical distribution on $Z_1, .., Z_n \overset{\text{iid}}{\sim} P_{\text{obs}}$

$$\widehat{\theta}_n^{\text{rob}} \in \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ \mathcal{R}_{k,\rho}(\theta; \widehat{P}_n) = \sup_{Q: D_{f_k}(Q \| \widehat{P}_n) \le \rho} \sum_{i=1}^{n} q_i \ell(\theta; Z_i) \right\}$$

Theorem (Duchi & N. '18)

For bounded Lipschitz losses, with probability at least $1 - e^{-t}$,

$$\mathcal{R}_{k,\rho}(\widehat{\theta}_n^{\text{rob}}; P_{\text{obs}}) - \min_{\theta \in \Theta} \mathcal{R}_{k,\rho}(\theta; P_{\text{obs}}) \lesssim \sqrt{t + d \log n} \cdot n^{-\frac{1}{k_* \vee 2}}$$

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

$k \in [2, \infty)$: parametric rate, $k \in (1, 2)$: slower rate

Fundamental lower bound

Theorem (Duchi & N. '18)

Linear function $\ell(\theta; Z) = \theta Z$ on $[-1, 1]$, $\mathcal{P}$ s.t. Z bounded

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

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

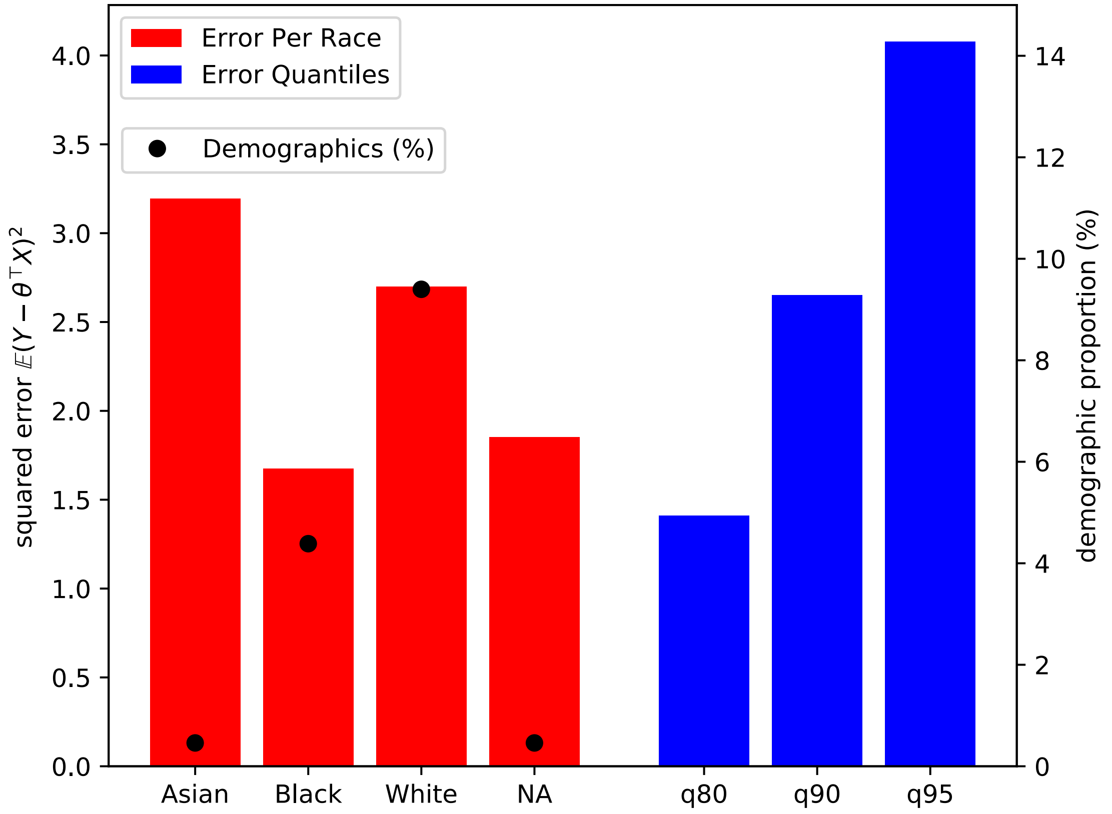
- Matching upper and lower bounds in n

  → Plug-in procedure is **optimal** in sample complexity!

- Statistical price of subpopulation performance
- Slow nonparametric rates unavoidable for $k \in (1, 2)$

