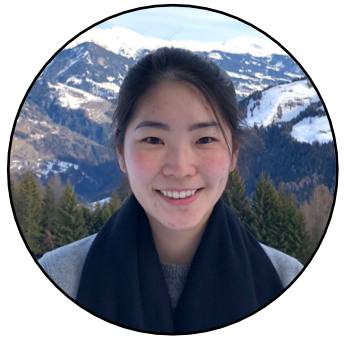
An Investigation of Why Overparameterization Exacerbates Spurious Correlations




Shiori Sagawa* Aditi Raghunathan* Pang Wei Koh* 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

ML model

label: bird type

waterbird vs landbird

Sagawa et al. (2020), 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 ✓

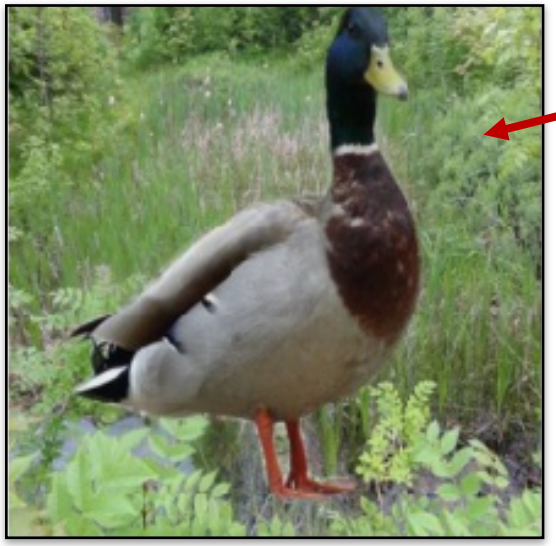
true label y: waterbird

Sagawa et al. (2020), 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

Sagawa et al. (2020), 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

Sagawa et al. (2020), Liu et al. (2015)

Models can latch onto spurious correlations

input x : face image

spurious correlation: gender

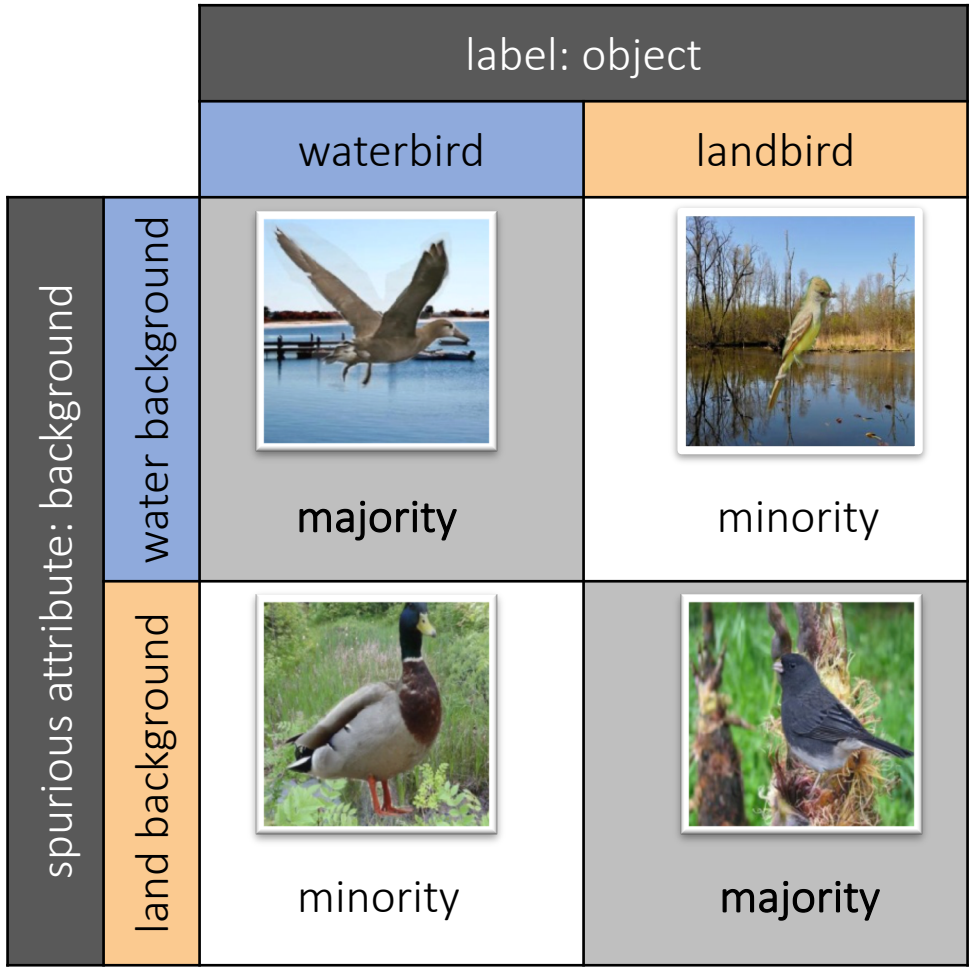
ML model

prediction $\hat{y}$: dark hair

true label y: blonde hair

Sagawa et al. (2020), Liu et al. (2015)

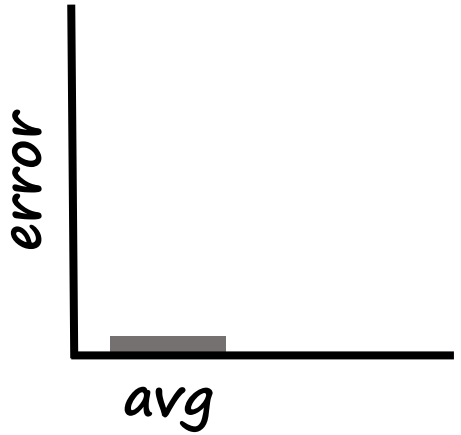
Models can latch onto spurious correlations



Sagawa et al. (2020)

Models perform well on average

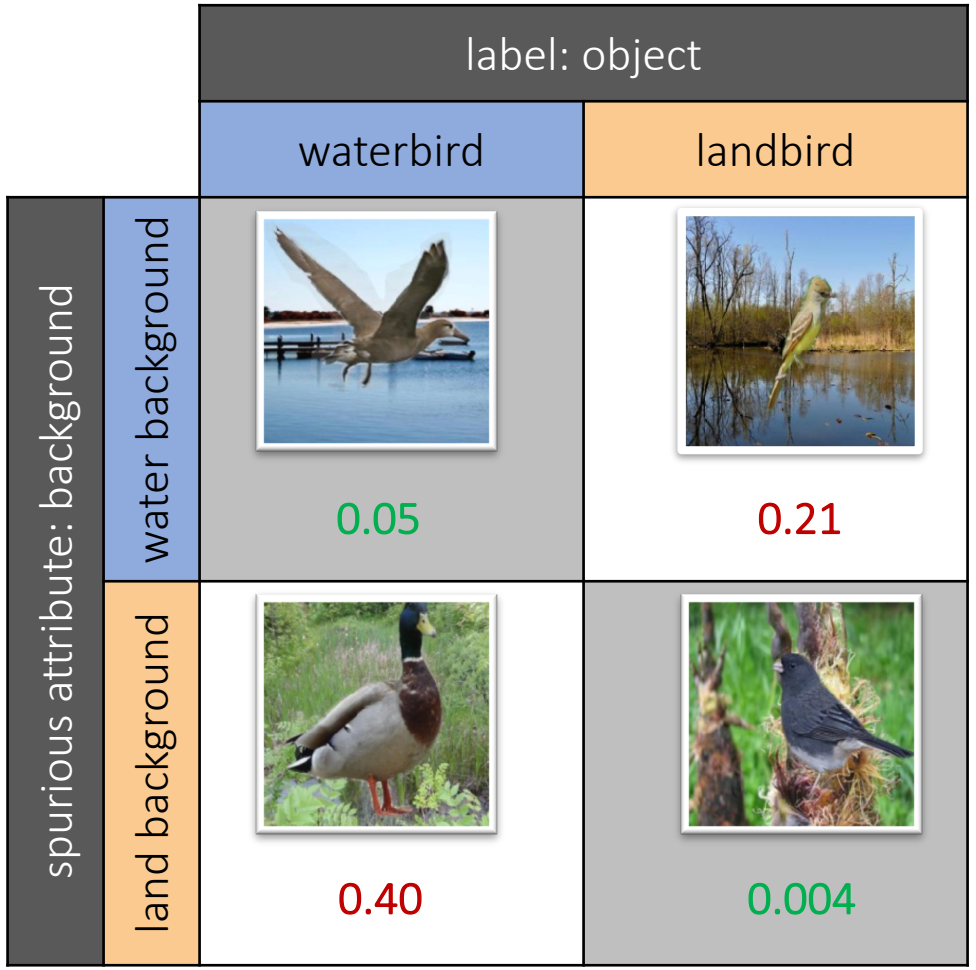
spurious attribute: background	label: object: waterbird	label: object: landbird
water background	 0.05	 0.21
land background	 0.40	 0.004

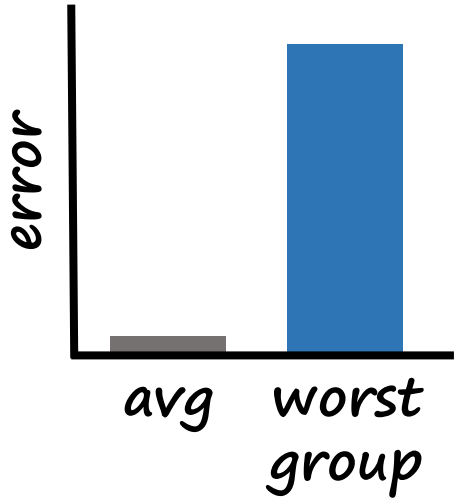


average error: 0.03

Sagawa et al. (2020)

But models can have high worst-group error





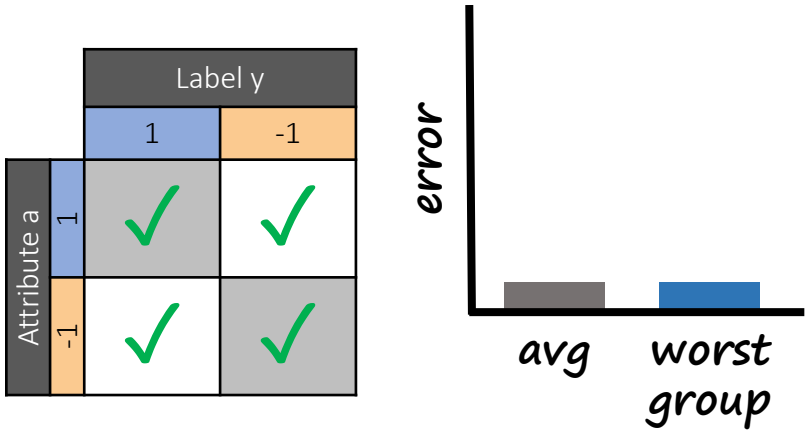
worst-group error: 0.40

Sagawa et al. (2020)

