An Investigation of Why Overparameterization Exacerbates Spurious Correlation

Authors: Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, Percy Liang

Presented by-
Ashish Singh,Yang Guo

Overview

1) What causes bias in Machine Learning?
2) Understanding spurious correlations with examples.
3) Background: Why the need for Overparameterization?
4) Problem Statement.
5) Empirical results from the experiment.
6) Analytical model and theoretical results.
7) Proposal of subsampling to mitigate the problem.
8) References

What causes bias in Machine Learning?

Skewed sample

Tainted examples

Sample size disparity

Proxies

Limited features

Suggested Reference:

NIPS 2017 Fairness in Machine Learning by Solon Barocas, Moritz Hardt

https://nips.cc/Conferences/2017/Schedule?showEvent=8734

B, Selbst (2016)

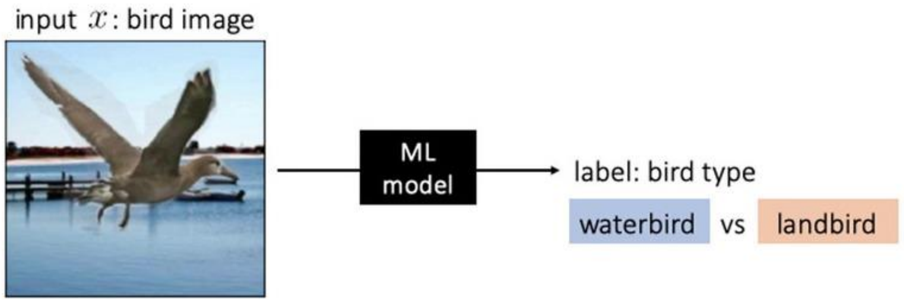
What causes bias in Machine Learning?

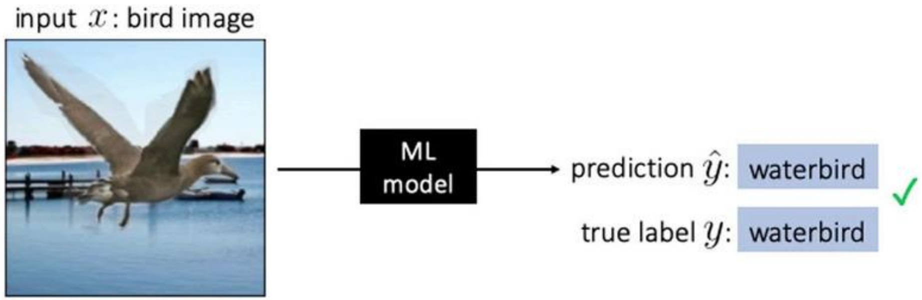
Spurious Correlations

misleading heuristics which might work on the majority group but doesn’t always holds true

Example: Spurious Correlations

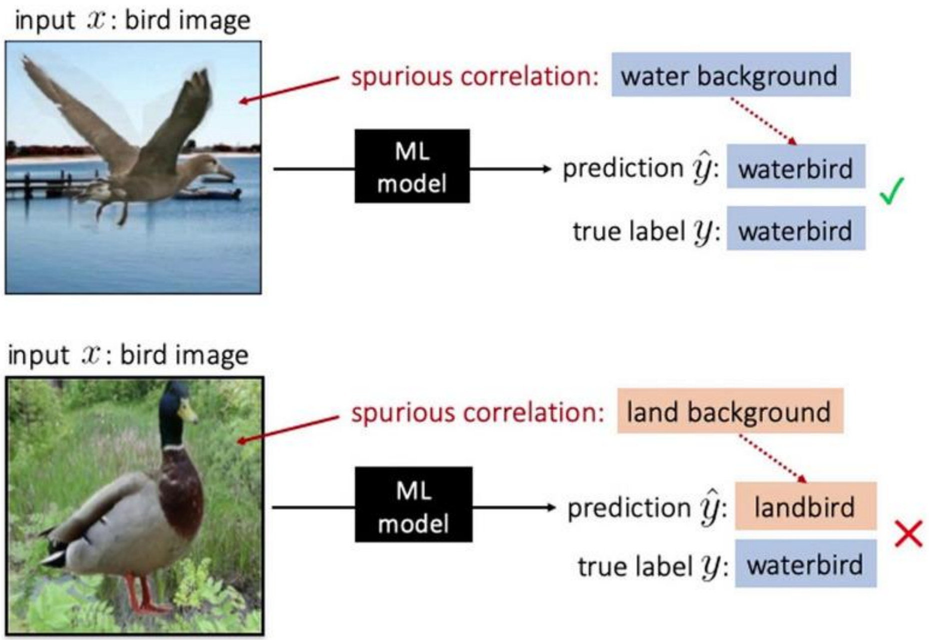
Here is an example considered in the paper (Waterbirds dataset).