Warfarin dosage

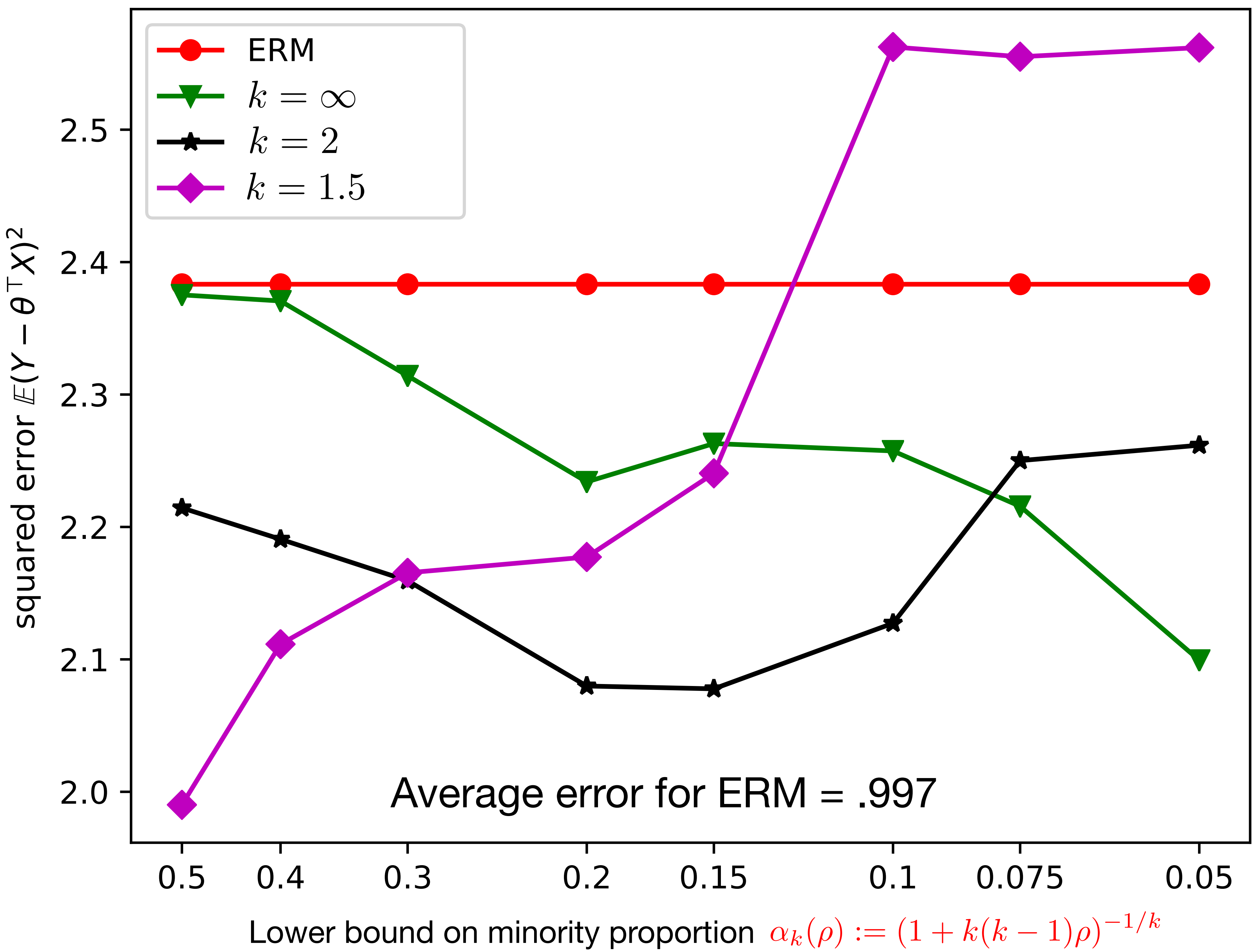
- Warfarin is the most widely used blood-thinner worldwide
- Y: therapeutic dosage
- X: demographics, **genetic** info
- Model: linear
  - Worked best out of polynomial regression, kernel