Approaches for improving worst-group error fail on high-capacity models

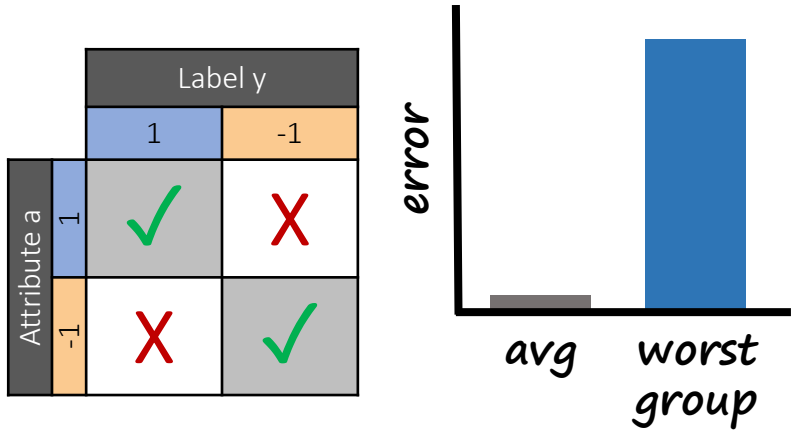
- Upweight minority groups: $\mathcal{R}_{\text{reweight}}(w) = \hat{E}_{(x,y,g)}\left[\frac{1}{\hat{p}_g}\ell(w,(x,y))\right]$

Low-capacity models



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

High-capacity models



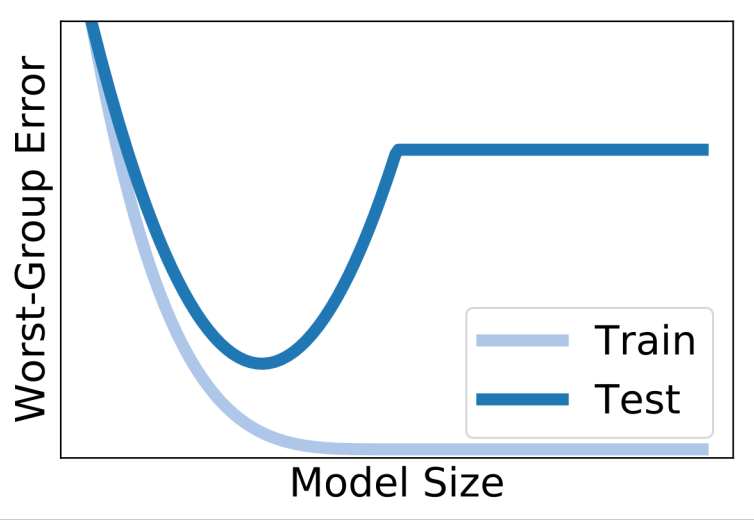
- Relies on spurious correlation
- High worst-group error

Sagawa et al. (2020)

Overparameterization hurts worst-group error for models trained with the reweighted objective

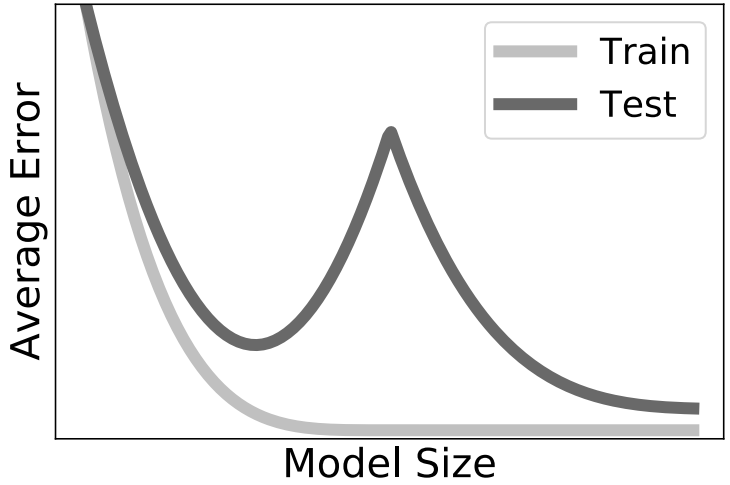
worst-group error

Overparameterized is *worse* than underparameterized



average error

Overparameterized is *better* than underparameterized

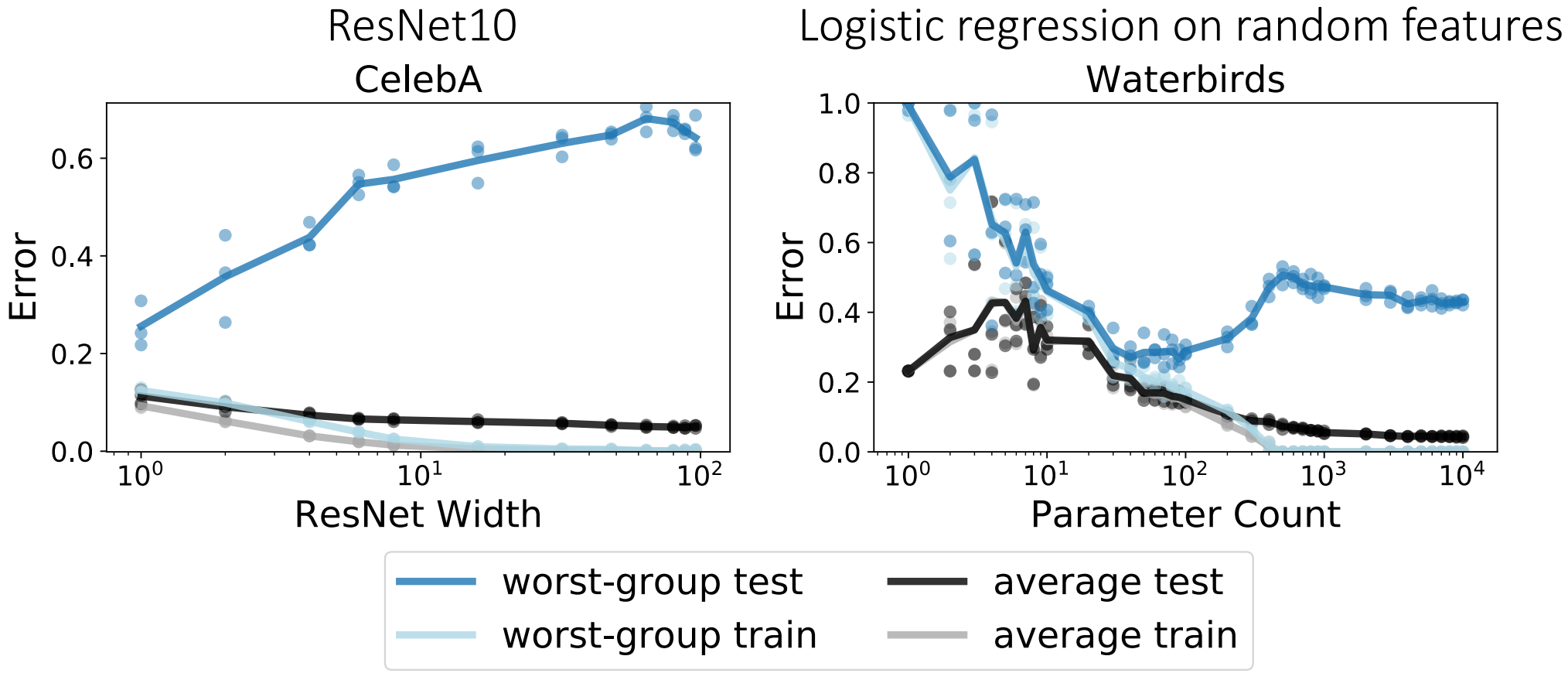


Our work: why does overparameterization exacerbate worst-group error?

