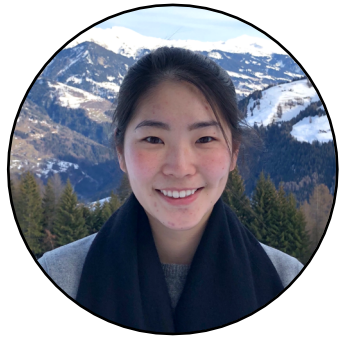
Distributionally Robust Neural Networks for Group Shifts:

On the Importance of Regularization for Worst-Case Generalization



Shiori Sagawa* Pang Wei Koh* Tatsunori B. Hashimoto Percy Liang

Models can latch onto spurious correlations

Misleading heuristics; might work on most training examples but may not always hold up

input x : bird image

label: bird type

waterbird vs landbird

Wah et al. (2011), Zhou et al. (2017)

Models can latch onto spurious correlations

Misleading heuristics; might work on most training examples but may not always hold up

input x : bird image

spurious correlation: water background

ML model

prediction $\hat{y}$: waterbird ✓

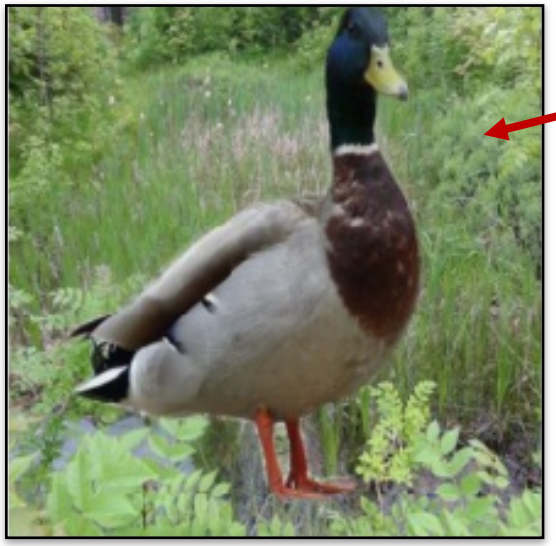
true label y: waterbird

Wah et al. (2011), Zhou et al. (2017)

Models can latch onto spurious correlations

Misleading heuristics; might work on most training examples but may not always hold up

input x : bird image



spurious correlation: land background

ML model

prediction $\hat{y}$: landbird

true label y: waterbird

Wah et al. (2011), Zhou et al. (2017)

Models can latch onto spurious correlations

input x : face image

ML model

label: hair color

blonde hair vs dark hair

Liu et al. (2015)