methods, neural networks, splines, boosting, bagging [IWPC '09]



- Loss: squared loss $\ell(\theta; X, Y) = (Y - \theta^\top X)^2$

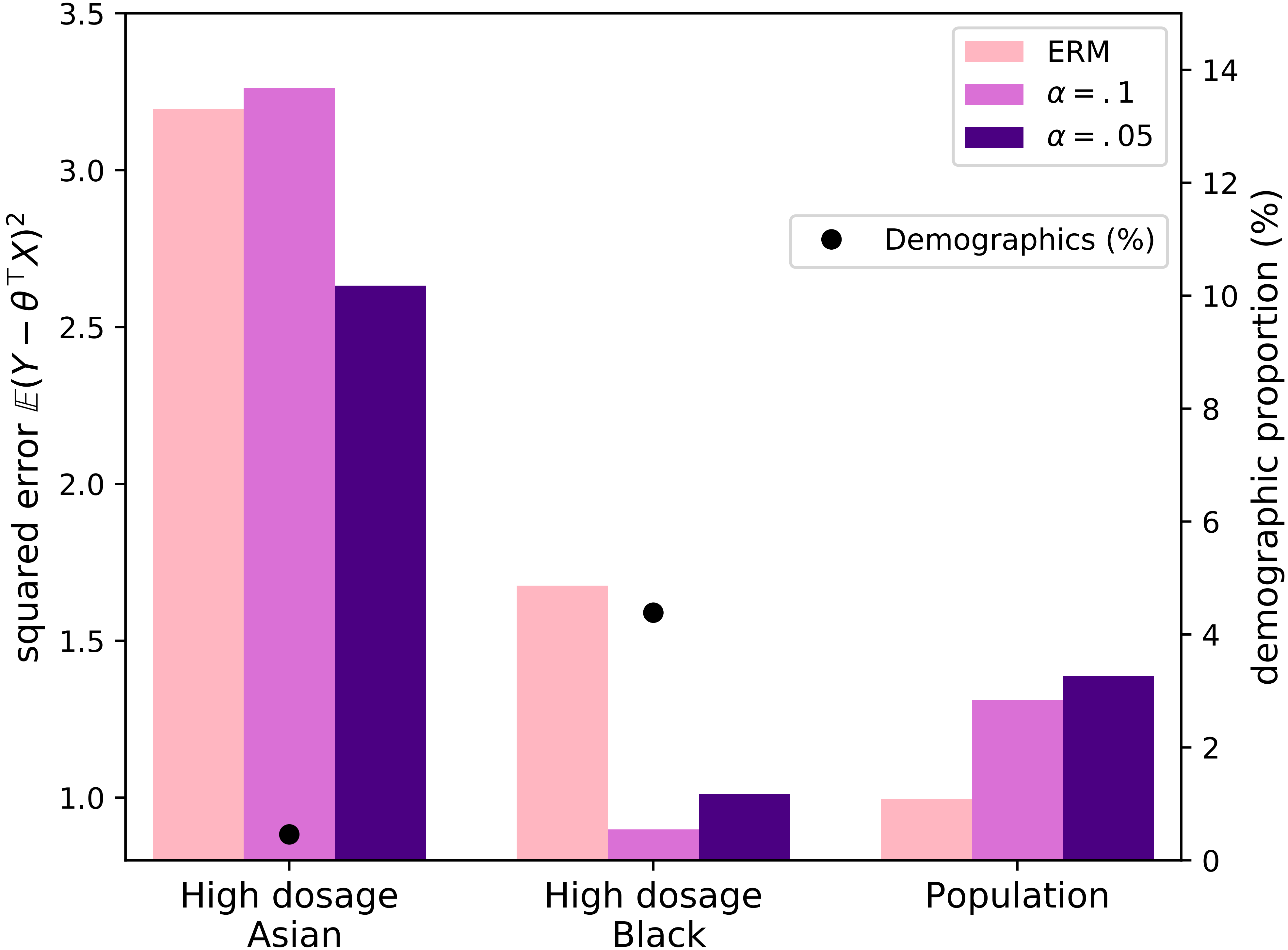
ERM suffered high prediction error on patients with high dosage