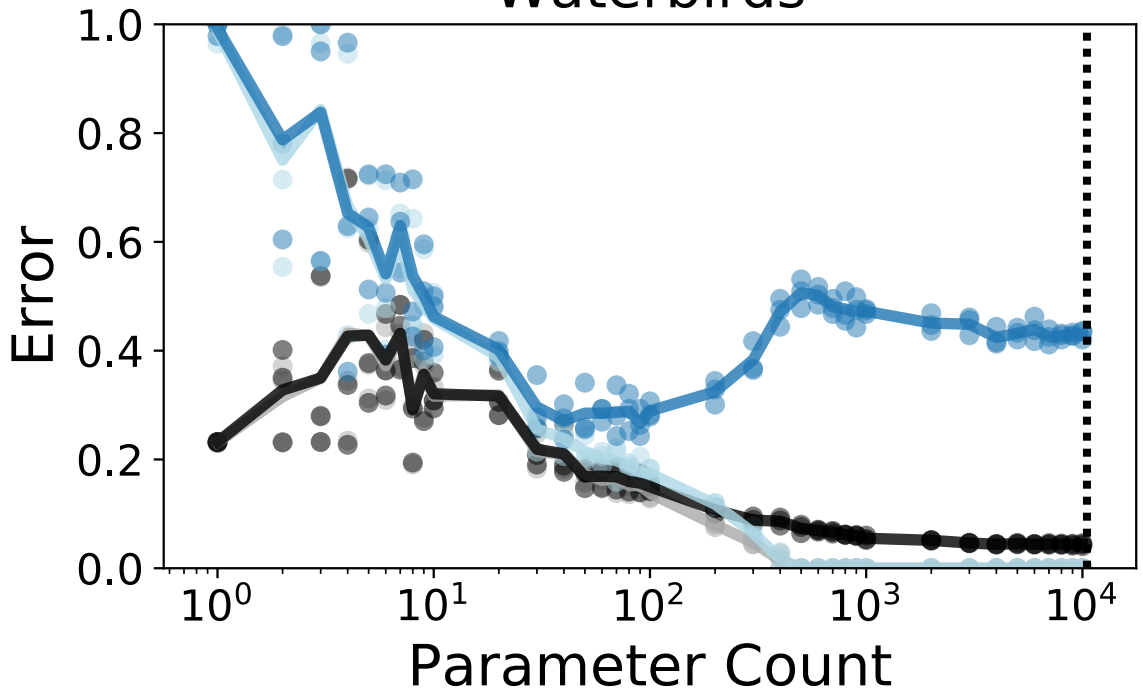
Overview

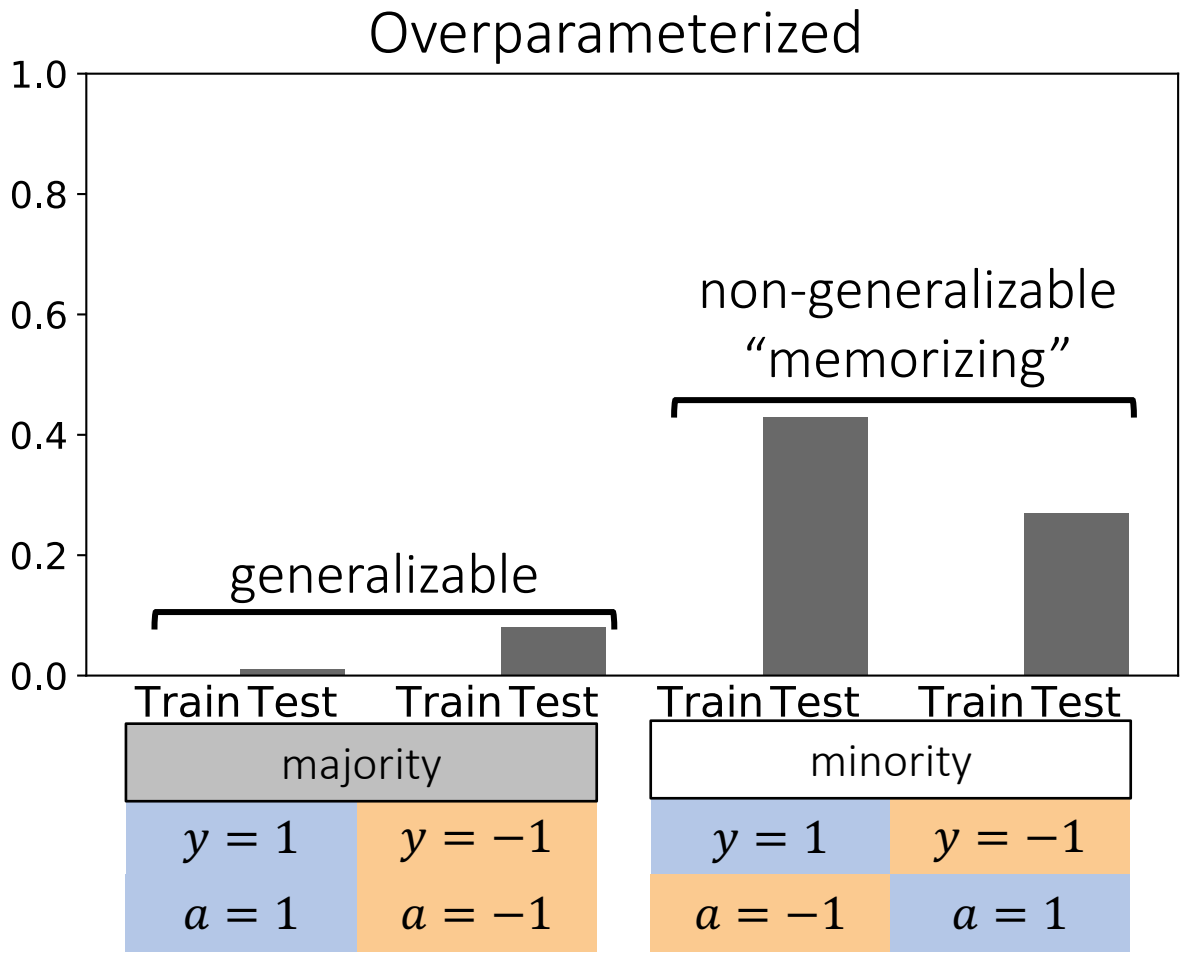
1. Empirical results
2. Analytical model and theoretical results
3. Subsampling

Overparameterization exacerbates worst-group error



Intuition: overparameterized models learn the spurious attribute and memorize minority groups

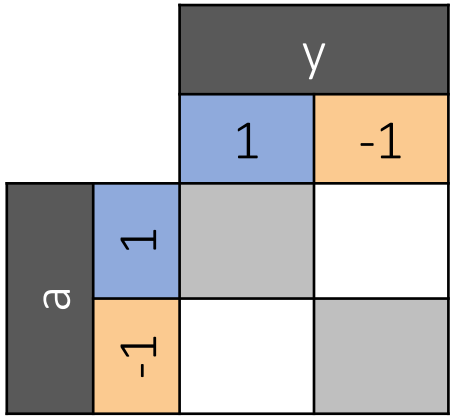




Overview

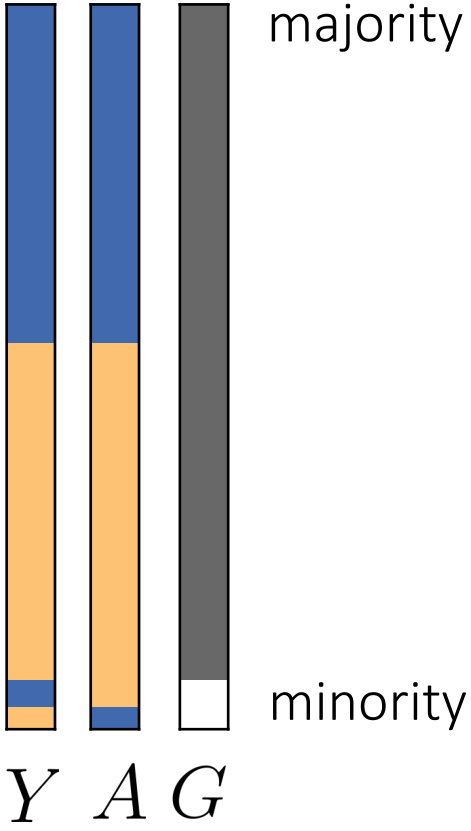
1. Empirical results
2. Analytical model and theoretical results
3. Subsampling

Toy example: data



Majority fraction

$$p_{\text{maj}} = \frac{n_{\text{maj}}}{n}$$



Toy example: data

$$x = [x_{\text{core}}, x_{\text{spu}}, x_{\text{noise}}]$$

$$x_{\text{core}} \in \mathbb{R}$$

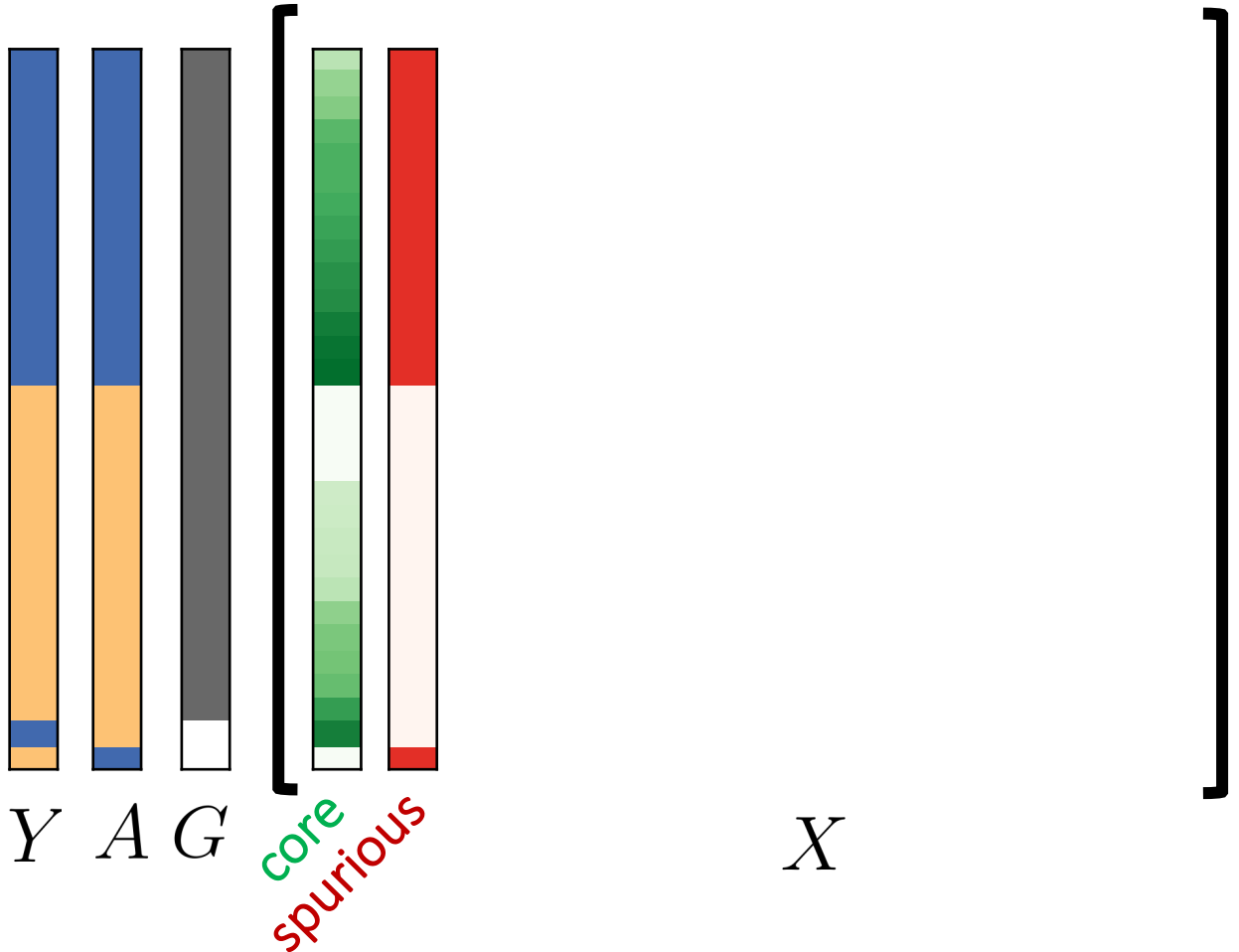
$$x_{\text{core}} \mid y \sim \mathcal{N}(y, \sigma^2_{\text{core}})$$

$$x_{\text{spu}} \in \mathbb{R}$$

$$x_{\text{spu}} \mid a \sim \mathcal{N}(a, \sigma^2_{\text{spu}})$$

Spurious-to-core information ratio (SCR)