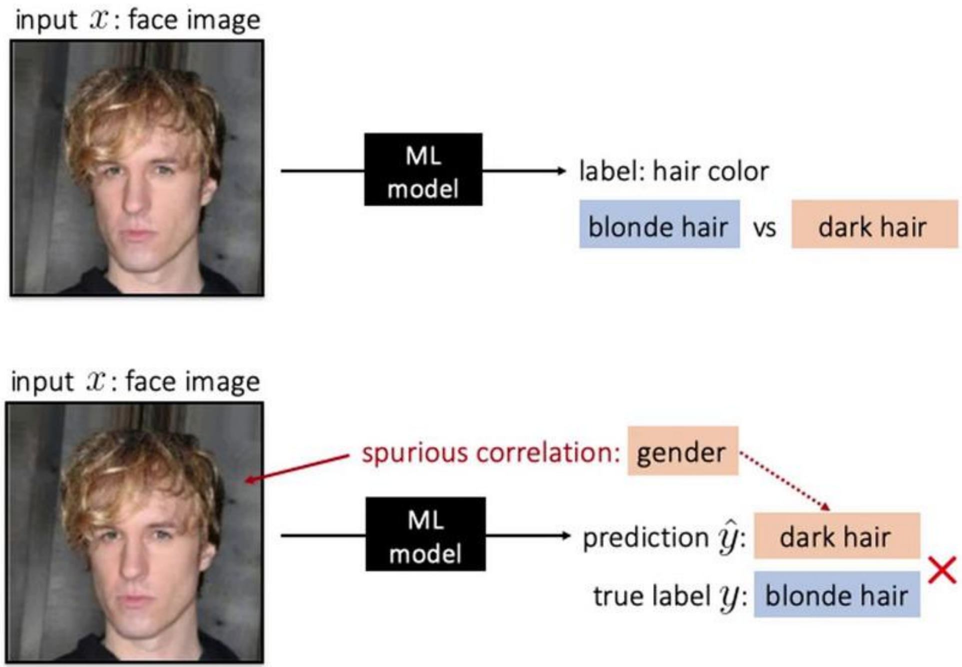
Example: Spurious Correlations

Here is an example considered in the paper (Waterbirds dataset).



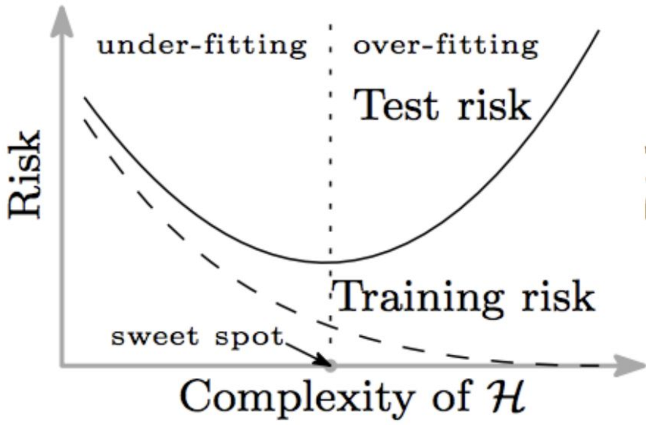
Example: Spurious Correlations

Here's another example considered in the paper (CelebA dataset).

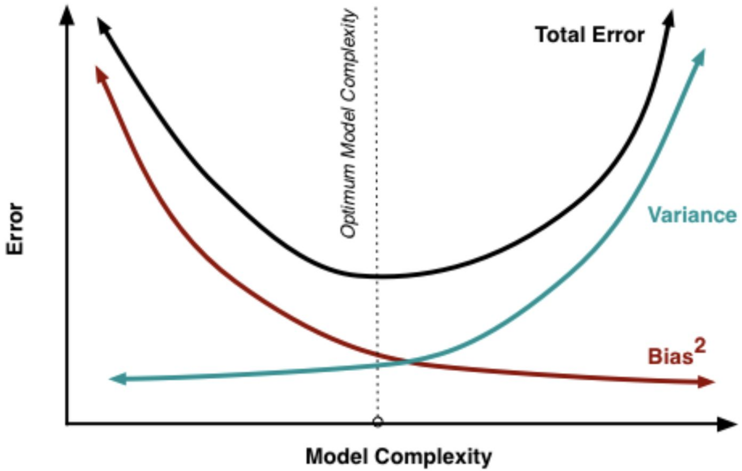


Background: Why the need for Overparameterization?

[Traditional wisdom]: Bias Variance Tradeoff w.r.t. Model complexity



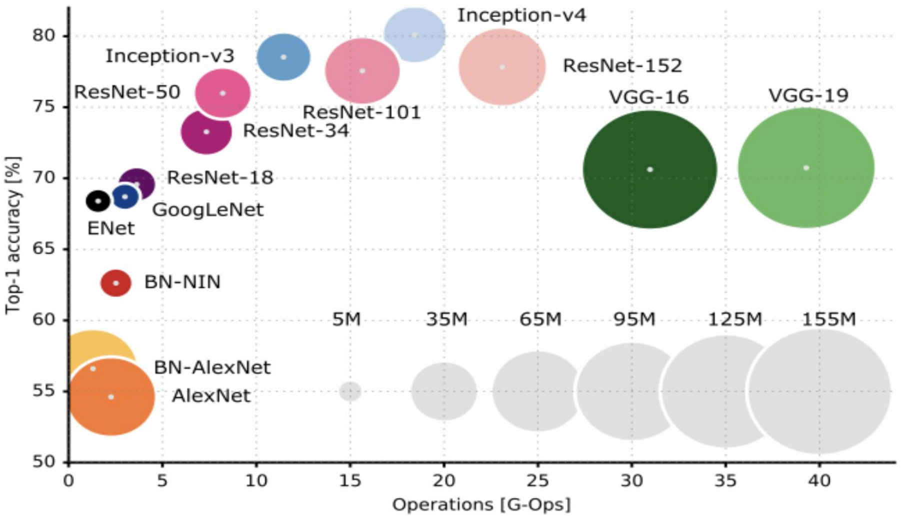
U-shaped “bias-variance” risk curve

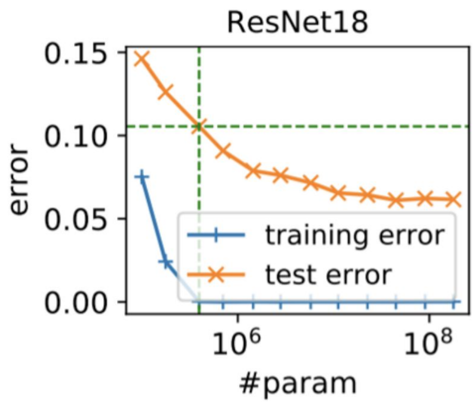


Belkin et al. 2018

Background: Why the need for Overparameterization?