High warfarin dosage (>49mg)



$$f_k(t) \approx t^k - 1$$

High warfarin dosage (>49mg)



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

Takeaway: **Improved performance on hard subpopulation, slight deterioration in average-case**

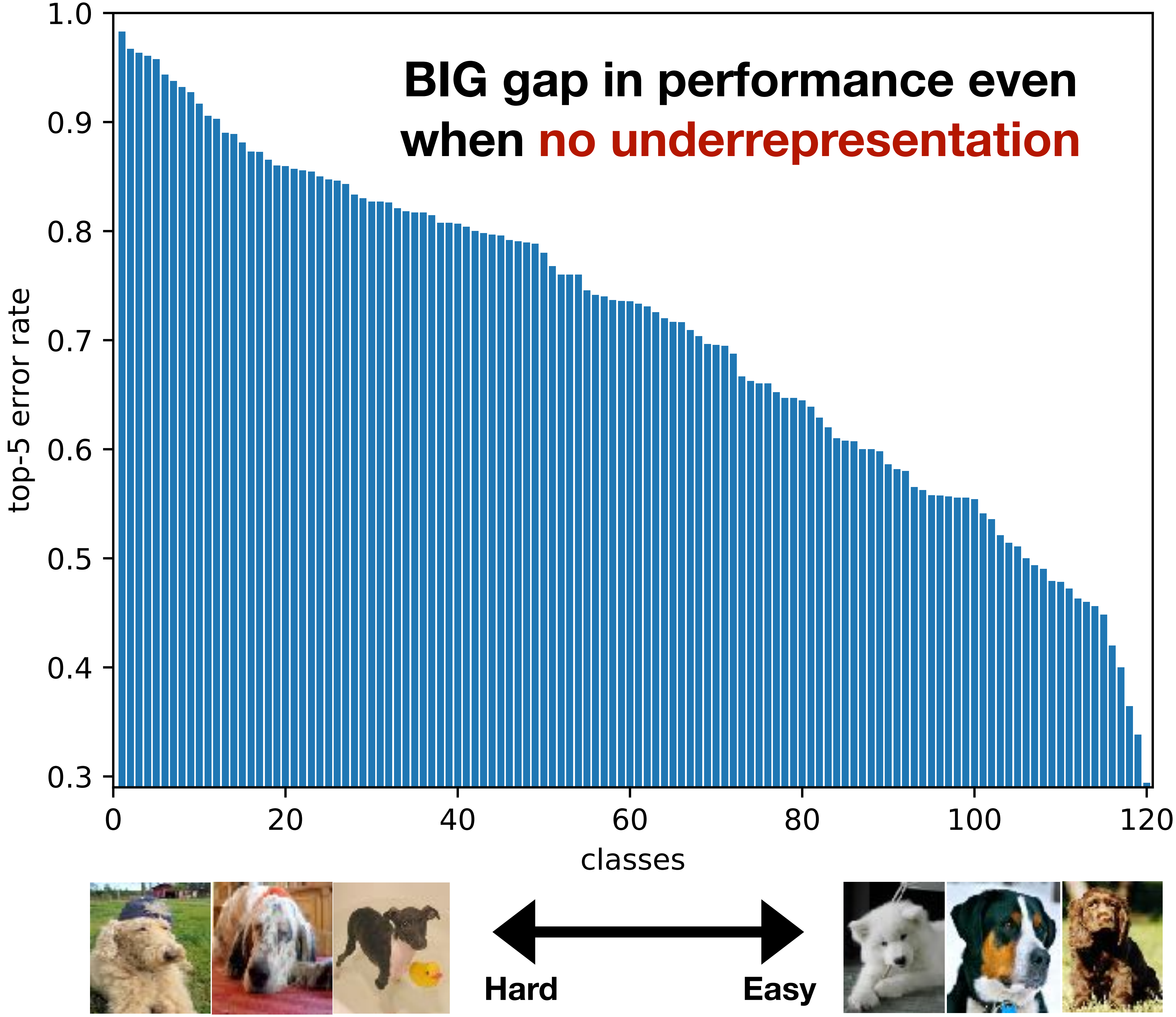
Fine-grained recognition

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

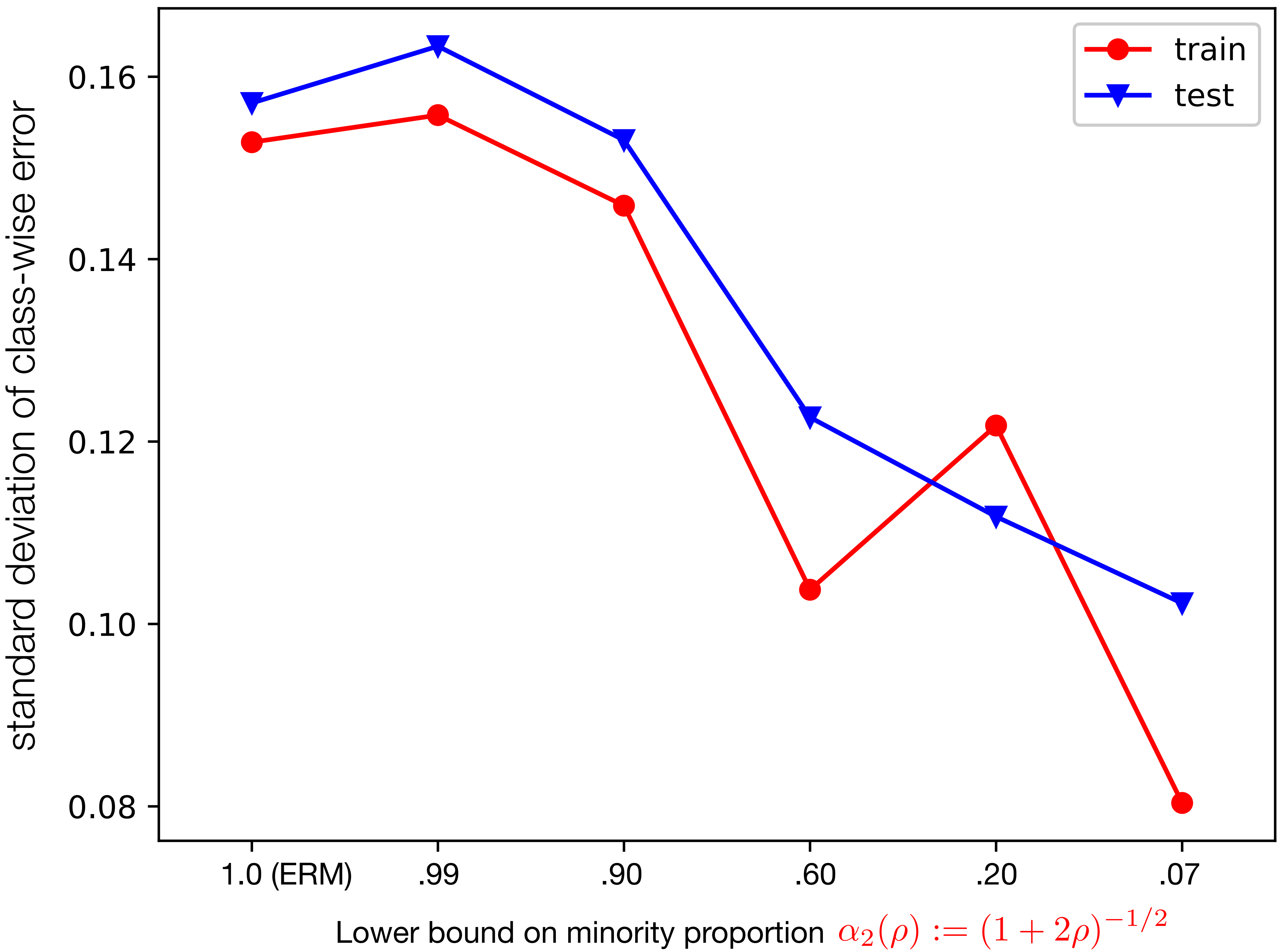
Stanford Dogs Dataset [Khosla et al. '11]