Models can latch onto spurious correlations

input x : face image

spurious correlation: gender

ML model

prediction $\hat{y}$: dark hair

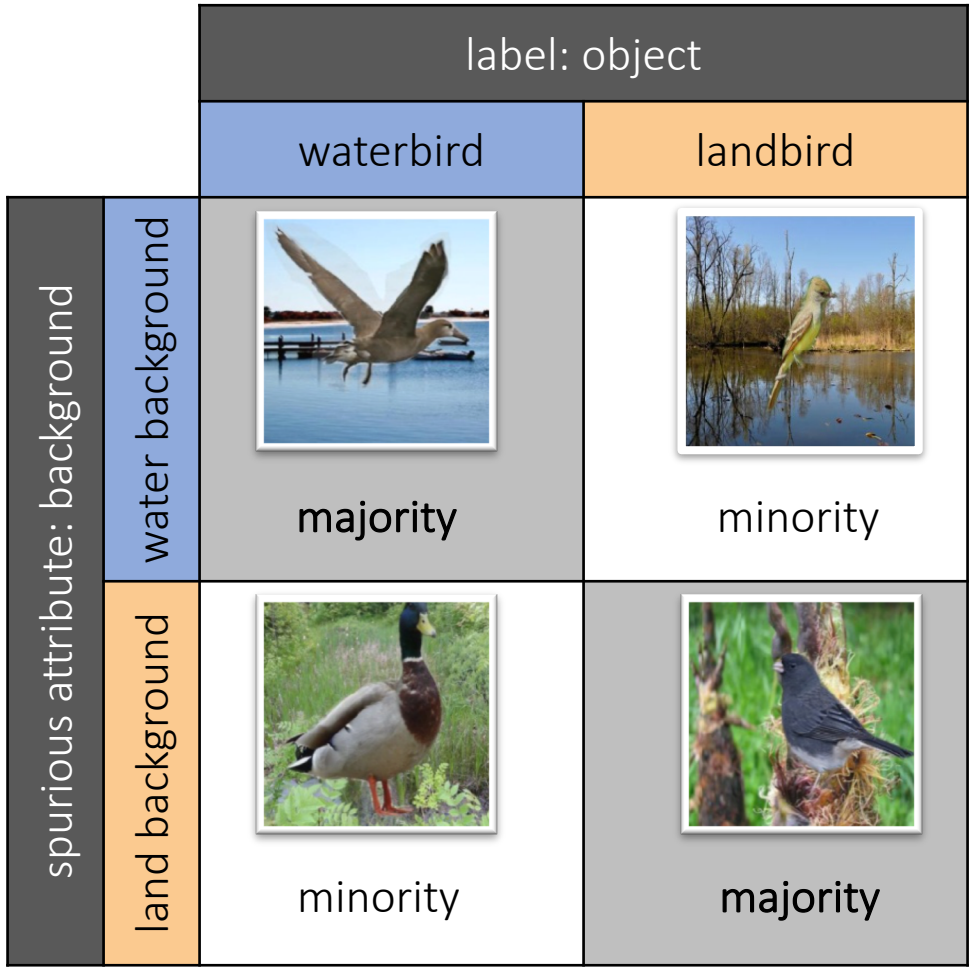
true label y: blonde hair

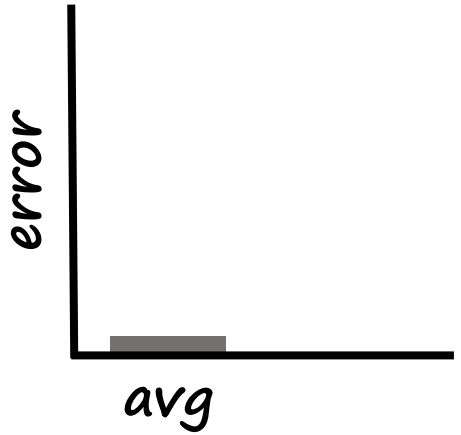
Liu et al. (2015)

Models can latch onto spurious correlations

spurious attribute: background		label: object: waterbird	label: object: landbird
	water background	 **majority**	 minority
	land background	 minority	 **majority**

Models perform well on average

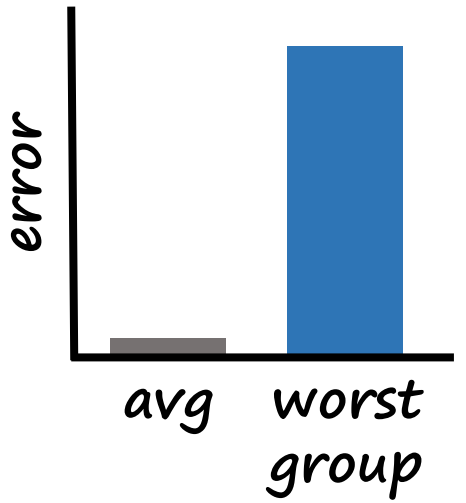




average error: 0.03

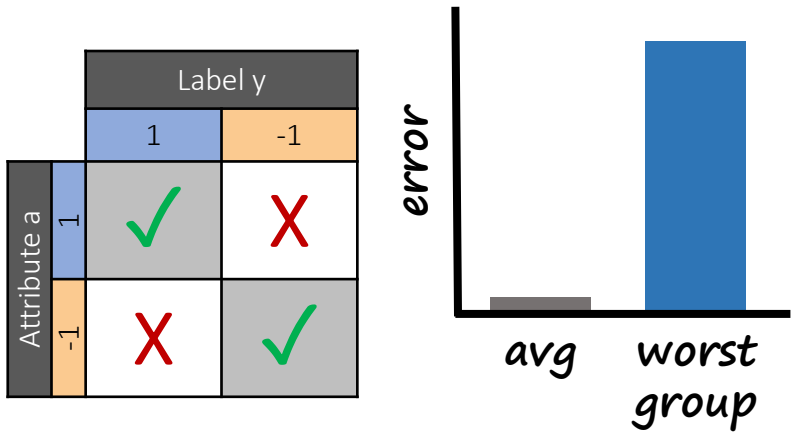
But models can have high worst-group error

spurious attribute: background	label: object: waterbird	label: object: landbird
water background	 0.05	 0.21
land background	 0.40	 0.004