$$r_{\text{s:c}} = \frac{\sigma^2_{\text{core}}}{\sigma^2_{\text{spu}}}$$



Toy example: data

$$x = [x_{\text{core}}, x_{\text{spu}}, x_{\text{noise}}]$$

$$x_{\text{core}} \in \mathbb{R}$$

$$x_{\text{core}} \mid y \sim \mathcal{N}(y, \sigma^2_{\text{core}})$$

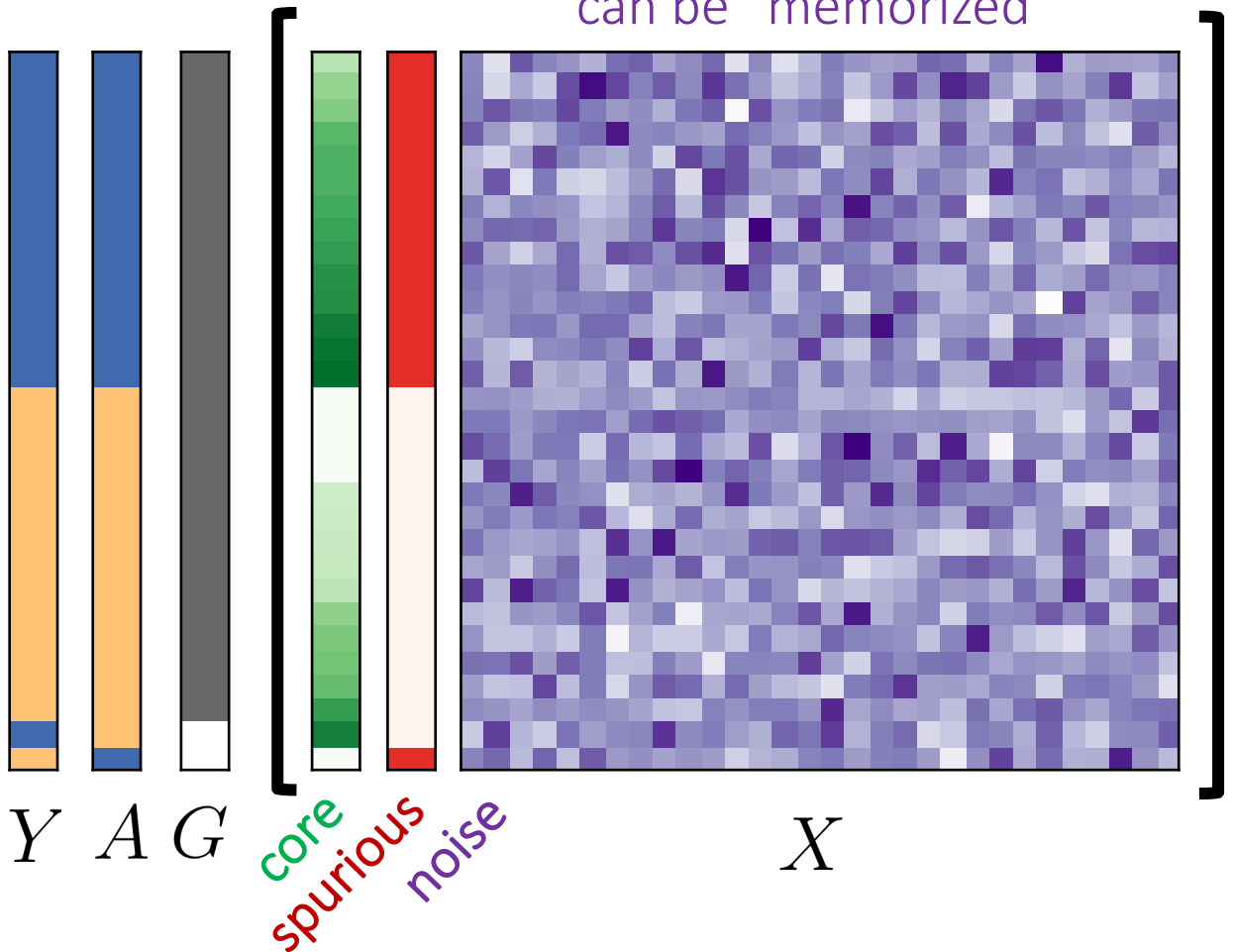
$$x_{\text{spu}} \in \mathbb{R}$$

$$x_{\text{spu}} \mid a \sim \mathcal{N}(a, \sigma^2_{\text{spu}})$$

$$x_{\text{noise}} \in \mathbb{R}^N$$

$$x_{\text{noise}} \sim \mathcal{N}\left(0, \frac{\sigma^2_{\text{noise}}}{N} I_N\right)$$

For large N>>n,
can be “memorized“

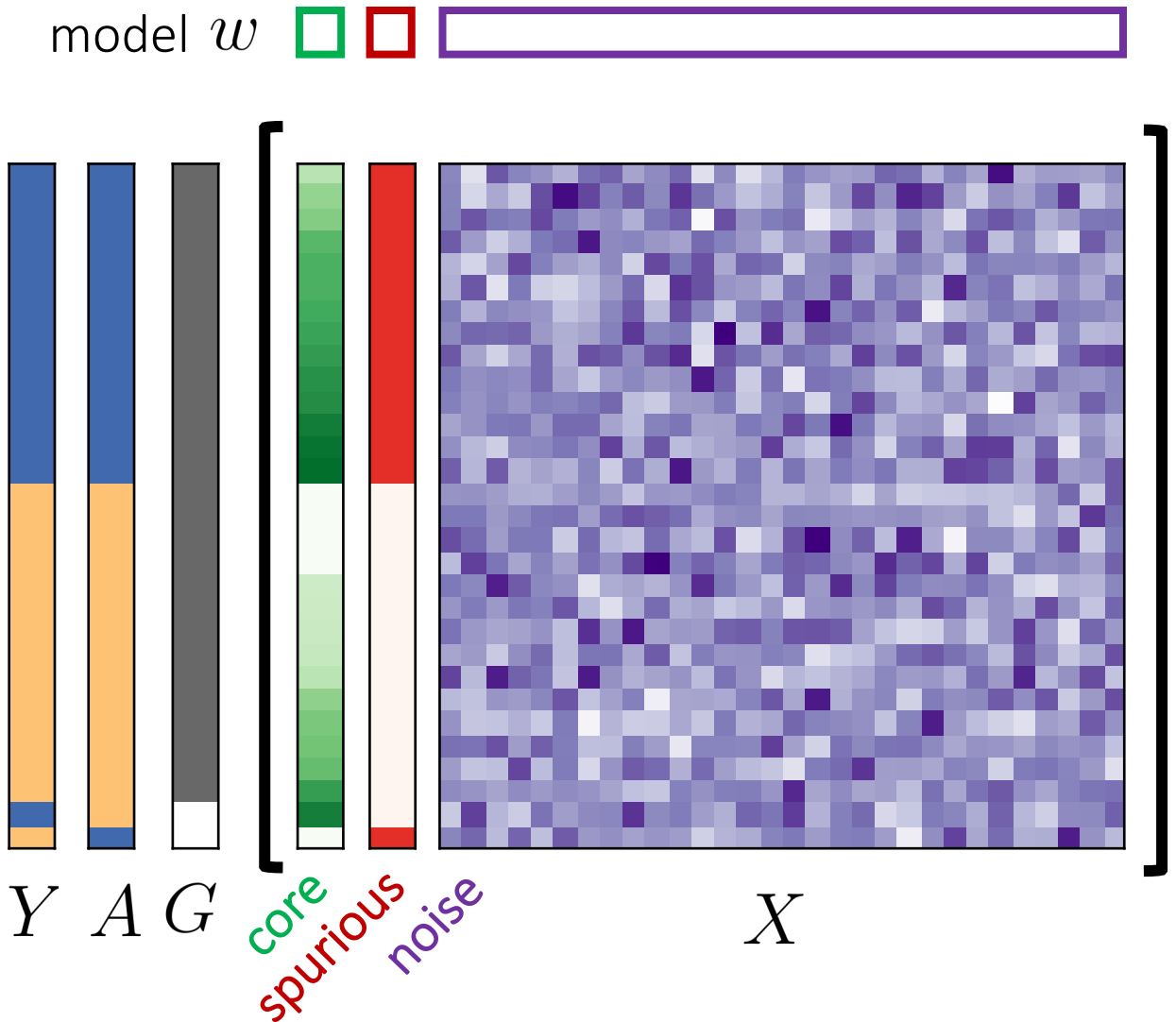


Toy example: linear classifier

- Logistic regression

- In overparameterized regime, equivalent to **max-margin classifier**

$$\hat{w}^{\text{mm}} = \arg\min \|w\|_2^2$$
$$\text{s.t. } y^{(i)} \left(w \cdot x^{(i)}\right) \geq 1 \; \forall i$$



Worst-group error is provably higher in the overparameterized regime

Theorem (informal). For any

High majority fraction

$$p_{\text{maj}} \geq \left(1 - \frac{1}{2001}\right)$$

$$\sigma^2_{\text{core}} \geq 1 \qquad \sigma^2_{\text{spu}} \leq \frac{1}{16 \log 100 n_{\text{maj}}},$$

High SCR

there exists N_0 such that for all $N > N_0$, with high probability,

$$Err_{\text{wg}}(\hat{w}^{\text{mm}}) \geq \frac{2}{3}$$

High worst-group error for overparameterized

However, with

$$p_{\text{maj}} = \left(1 - \frac{1}{2001}\right) \qquad \sigma^2_{\text{core}} = 1 \qquad \sigma^2_{\text{spu}} = 0$$

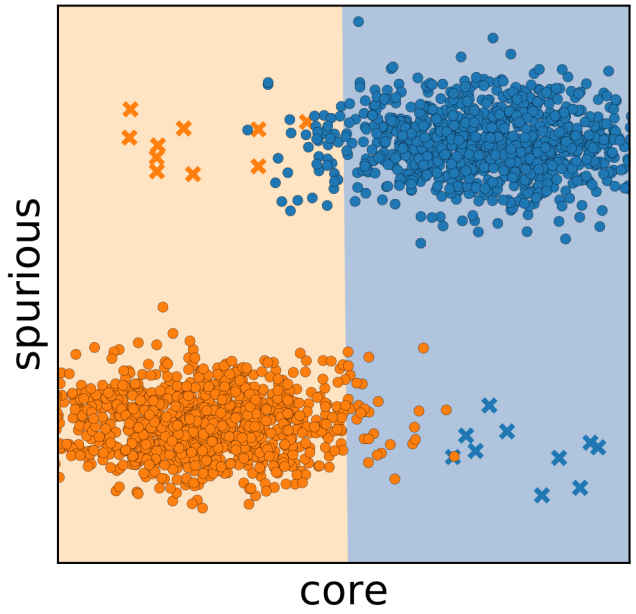
and $N = 0$ in the asymptotic regime with $n_{\text{maj}}, n_{\text{min}} \to \infty$,

$$Err_{\text{wg}}(\hat{w}^{\text{rw}}) \leq \frac{1}{4}$$

Low worst-group error for underparameterized

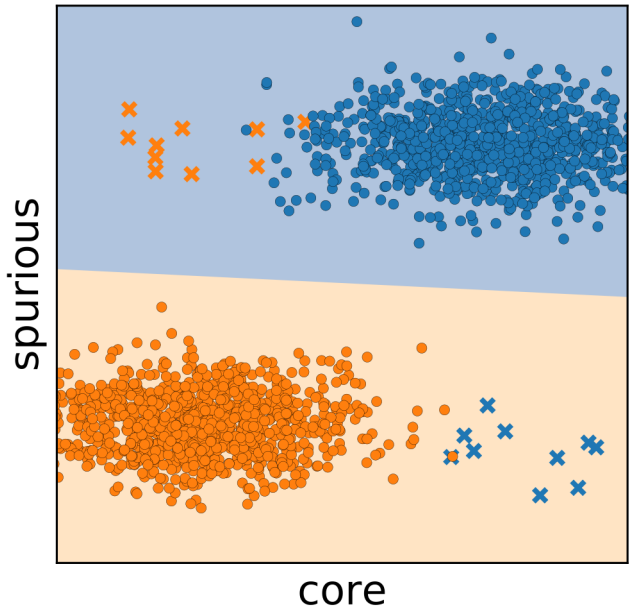
Underparameterized models need to learn the core feature to achieve low reweighted loss