Overparameterized model: # Parameters > # Data points

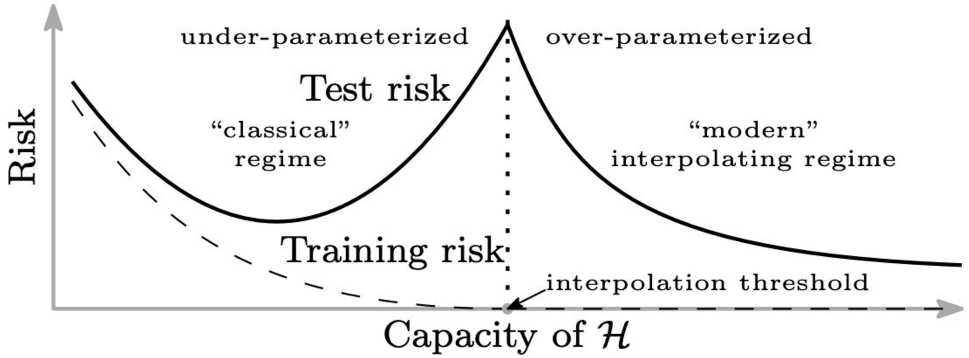




Neyshabur et al. 2018

Background: Why the need for Overparameterization?

After a certain threshold, the model becomes implicitly regularized by running SGD since the model tries to interpolate between points as smoothly as possible during the local search process.

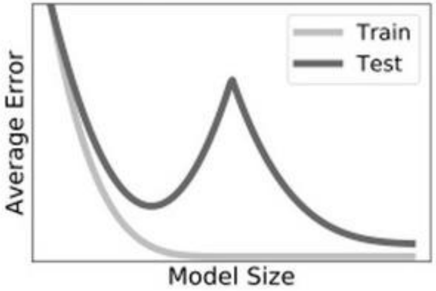


Inductive bias of SGD-type algorithm leads to the success of over-parameterized model like neural networks

Belkin et al. 2018

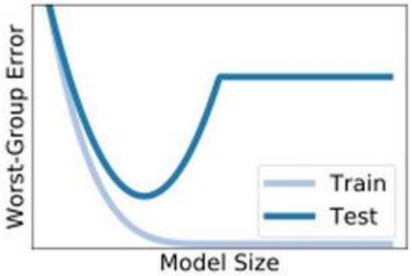
Overparameterization hurts worst group error when there are spurious correlations

Average Error



Why Overparameterization exacerbates worst-group error?

Worst-Group Error



Overparameterized is **better** than the underparameterized in **average error**

Overparameterized is **worse** than the underparameterized in **worst-group error**

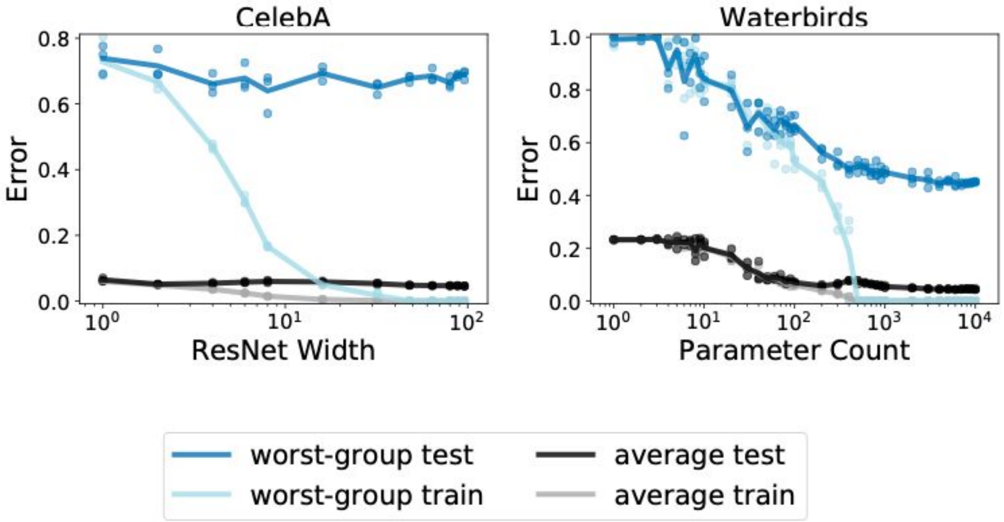
Empirical Setup: Models

Models used:

1) For CelebA dataset {hair color, gender}, ResNet10 model and model size is varied by increasing the network width from 1 to 96.
2) For Waterbirds dataset, logistic regression is used over random projections. The model size is varied by varying the number of the projections from 1 to 10000.

Empirical Setup: Verifying results from previous work

Training models via ERM have poor worst-group test error regardless of whether they are under- or overparameterized.