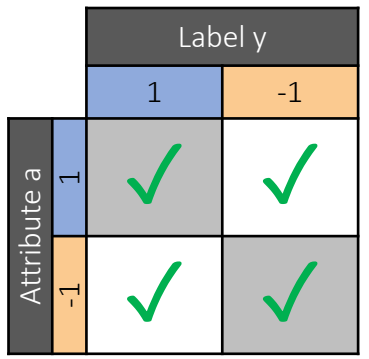


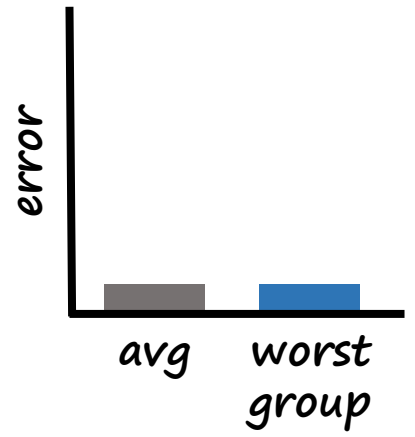
worst-group error: 0.40

Goal: low worst-group error



- Relies on spurious correlation
- High worst-group error



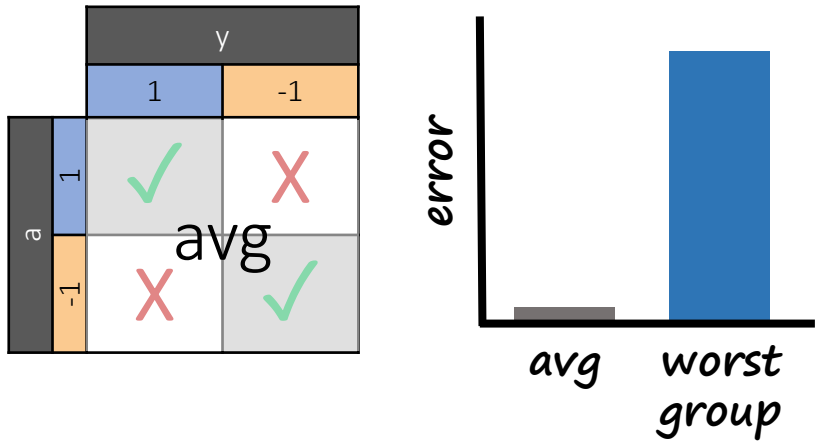


- More robust to spurious correlation
- Low worst-group error

Our approach: minimize the worst-group loss

Standard (ERM): average loss

$$\mathcal{R}_{\text{ERM}}(w) = \hat{\mathbb{E}}_{(x,y,g)}\left[\ell(w;(x,y))\right]$$

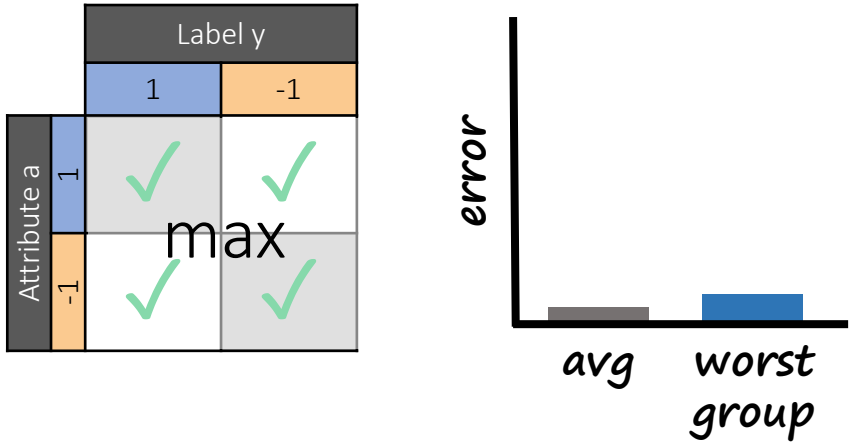


Group DRO: worst-group loss

$$\mathcal{R}_{\text{gDRO}}(w) = \max_{g' \in \mathcal{G}} \hat{\mathbb{E}}_{(x,y,g)}\left[\ell(w;(x,y)) \mid g = g'\right]$$

worst-group

average loss for each group g'