No underrepresentation:
same number of images per class

ERM error rate

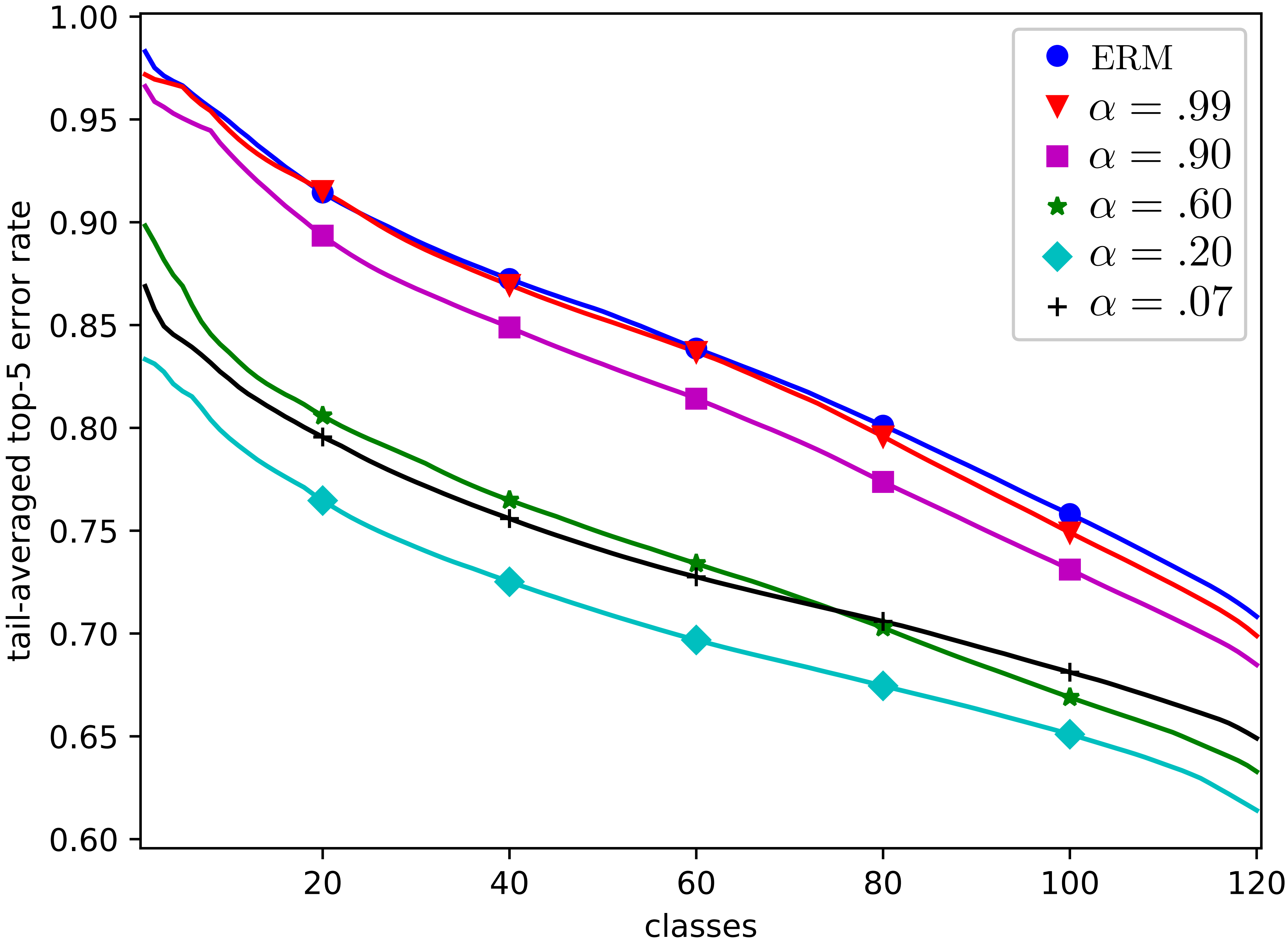


Variation in error over 120 class



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

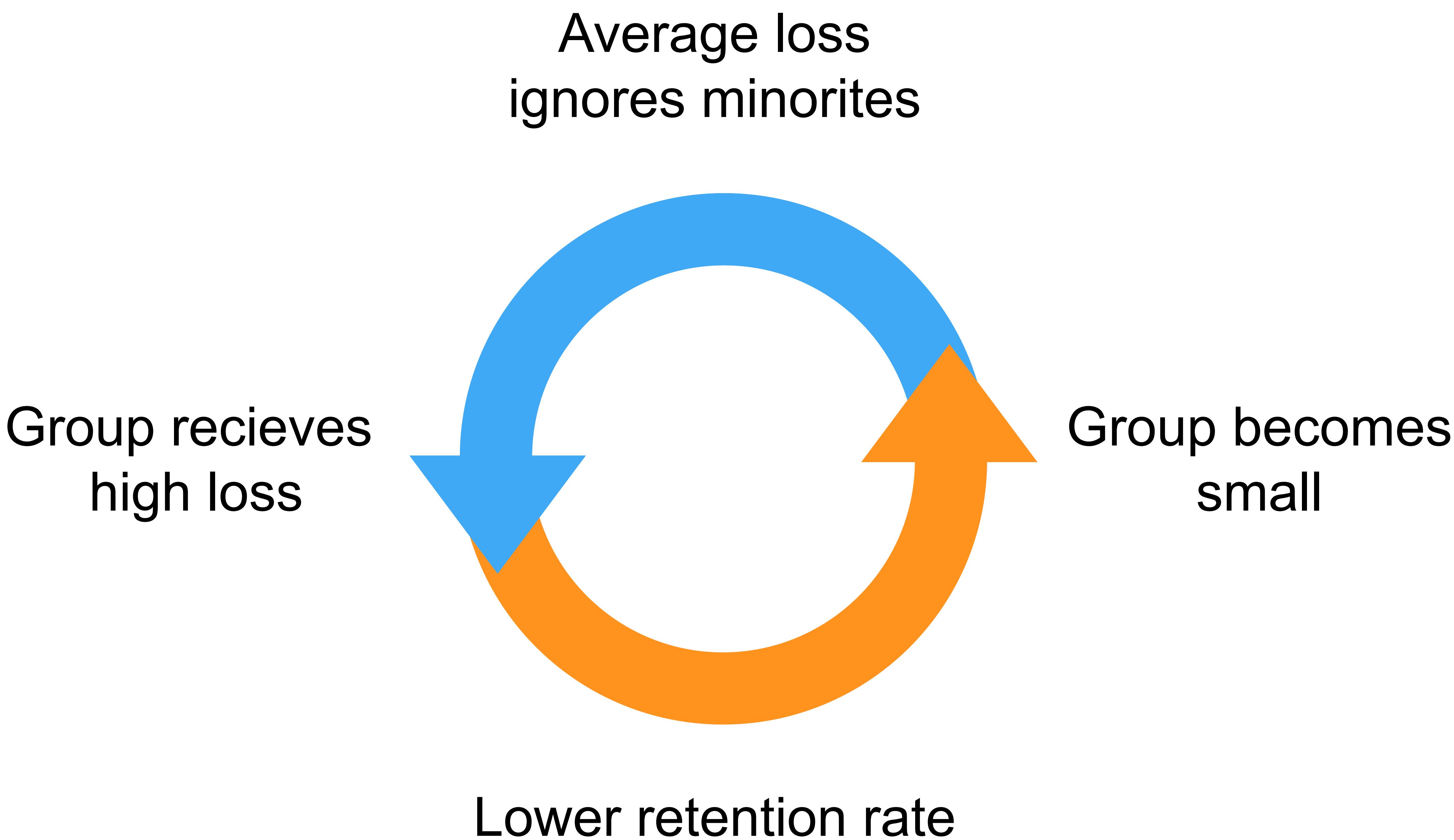
Worst x-classes



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

Takeaway: Guarantee uniform performance across dog breeds