learning core



✓ low reweighted loss

learning spurious

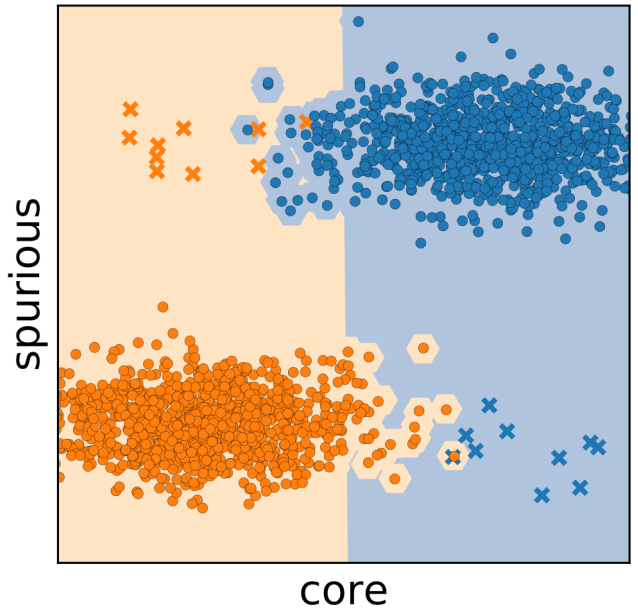


X high reweighted loss

In overparameterized regime, minimum-norm inductive bias favors less memorization

Norm scales with the number of points "memorized"

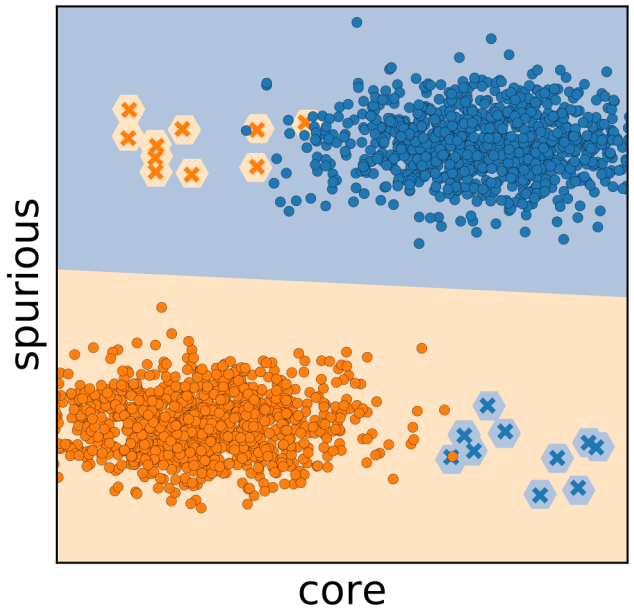
learning core
memorizing outliers



many examples memorized

X high norm

learning spurious
memorizing minority

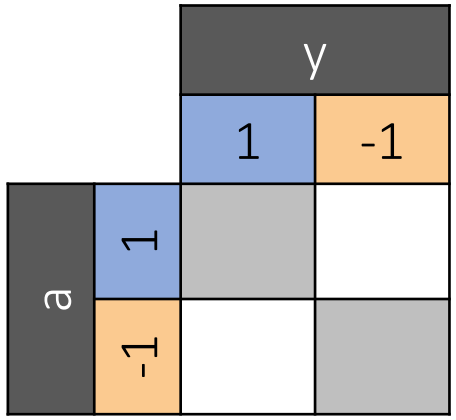


few examples memorized

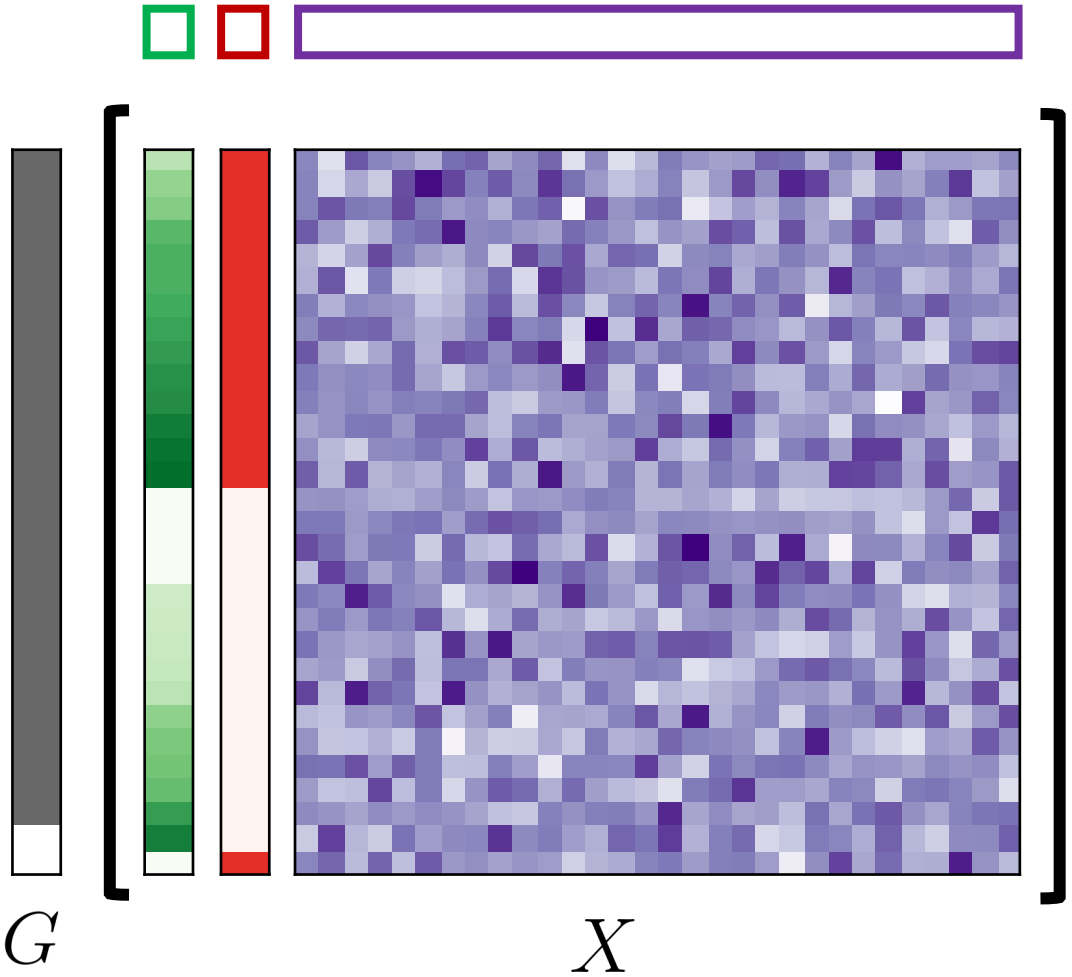
✓ low norm

Intuition: memorizing as few examples as possible under the min-norm inductive bias

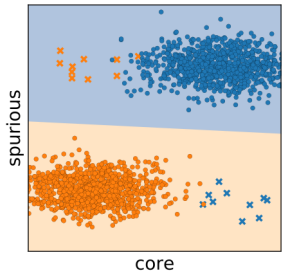
model $w = [w_{\text{core}}, w_{\text{spu}}, w_{\text{noise}}]$



Train error



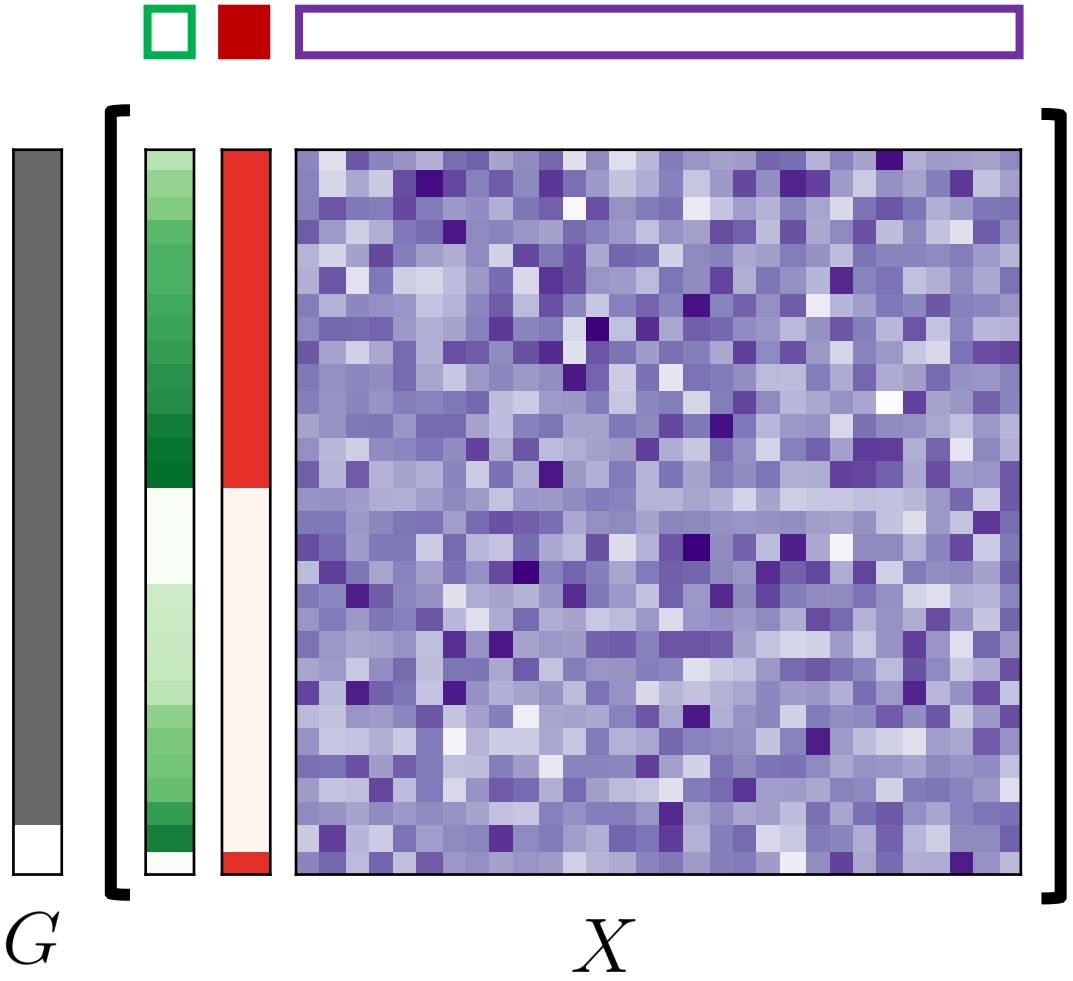
Learn spurious → memorize minority, low norm