Empirical Setup: Reweighted Objective

New objective function:

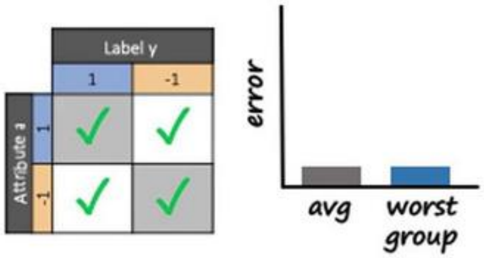
Upweighting the minority groups: $\mathcal{R}_{\text{reweight}}(w) = \hat{E}_{(x,y,g)}\left[\frac{1}{\hat{p_g}}\ell(w,(x,y))\right]$

Another approach: Group DRO but for simplicity upweighting is considered here.

Prior work shows approaches for improving worst-group error fail on high capacity models

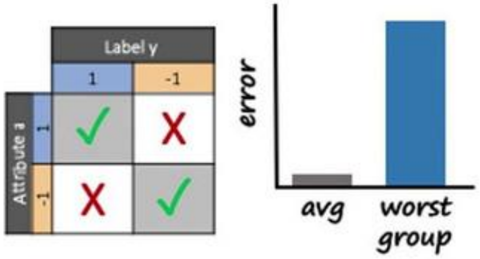
Upweighting the minority groups: $\mathcal{R}_{\text{reweight}}(w) = \hat{E}_{(x,y,g)}\left[\frac{1}{\hat{p}_g}\ell(w,(x,y))\right]$

Low-capacity Models



More robust to spurious correlations
Low worst-group error

High-capacity Models

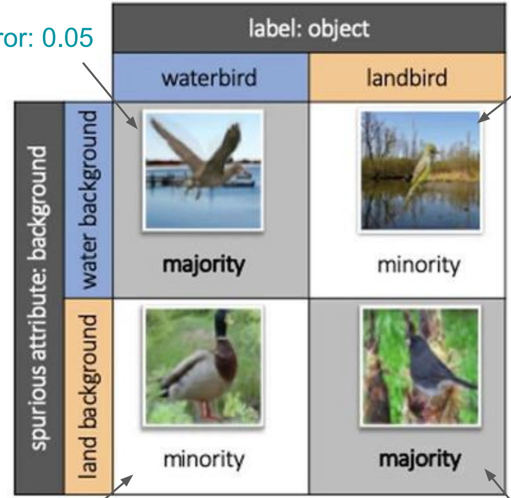


Relies on spurious correlations
High worst-group error

Empirical Results: Overparameterization exacerbates worst-group even when trained with reweighted objective

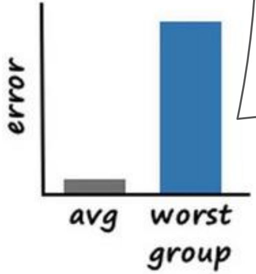
error: 0.05

error: 0.21



error: 0.40

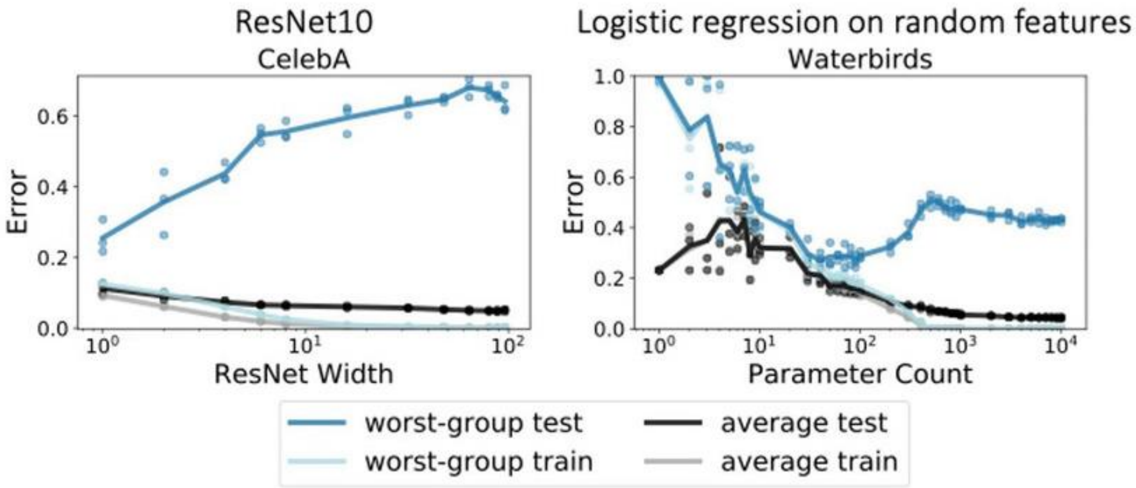
error: 0.004



Model performs well on average
but can have high worst group error

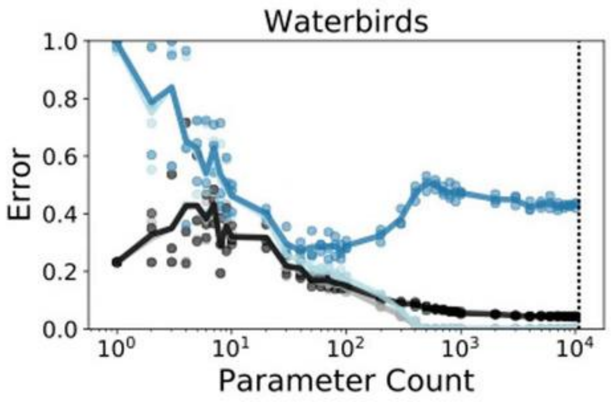
average error: 0.03
worst-group error: 0.40