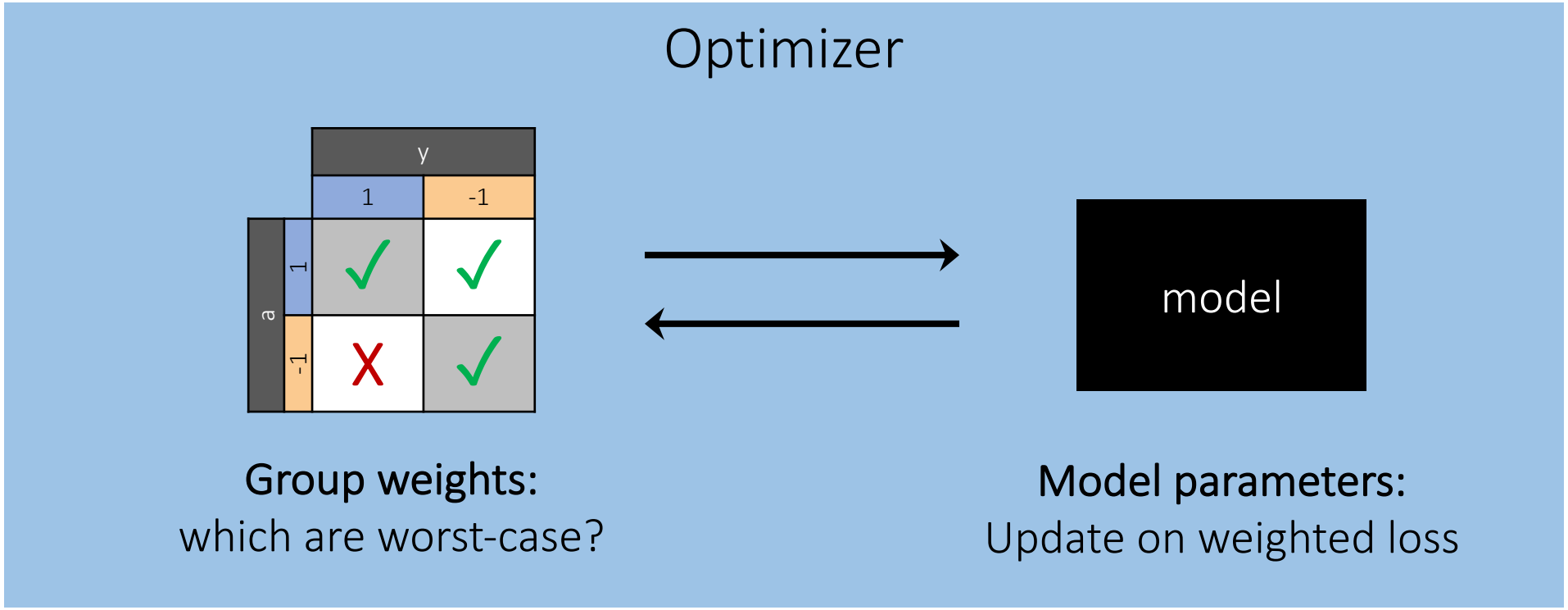


Train: known groups for each example

Test: unknown groups

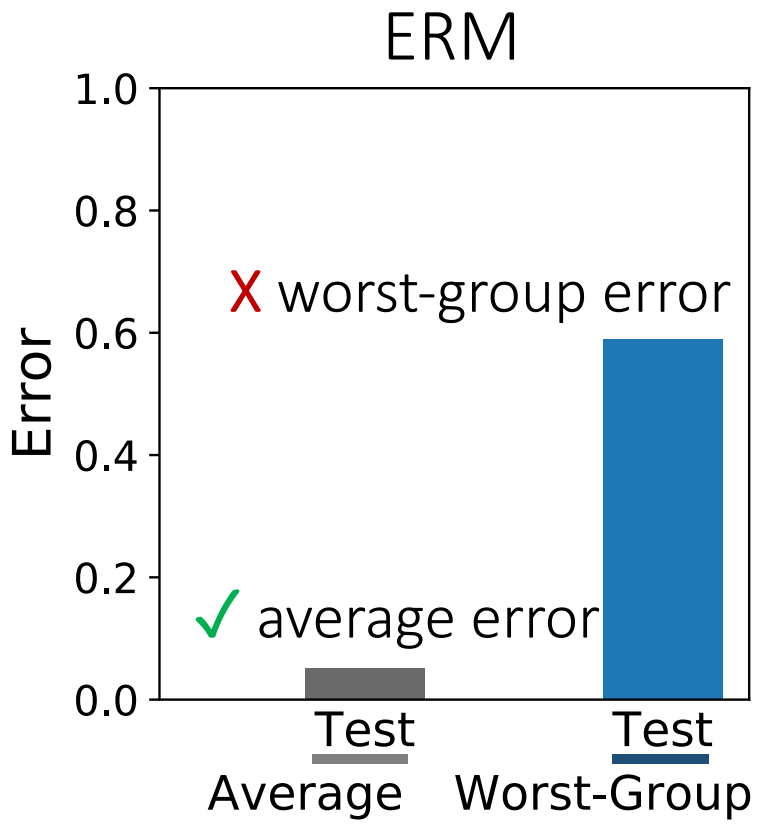
Hu, Niu, Sato, Sugiyama (2018), Oren, Sagawa, Hashimoto, Liang (2019)