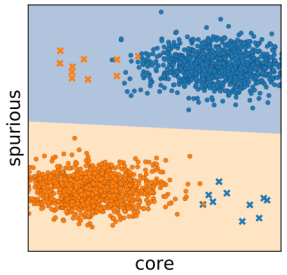
model $w = [w_{\text{core}}, w_{\text{spu}}, w_{\text{noise}}]$

a \ y	1	-1
1	0	1
-1	1	0

Train error



Learn spurious → memorize minority, low norm

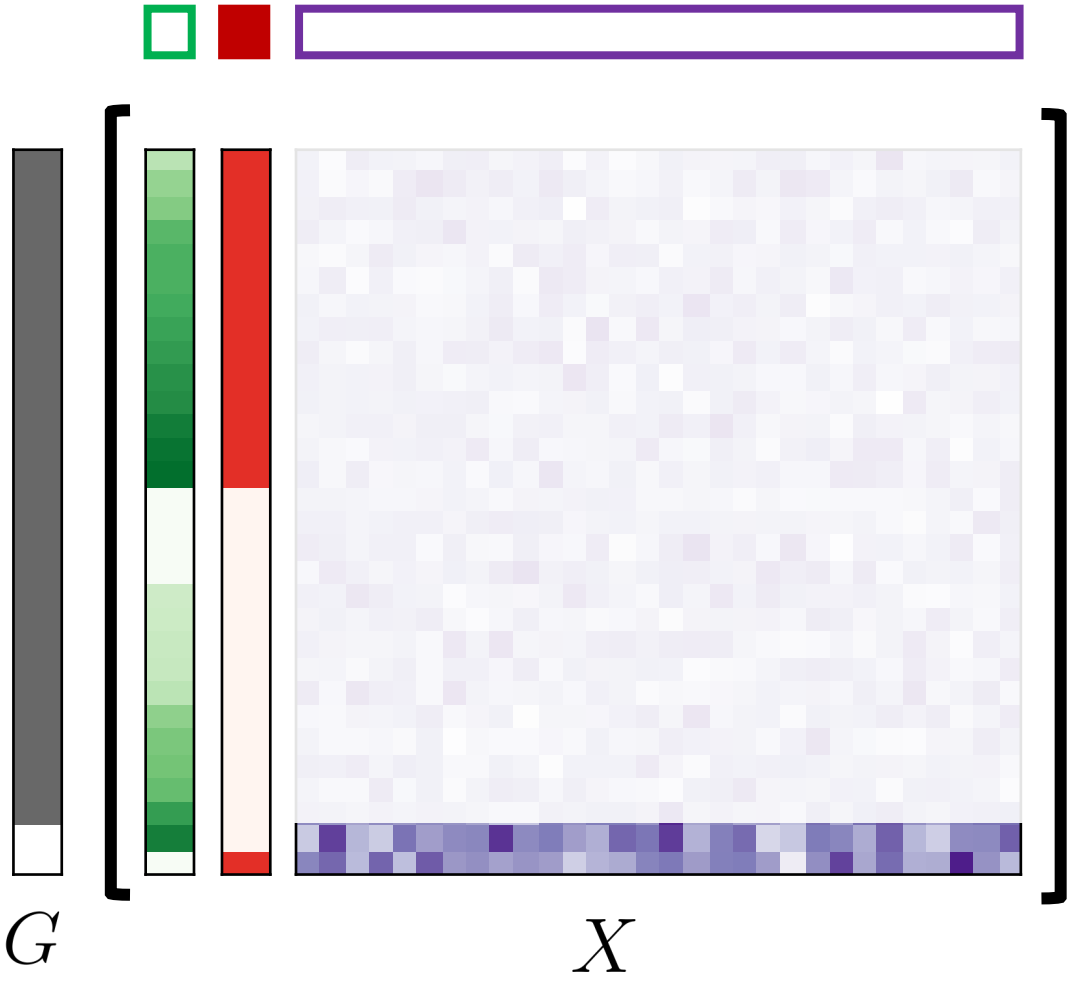


model $w = [w_{\text{core}}, w_{\text{spu}}, w_{\text{noise}}]$

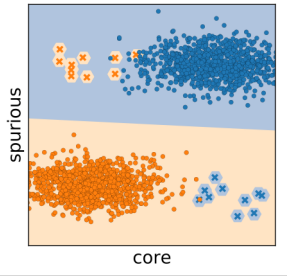
a \ y	1	-1
1	0	1
-1	1	0

Train error

$O(n_{\text{min}})$ points to memorize



Learn spurious → memorize minority, low norm

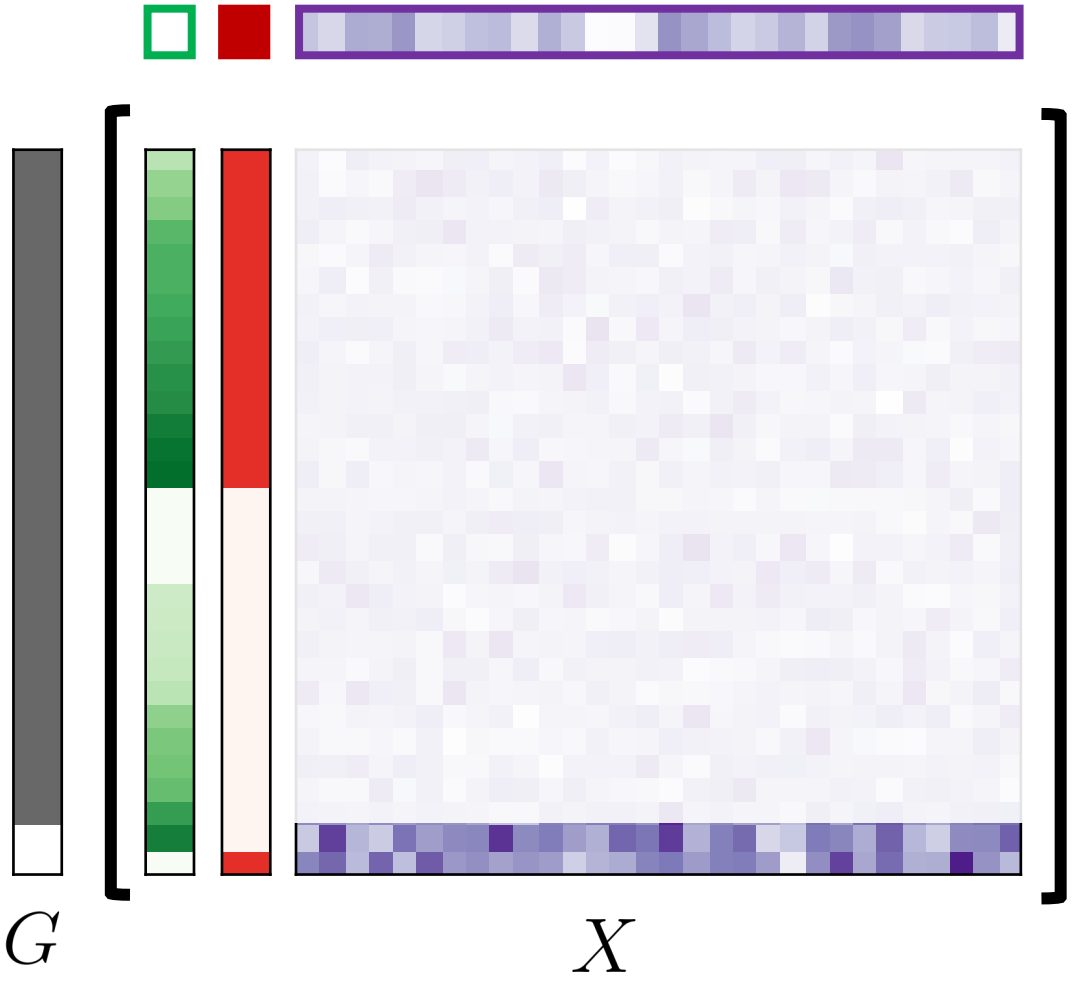


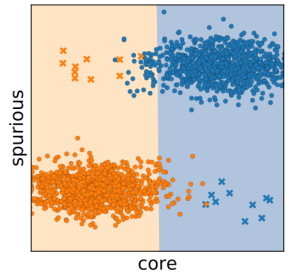
model $w = [w_{\text{core}}, w_{\text{spu}}, w_{\text{noise}}]$ ✓ low norm