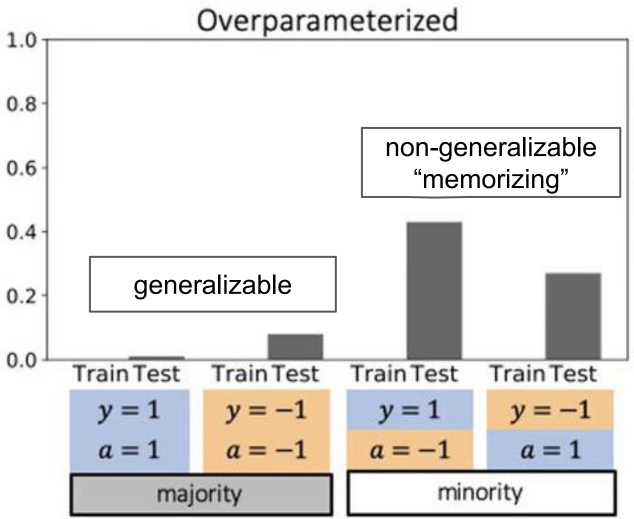
Empirical Results: Overparameterization exacerbates worst-group even when trained with reweighted objective



(when trained to minimize average loss, observing worst-group error across model sizes)

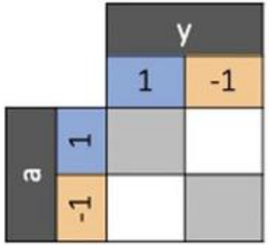
Hypothesis: Overparameterized models learn the spurious attribute and memorize minority groups





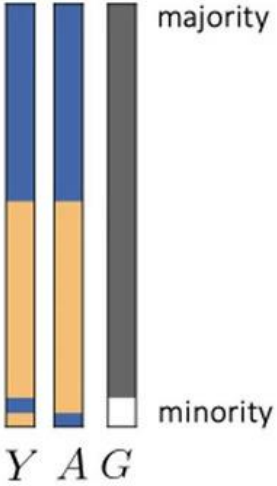
Overparameterized models learn the spurious features and memorize the minority

Analytical Model and Theoretical Results: Toy example data



Majority fraction

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



Analytical Model and Theoretical Results: Toy example data

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

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

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

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

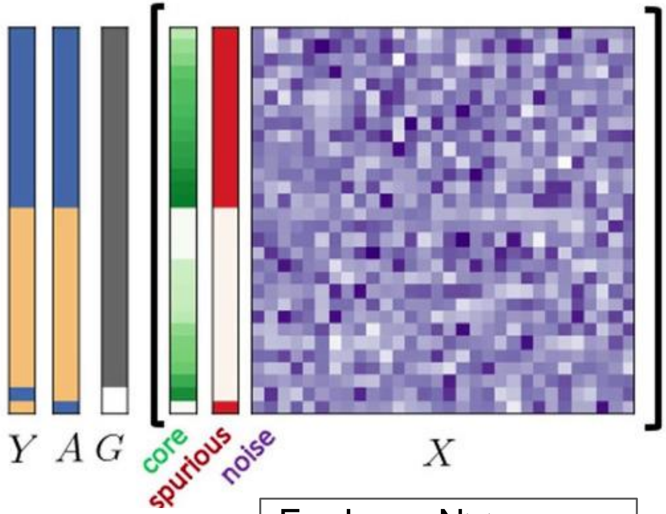
$$x_{\text{spu}} \mid a \sim \mathcal{N}(a, \sigma^2_{\text{spu}})$$

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

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

SCR:

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



For large N>>n,
can be “memorized”

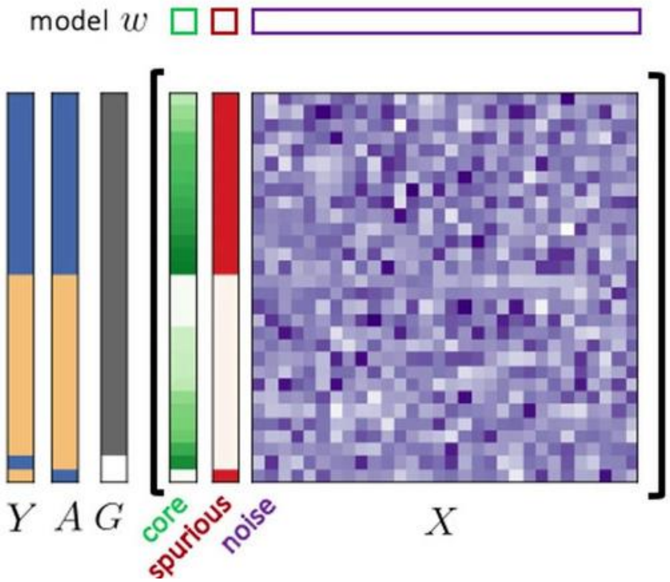
Analytical Model and Theoretical Results: Linear Classifier

Linear Classifier

$\hat{w}^{\text{rw}}$ minimizes reweighted logistic loss.