Optimization algorithm for Group DRO

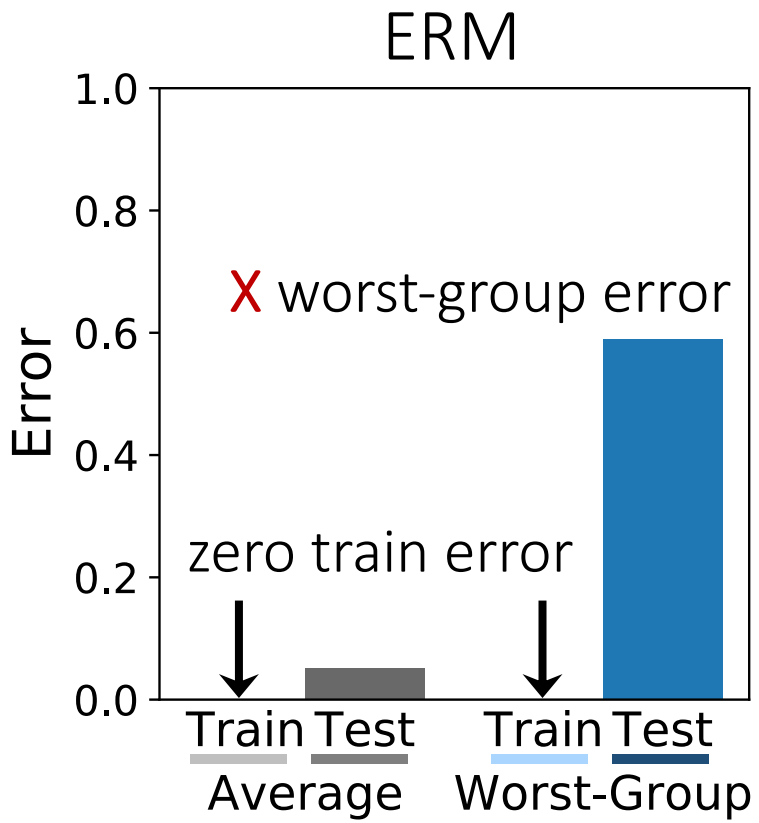


✓ Scalable
✓ Theoretical guarantees
✓ Similar # iterations to convergence as ERM

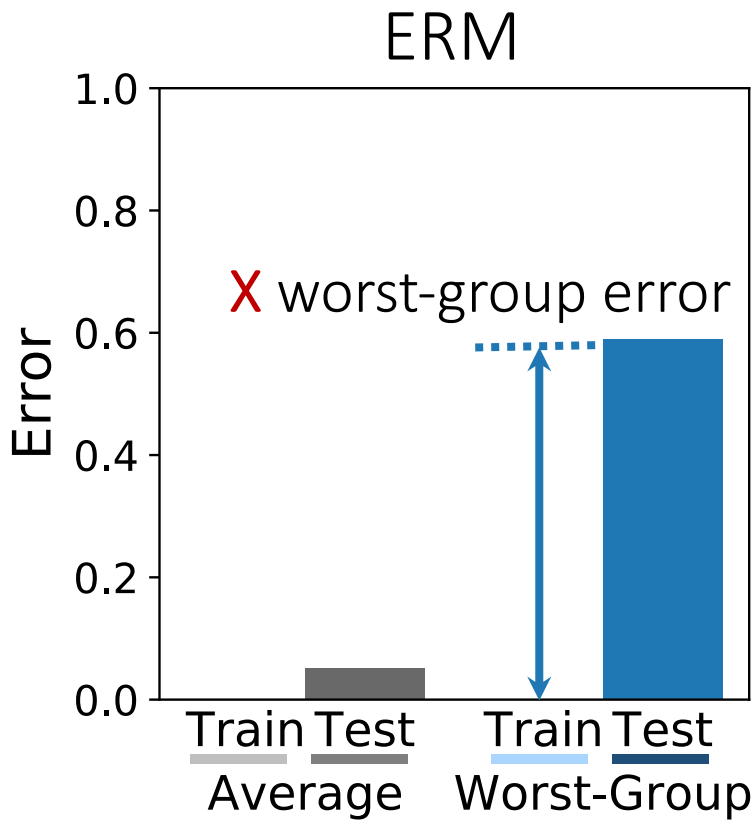
Attempt 1: ERM → high worst-group test error