a \ y	1	-1
1	0	0
-1	0	0

Train error

$O(n_{\text{min}})$ points to memorize



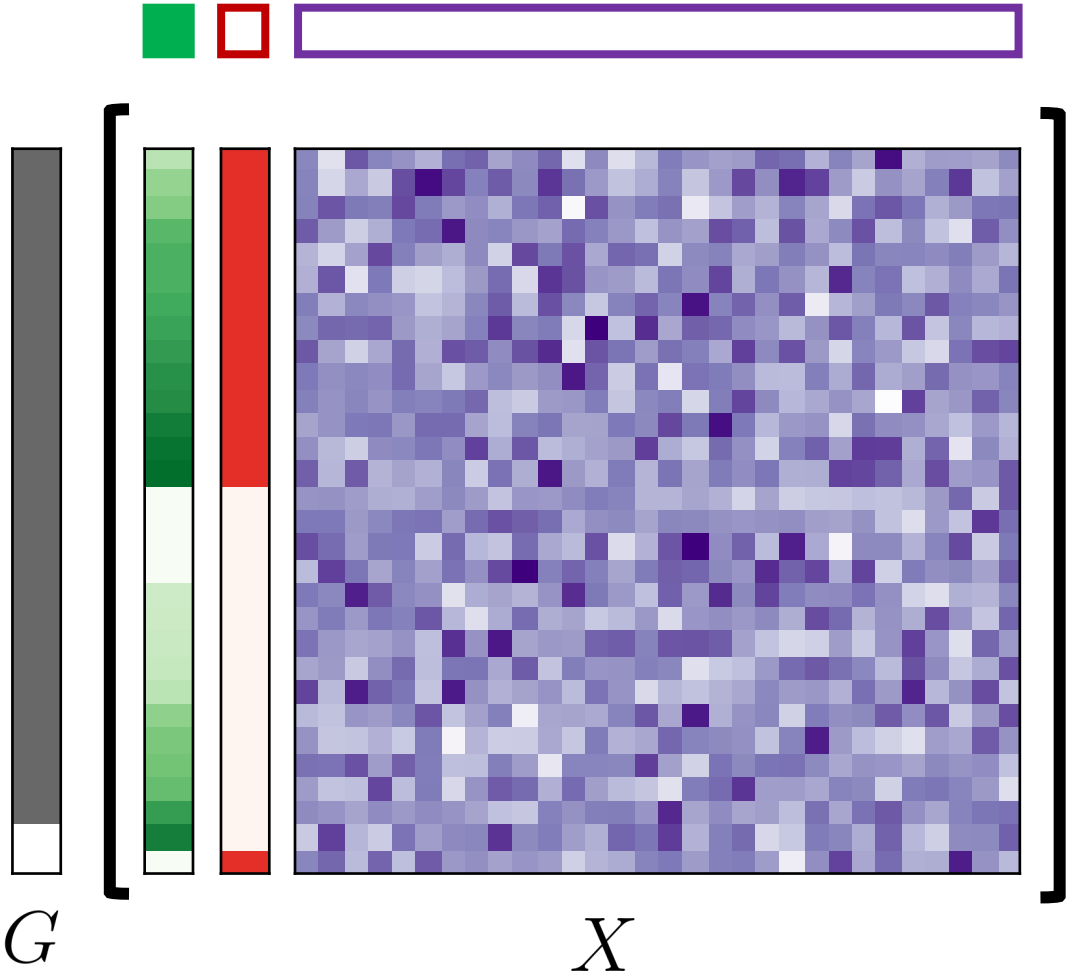


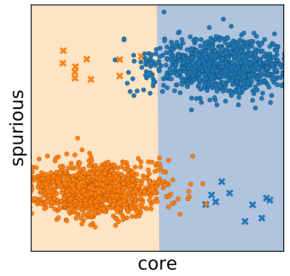
Learn core → memorize more, high norm

model $w = [w_{\text{core}}, w_{\text{spu}}, w_{\text{noise}}]$

a \ y	1	-1
1	>0	>0
-1	>0	>0

Train error





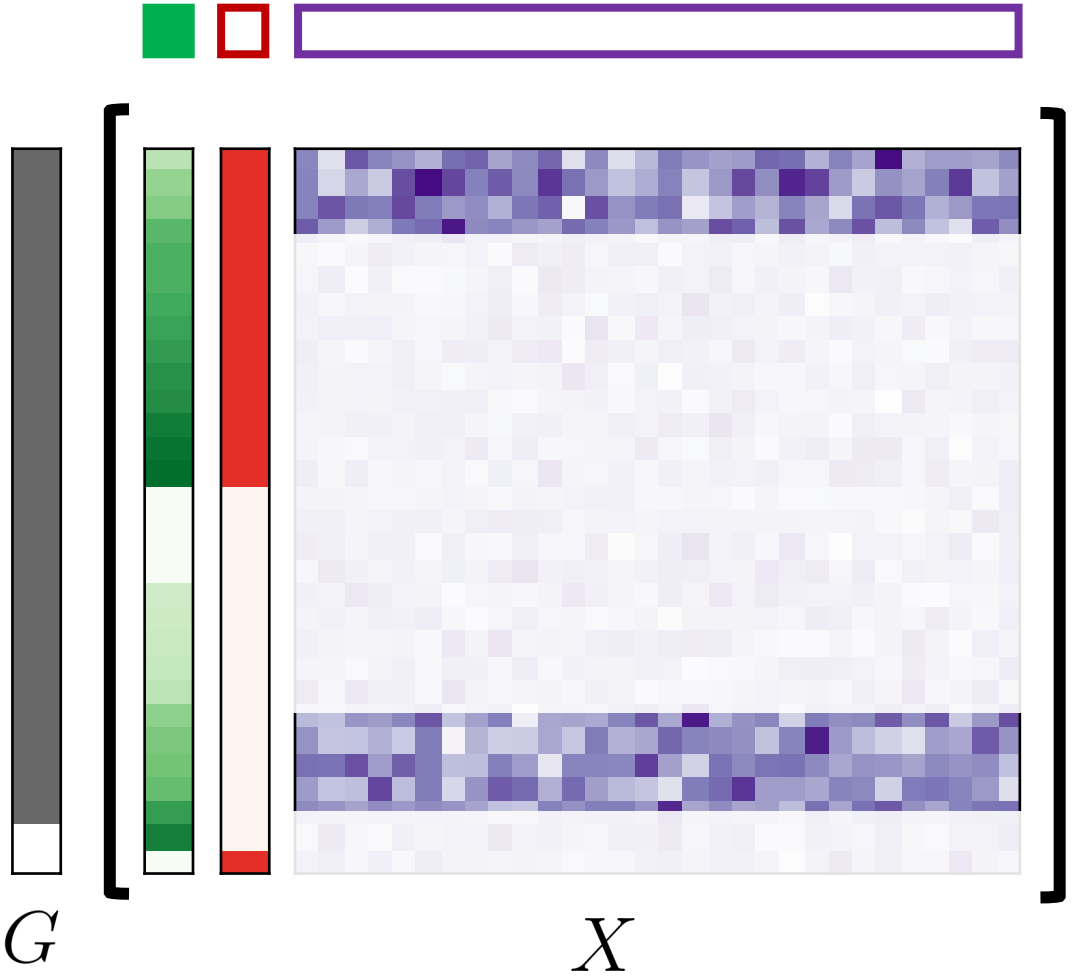
Learn core → memorize more, high norm

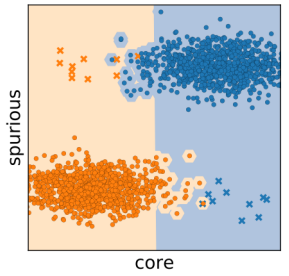
model $w = [w_{\text{core}}, w_{\text{spu}}, w_{\text{noise}}]$