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 probably higher in the overparameterized regime

Theorem (informal). For any

High majority fraction $p_{\text{maj}} \geq \left(1 - \frac{1}{2001}\right)$ $\sigma^2_{\text{core}} \geq 1$ $\sigma^2_{\text{spu}} \leq \frac{1}{16 \log 100 n_{\text{maj}}}$, 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

Notations

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

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

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

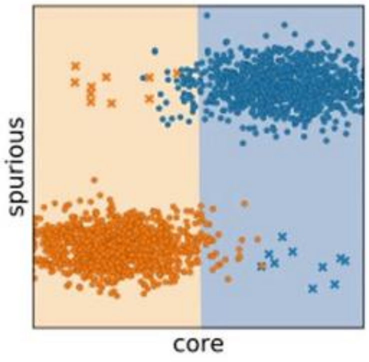
$x_{\text{spu}} \in \mathbb{R}$

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

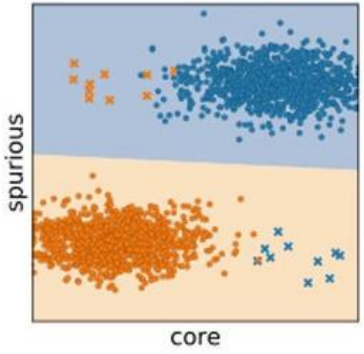
$x_{\text{noise}} \in \mathbb{R}^N$

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

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



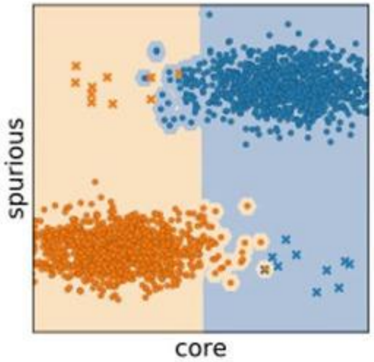
learning core features
low reweighted loss



learning spurious features
high reweighted loss

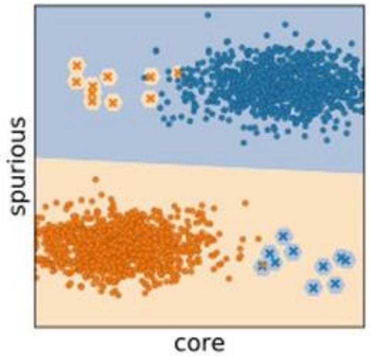
Sagawa et al. 2020

Hypothesis: Overparameterized models learn the spurious attribute and memorize minority groups