Attempt 1: ERM → high worst-group test error

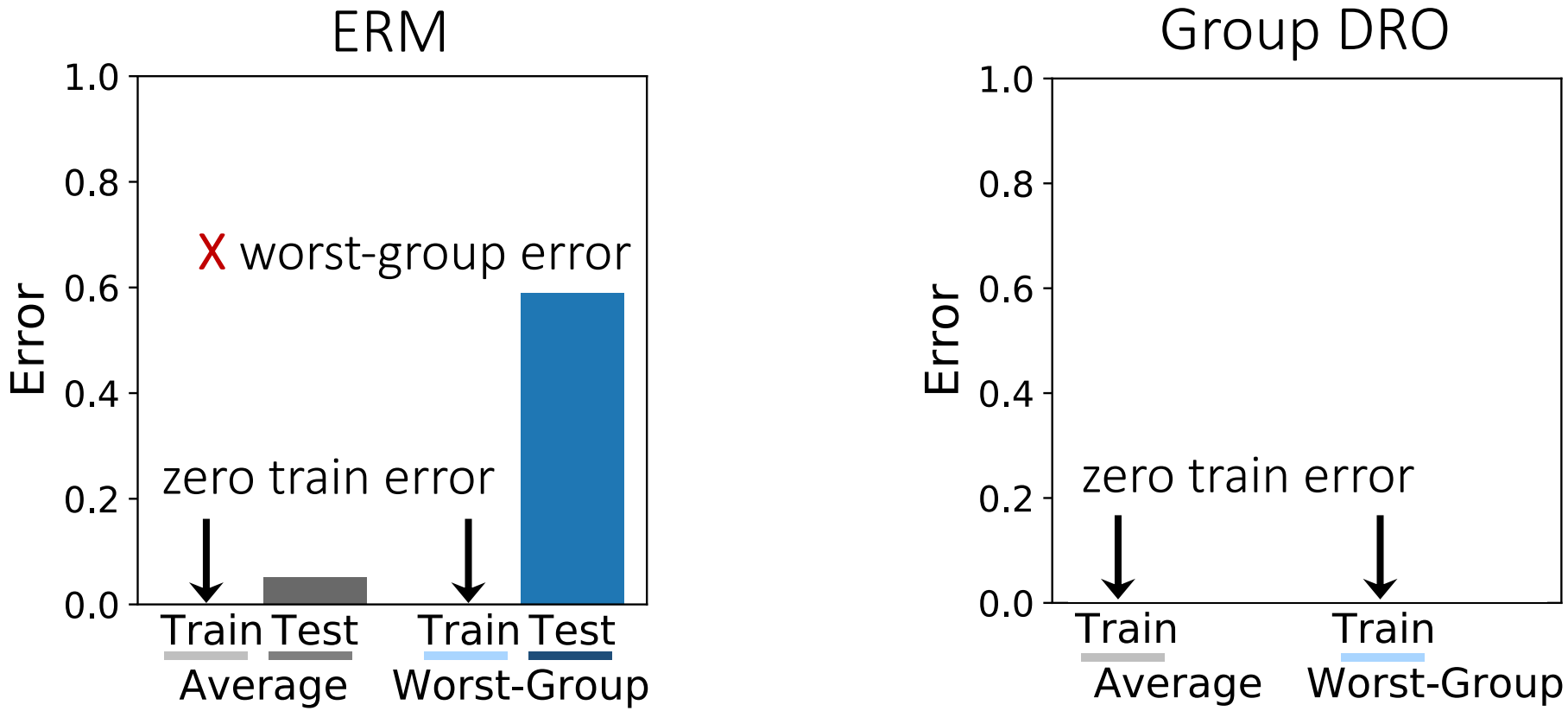


Attempt 1: ERM → high worst-group test error



worst-group error is high because of poor generalization

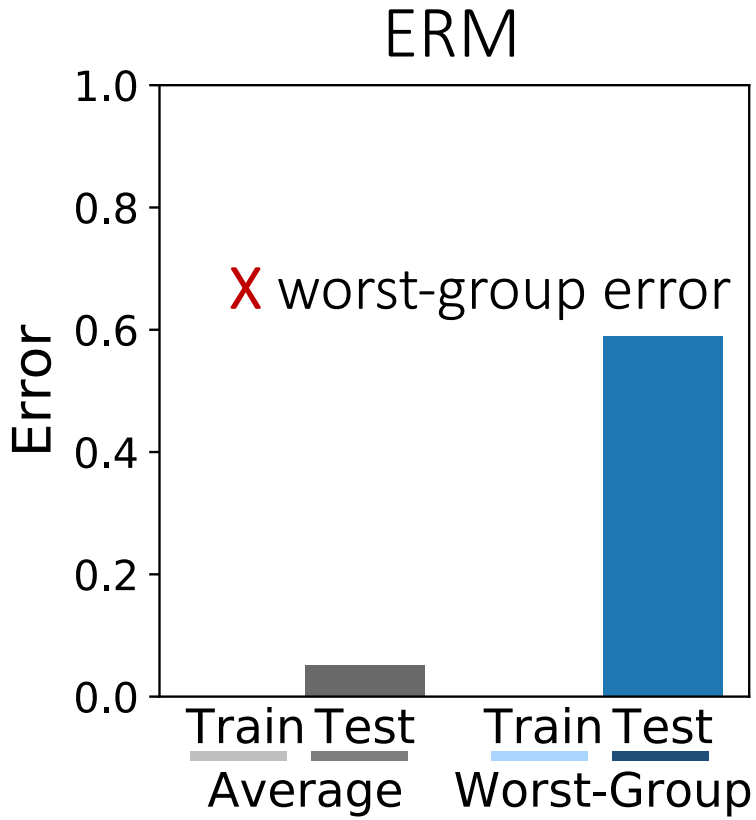
Attempt 1: zero training error → ERM ≈ group DRO

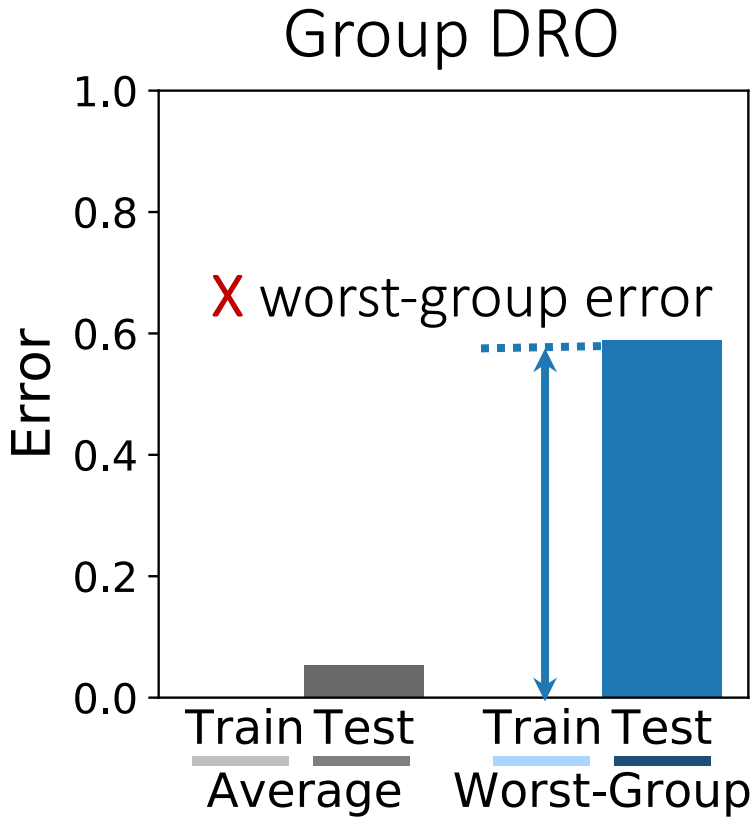


worst-group error is high because of poor generalization

but group DRO only controls *training* error!

Attempt 1: poor generalization → group DRO fails





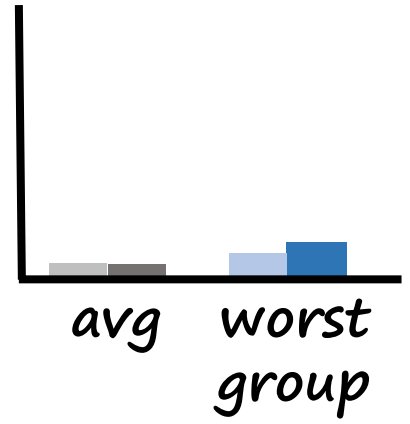
worst-group error is high because of poor generalization

but group DRO only controls *training* error!