a \ y	1	-1
1	>0	>0
-1	>0	>0

Train error

$O(n)$ points
to memorize





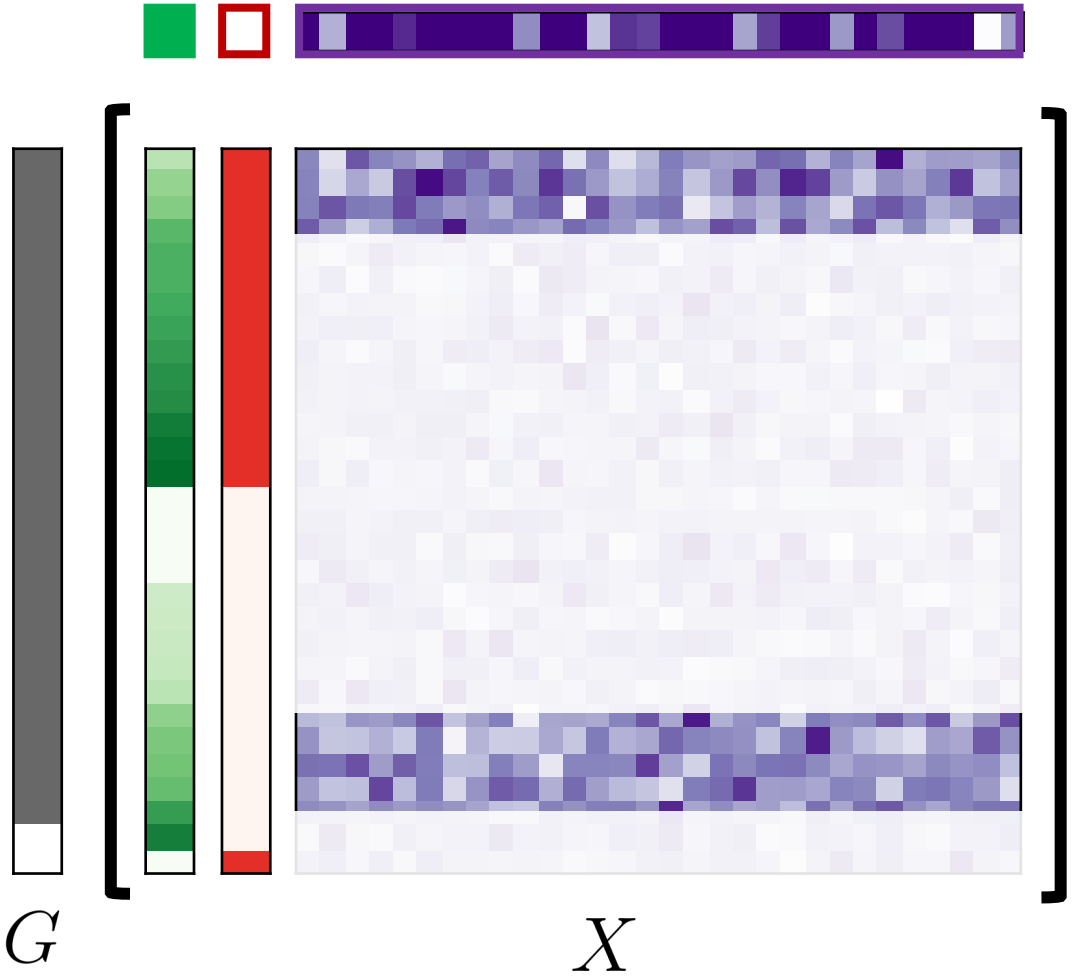
Learn core → memorize more, high norm

model $w = [w_{\text{core}}, w_{\text{spu}}, w_{\text{noise}}]$ X high norm

a \ y	1	-1
1	0	0
-1	0	0

Train error

$O(n)$ points
to memorize

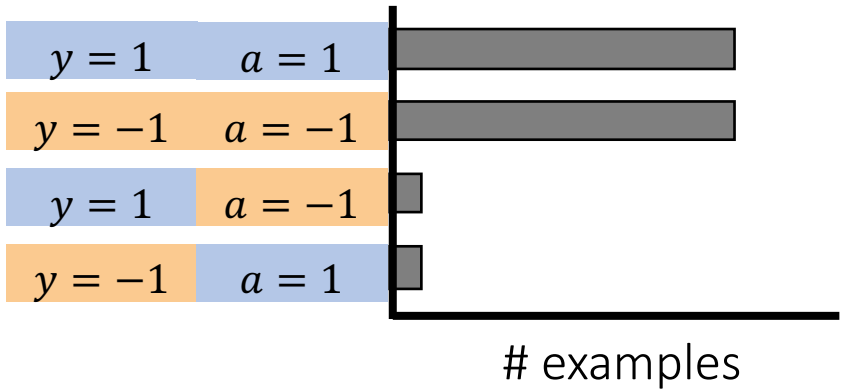


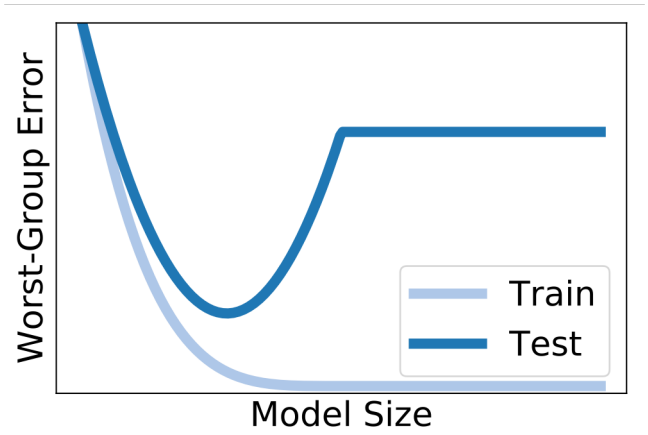
Overview

1. Empirical results
2. Simulations on synthetic data
3. Subsampling

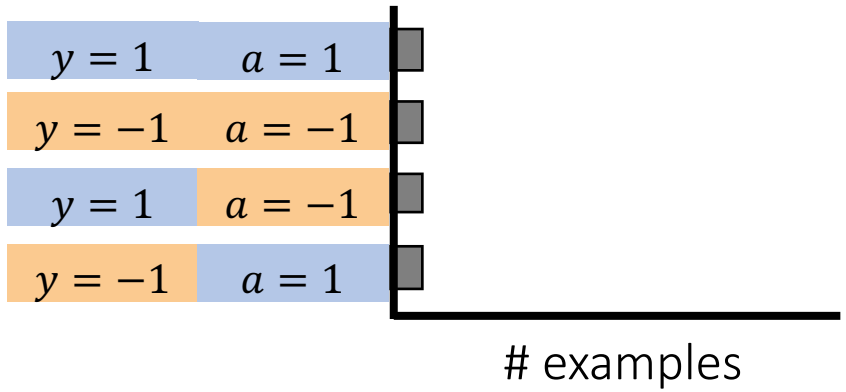
Reweighting vs subsampling

upweighting





subsampling

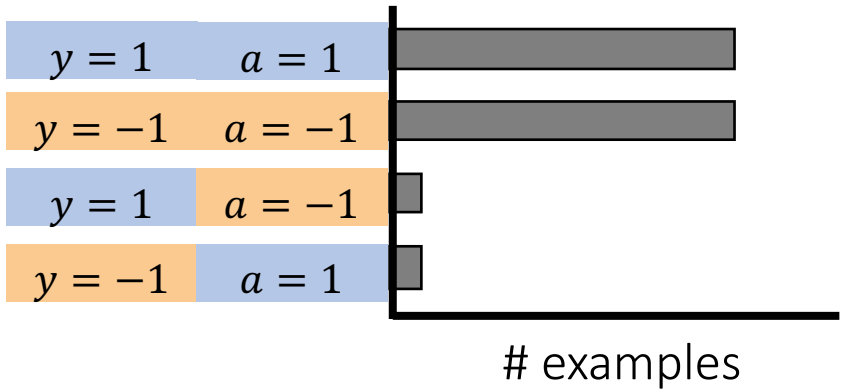


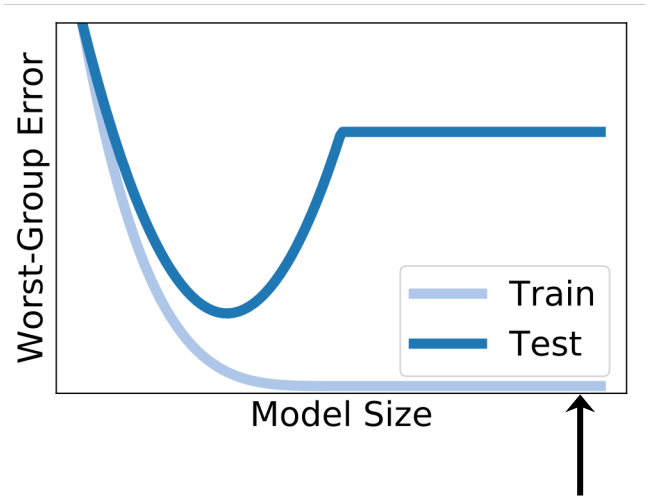
- Reduces majority fraction
- Lowers memorization cost of learning the core feature

Chawla et al. (2011)

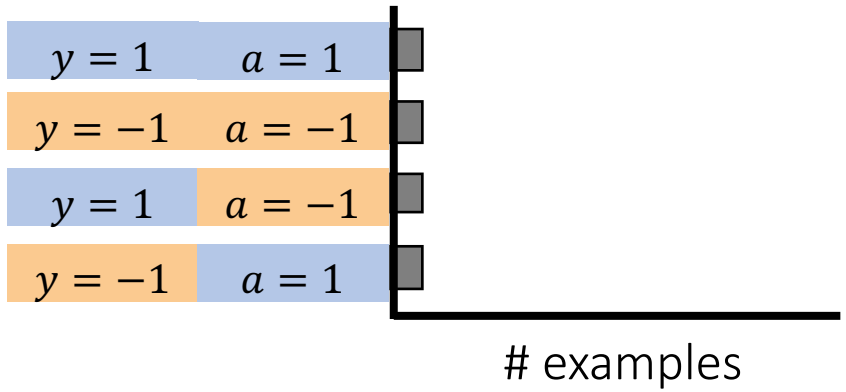
Reweighting vs subsampling

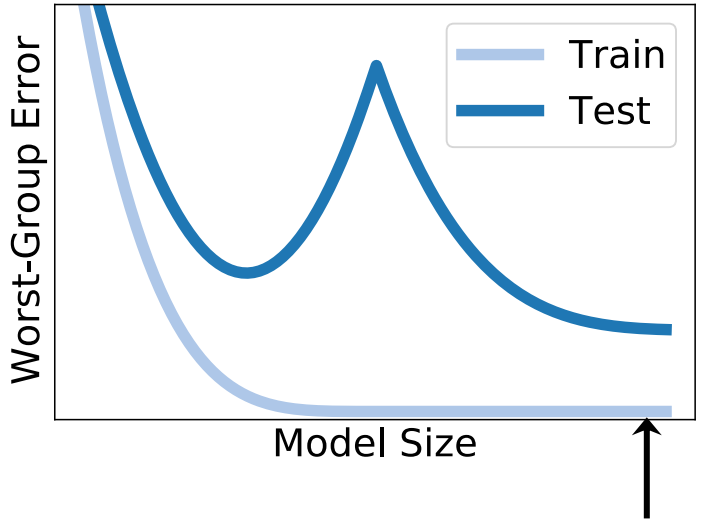
upweighting





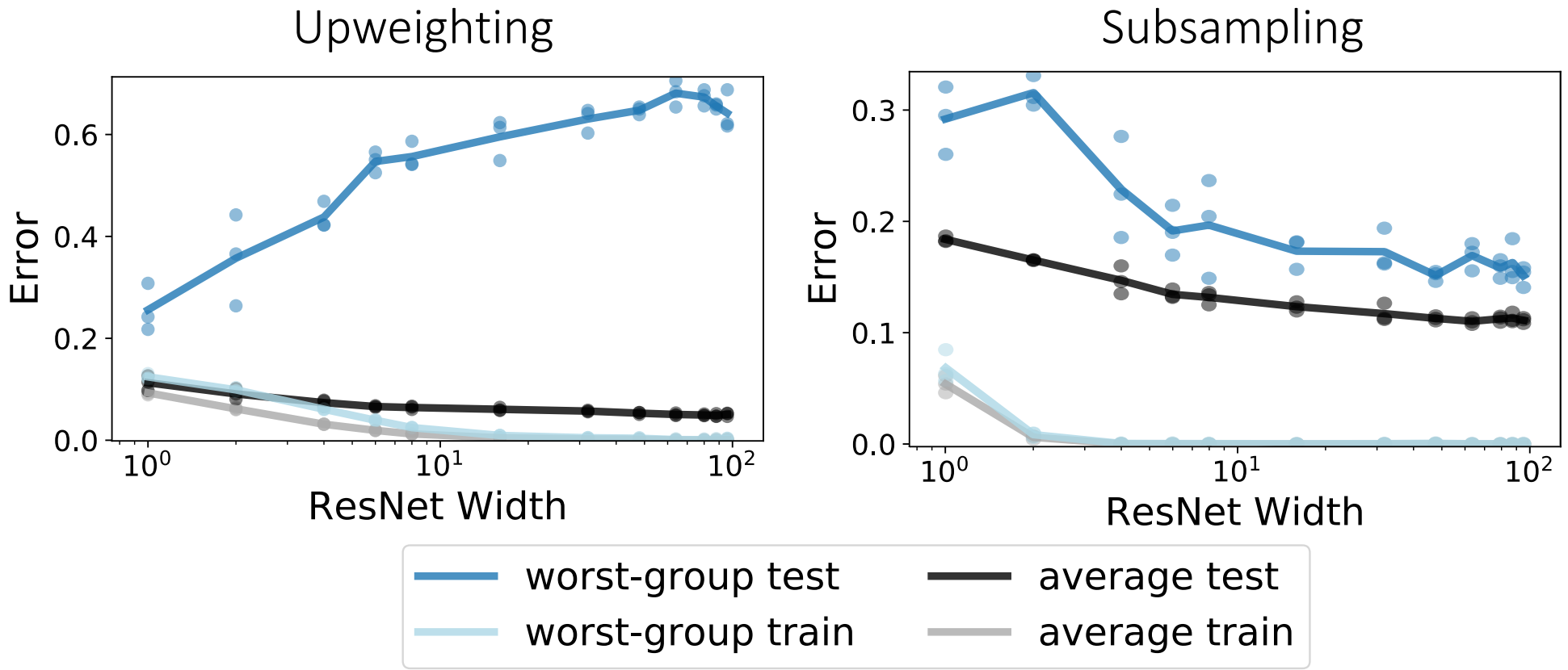
subsampling





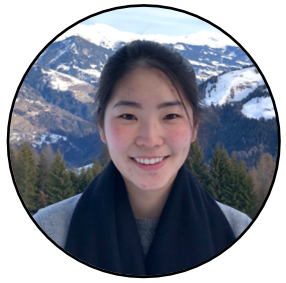
Chawla et al. (2011)

Subsampling the majority group → overparameterization helps worst-group error



Potential tension between using all of the data vs. using large overparameterized models.
Both help average error, but can't have both for good worst-group error.

Thanks!



Shiori Sagawa*

Aditi Raghunathan*

Pang Wei Koh*

Percy Liang

Thank you to Yair Carmon, John Duchi, Tatsunori Hashimoto, Ananya Kumar, Yiping Lu, Tengyu Ma, and Jacob Steinhardt

Funded by Open Philanthropy Project Award, Stanford Graduate Fellowship, Google PhD Fellowship, Open Philanthropy Project AI Fellowship, and Facebook Fellowship Program.