learning core features
memorizing outliers

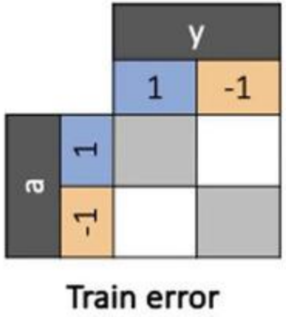
many examples to memorize

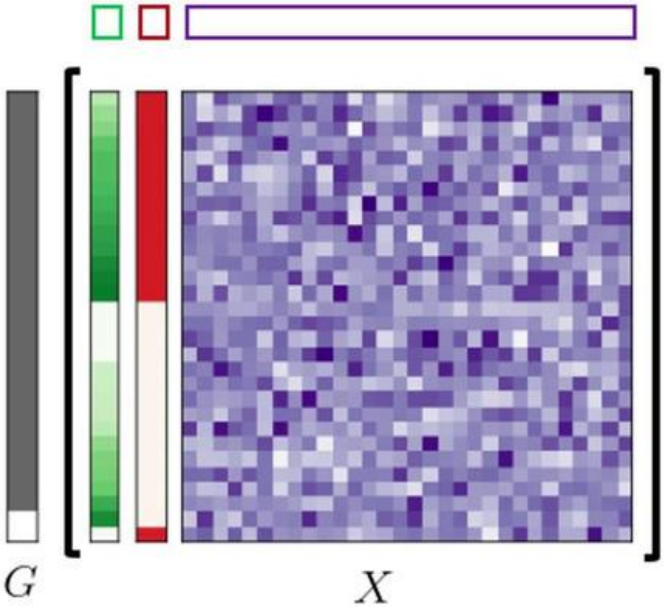


learning spurious features
memorizing minority

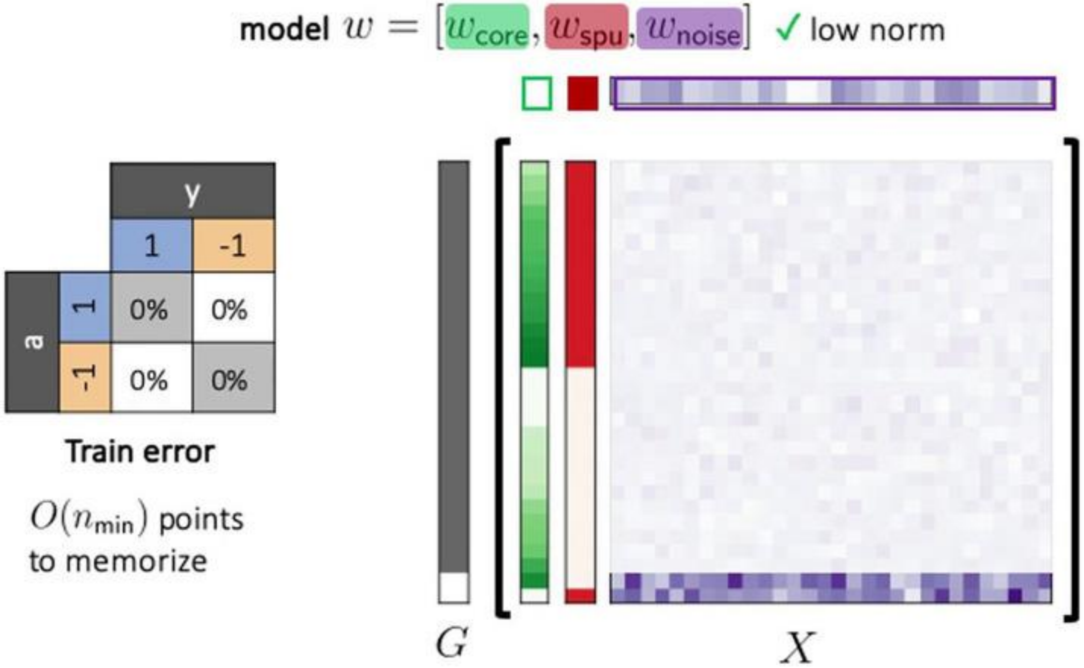
few examples to memorize

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

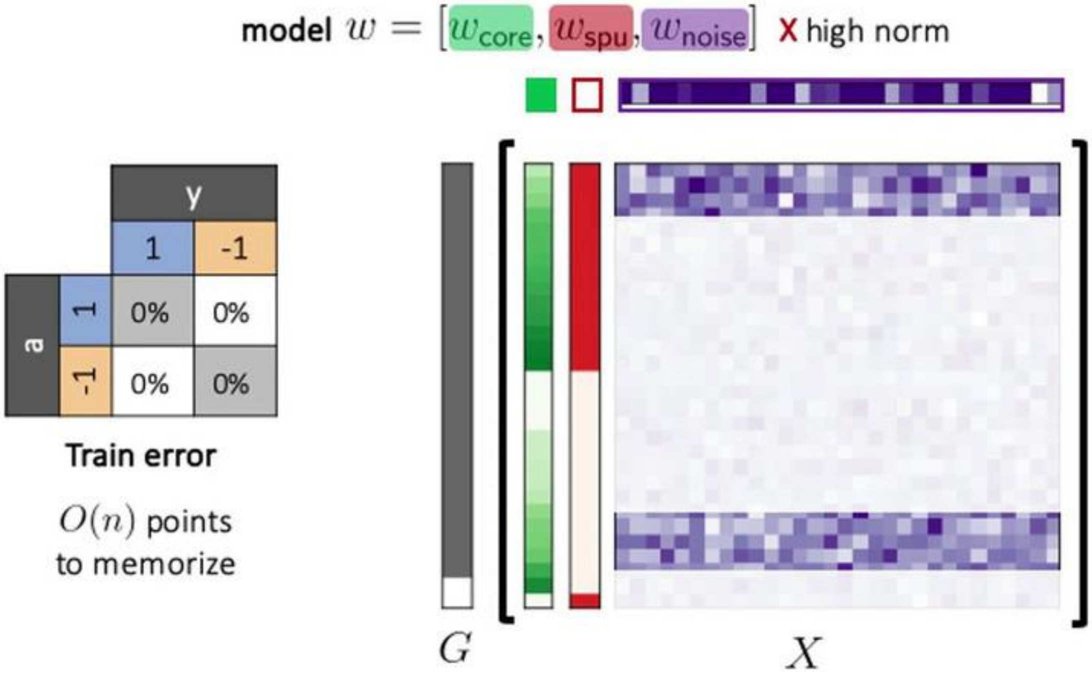




Learn spurious features - memorize minority, low norm

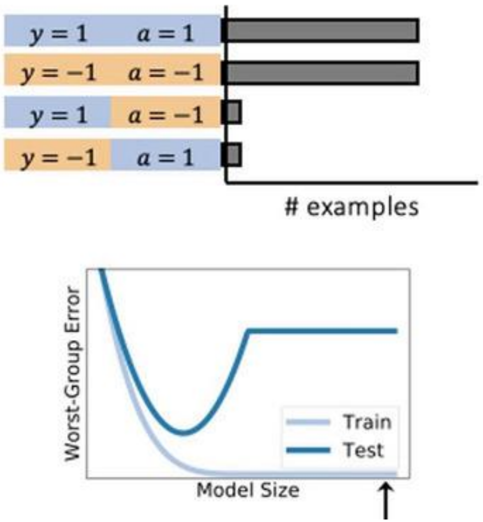


Learn core features - memorize more, high norm

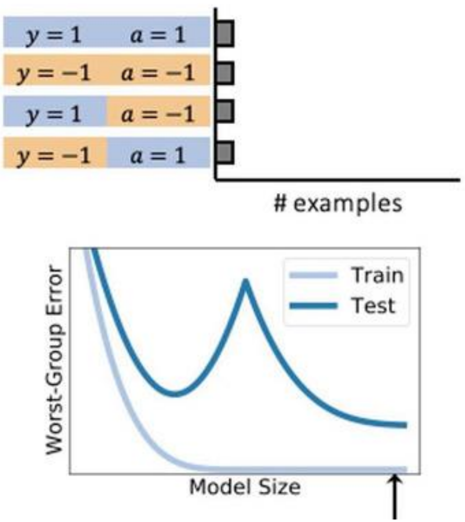


Proposed Subsampling: Reweighting vs Subsampling

Reweighting

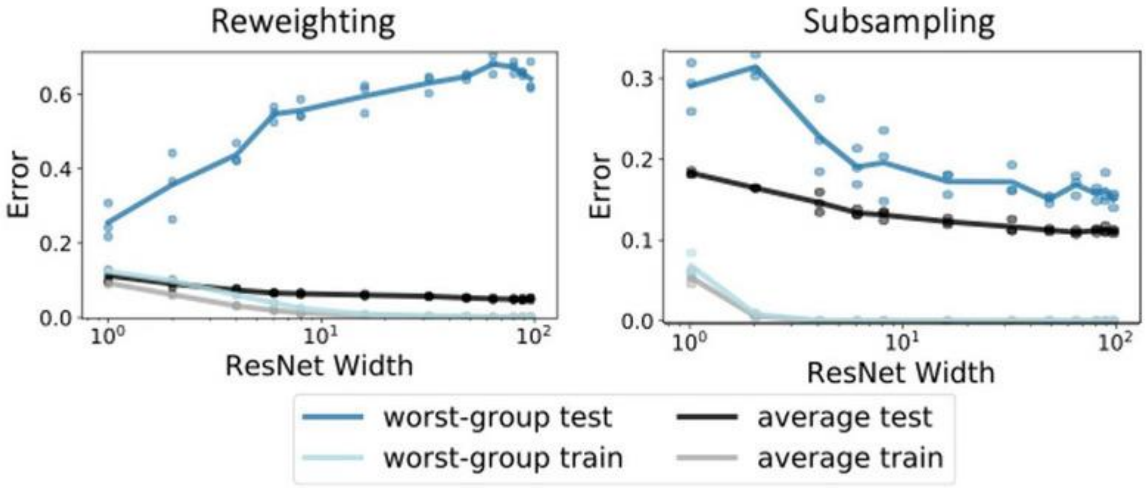


Subsampling



Reduces Majority fraction
Lowers the memorization cost of learning the core features

Proposed Subsampling: Overparameterization helps worst-group error after subsampling



This results in a conflict of whether to use all of the data vs large overparameterized models.
Both help average error, but together they are not good for worst-group error.

References

1. Reconciling modern machine learning practice and the bias-variance trade-off [Belkin et al. 2018]
2. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization [Sagawa et al. 2020]
3. An investigation of why overparameterization exacerbates spurious correlations [Sagawa et al. 2020]
4. Towards Understanding the Role of Over-Parameterization in Generalization of Neural Networks [Neyshabur 2018]

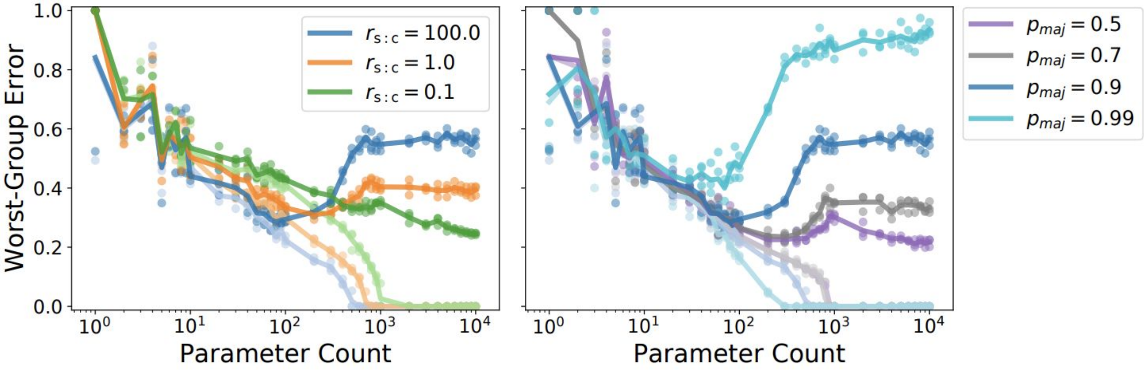
Thanks!

Quiz Questions