New challenge: train error ≉ test error on worst group

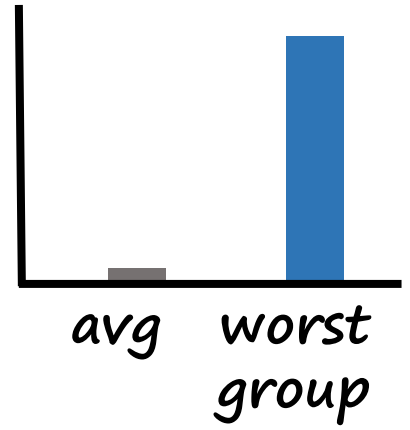
Prior work: train error ≈ test error for worst-case group

- Small convex or generative models



Our setting: high worst-group test error despite zero train error

- State-of-the-art neural networks

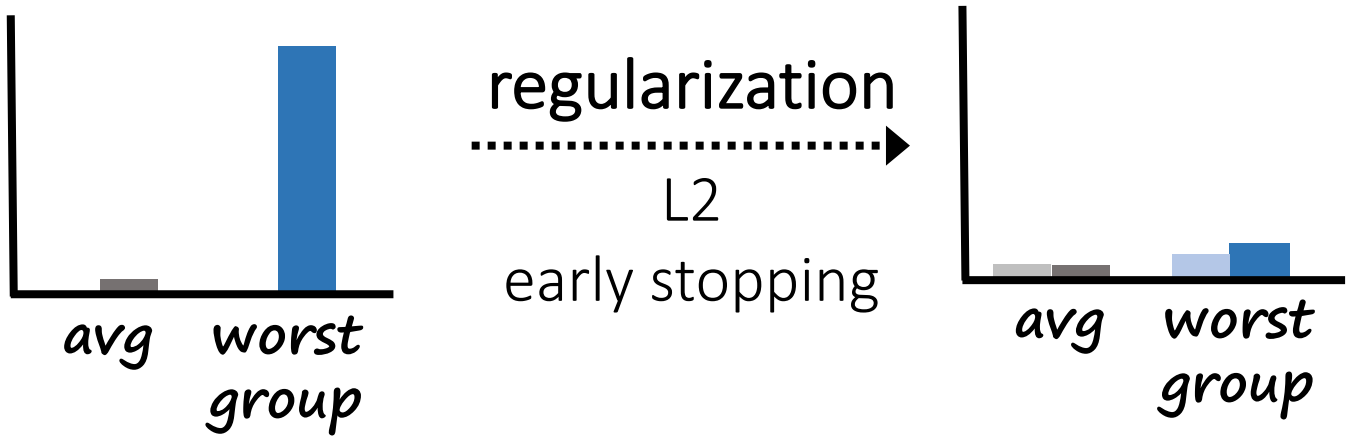


Ben-Tal, et al. (2013), Duchi and Namkoong (2018), Hu, Niu, Sato, Sugiyama (2018), Oren, Sagawa, Hashimoto, Liang (2019)