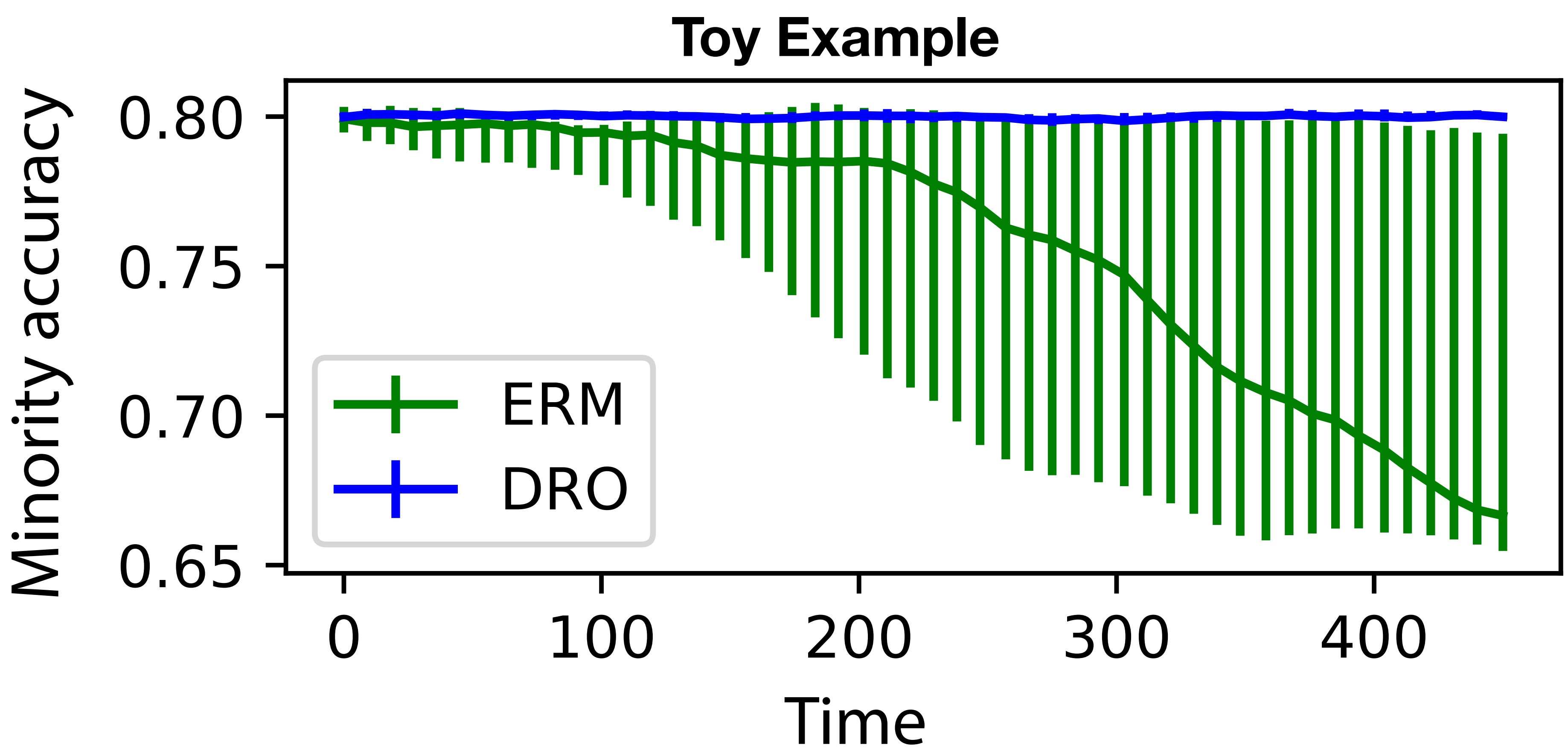
Repeated loss minimization



Problem: Degradation over time

Problem: Degradation over time

Small disparities can amplify to exacerbate subpopulation performance



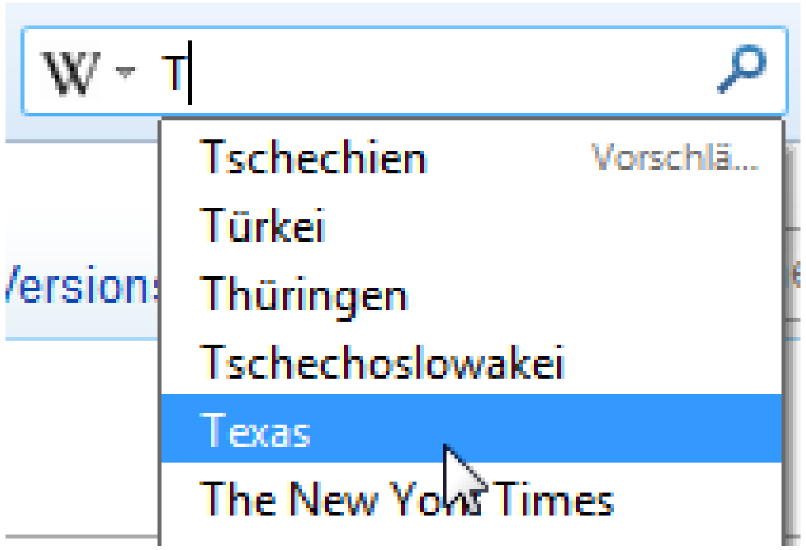
"Theorem" (HS<u>N</u>L'18) Under general user retention dynamics,

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)