1. Which of the following properties for the training data will make overparameterization hurt the worst-group error?
   A. Higher majority fraction
   B. Lower majority fraction
   C. Higher spurious-core information ratio
   D. Lower spurious-core information ratio

A, C

Reason:

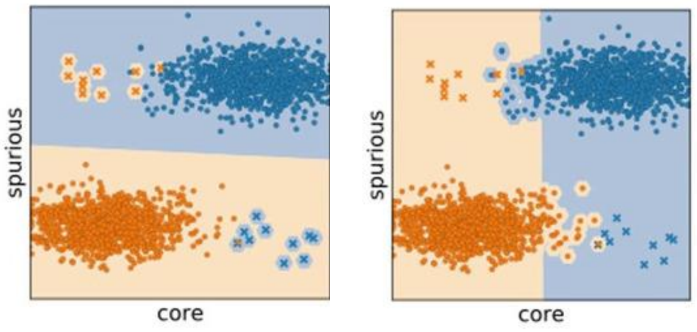


Quiz Questions

2. What is the reason that subsampling outperforms reweighting under the overparameterized regime?
    A. Lower the memorization cost of the core feature by reducing the majority fraction
    B. Lower the memorization cost of the core feature by increasing the majority fraction
    C. Lower the memorization cost of the spurious feature by reducing the majority fraction
    D. Lower the memorization cost of the spurious feature by increasing the majority fraction

A

Reason: Because the overparameterized model is able to memorize the minority training data, if we assign higher weight for these points, the model will still have the exact same loss. In comparison, subsampling makes is less expensive to memorize the outliers.



Quiz Questions

3. Under the overparameterized setting, minimum norm inductive bias will favor which of the followings:

   A. Memorizing the outliers in the majority group
   B. Memorizing the training points in the minority group
   C. Memorizing the complete training set in the majority group
   D. Memorizing the training data by balancing the groups in the training data

B

Reason: The overparameterized model will prefer the memoring the training points in the minority group as it will have less number of points to be memorized.