Approach: regularization + group DRO

Problem: zero *train* error, but high worst-group *test* error

Solution: regularization

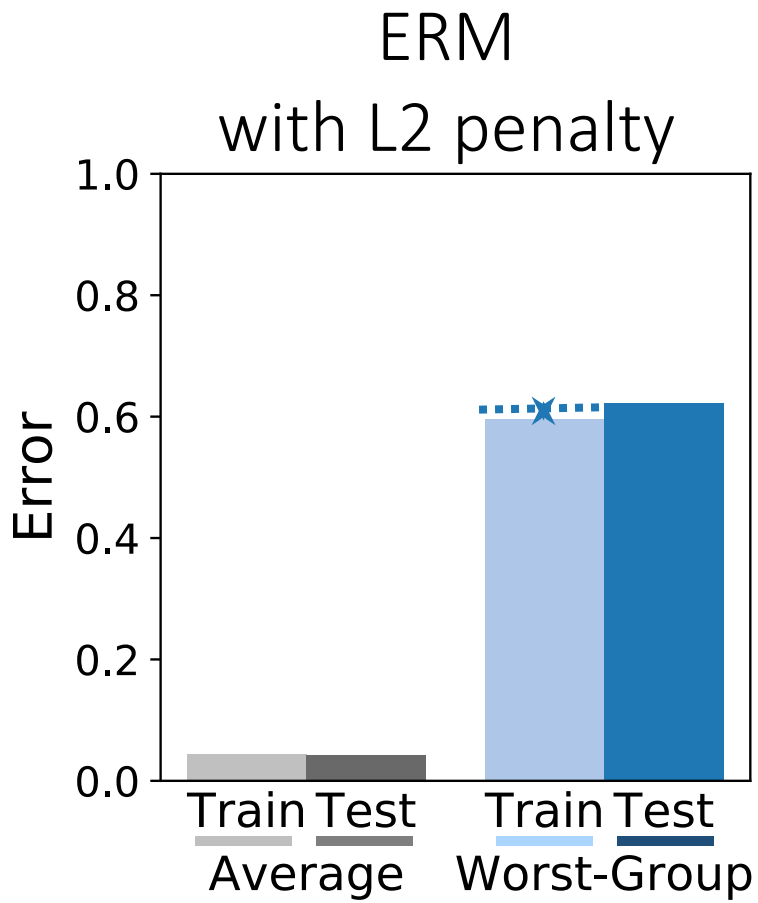


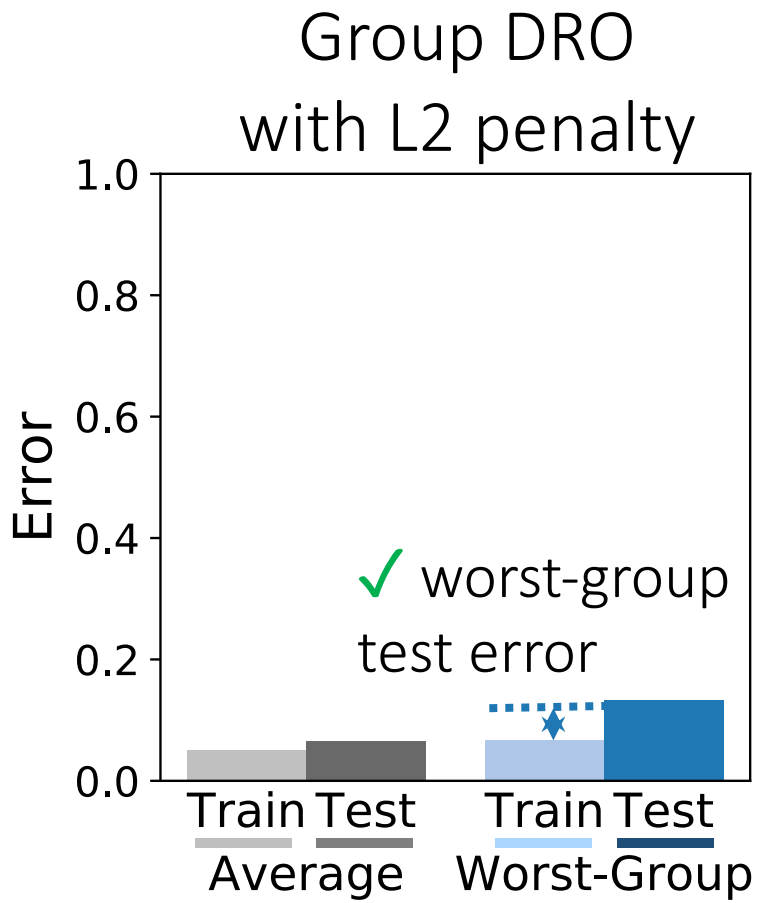
Counterintuitive with respect to recent trends:

More complex models with zero training error → better average error

Hoffer, Hubara, Soudry (2017), Belkin et al. (2019), Nakkiran et al. (2020)

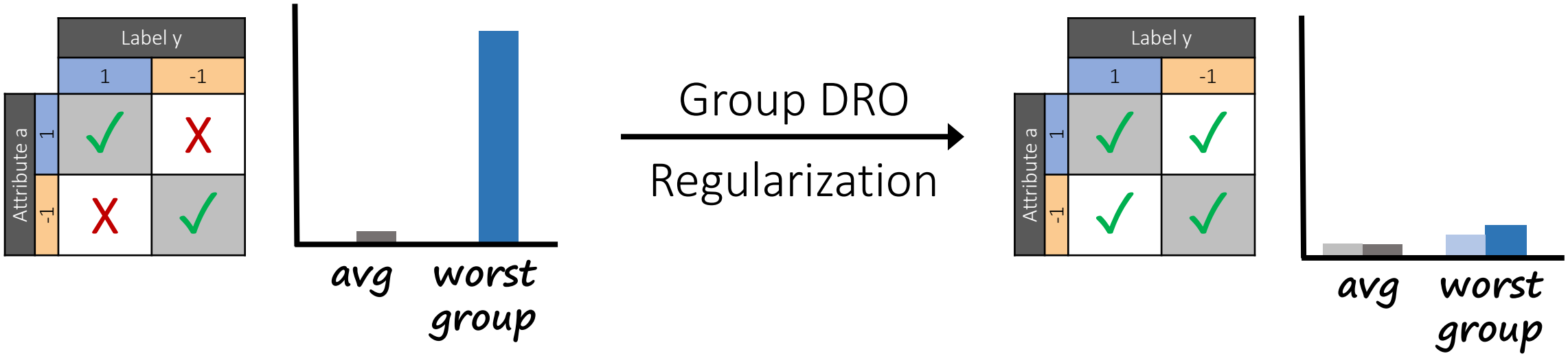
Attempt 2: regularization + group DRO works



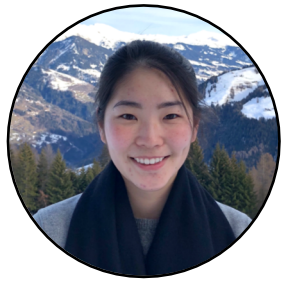


Group DRO + regularization mitigates the spurious correlation problem

Goal: low worst-group error



Thanks!



Shiori Sagawa*

Pang Wei Koh*

Tatsunori B. Hashimoto

Percy Liang

Thank you to Shyamal Buch, Yair Carmon, Zhenghao Chen, John Duchi, Jean Feng, Christina Heinze-Deml, Robin Jia, Daphne Koller, Ananya Kumar, Tengyu Ma, Jesse Mu, Hongseok Namkoong, Emma Pierson, and Fanny Yang

Funded by Toyota Research Institute (TRI), Open Philanthropy Project Award, Stanford Graduate Fellowship, and Facebook Fellowship Program.