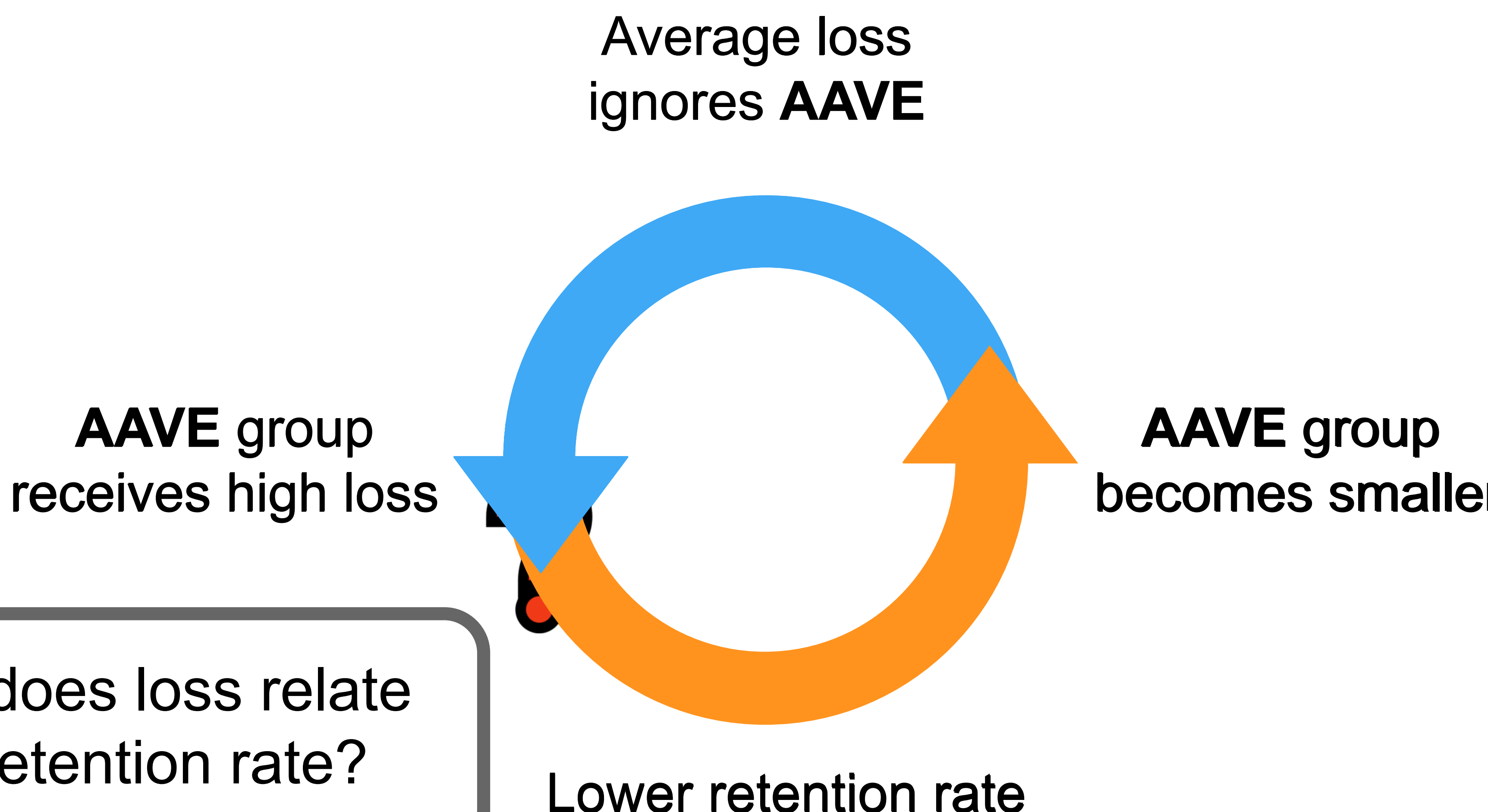
If u wit me den u pose to RESPECT ME

Standard American English (SAE)

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

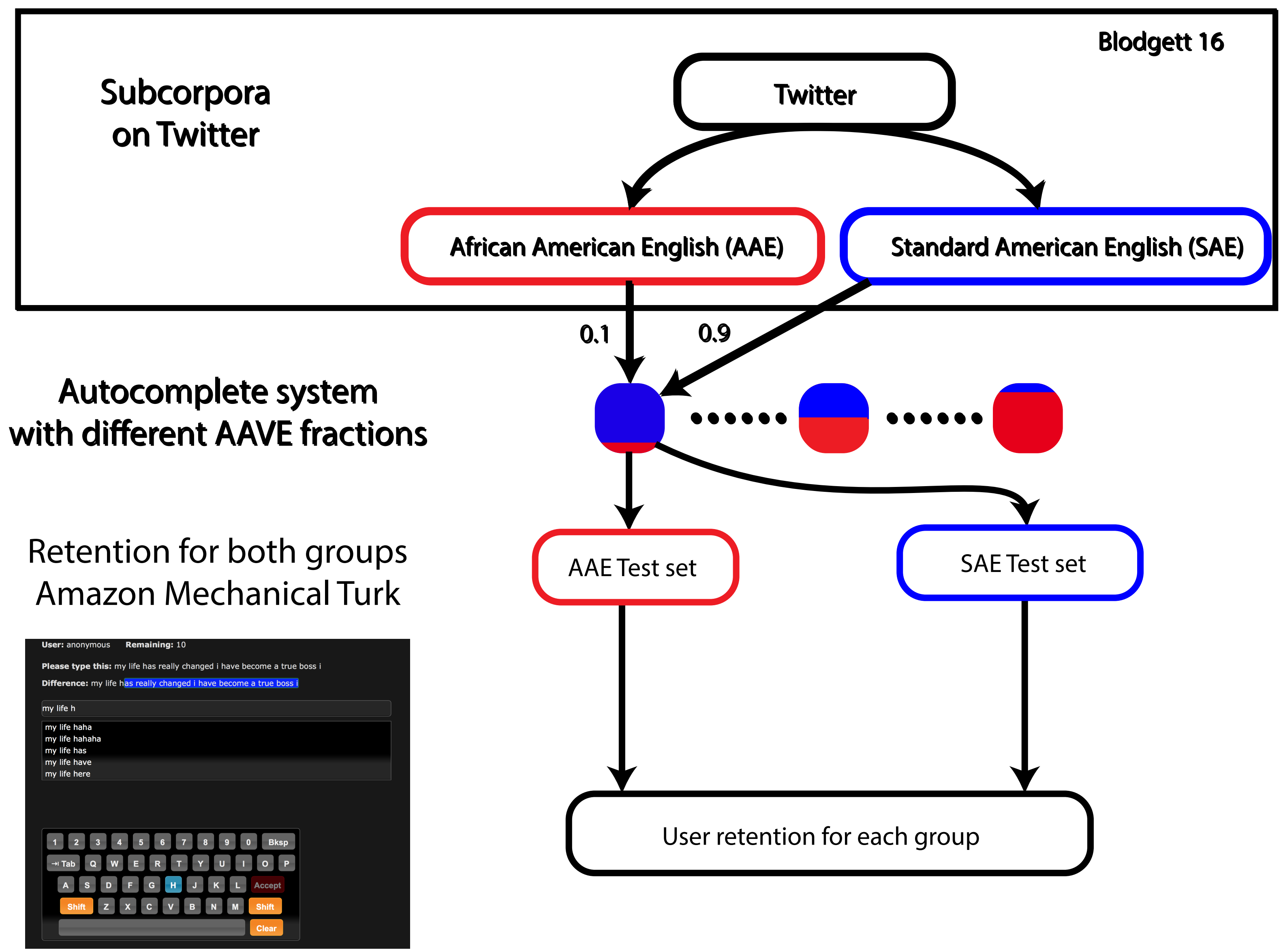
Experiment: Auto-complete

Retention feedback loop

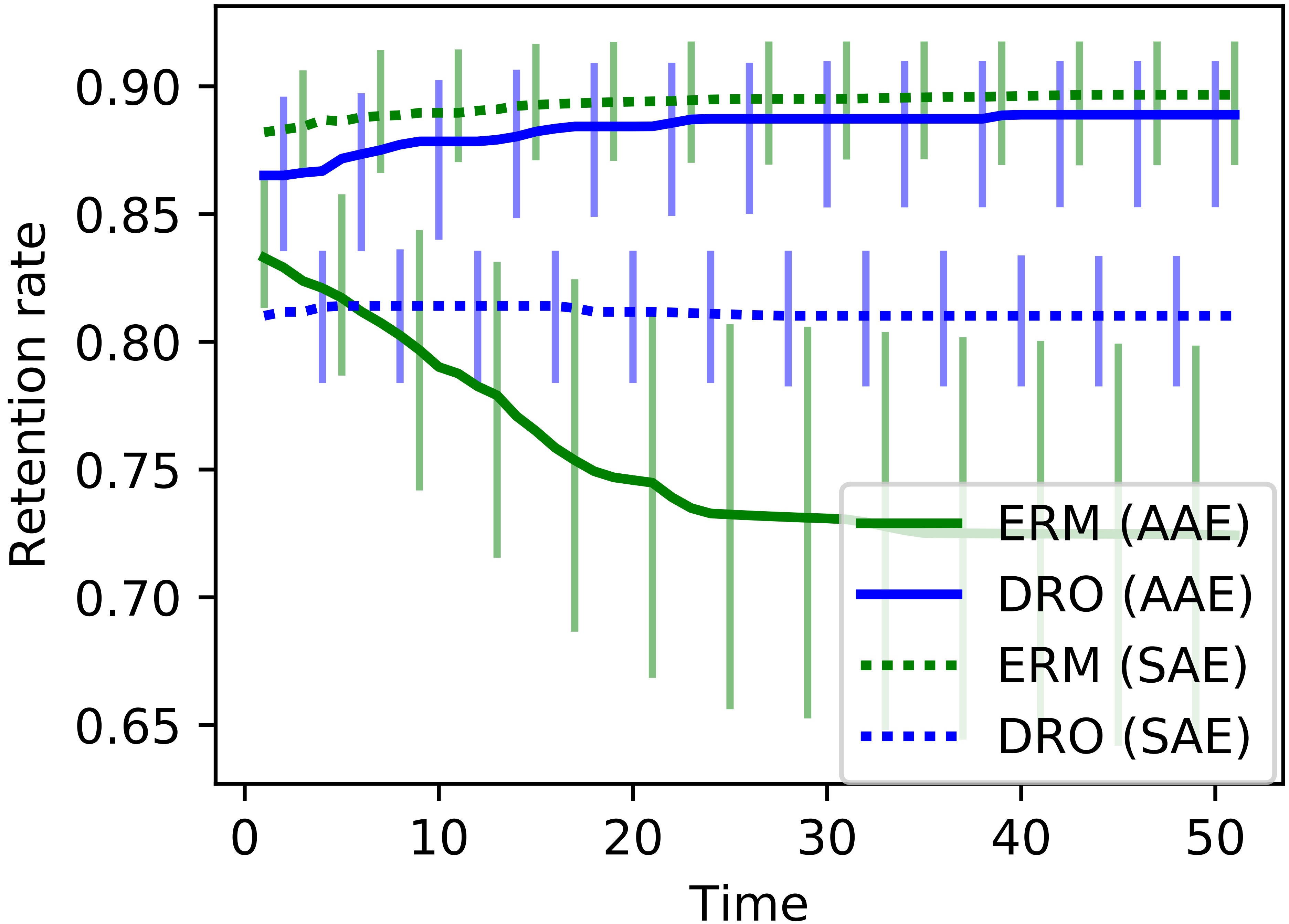


How does loss relate to retention rate?

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

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

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

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

DRO over covariate shift

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

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

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

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

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

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