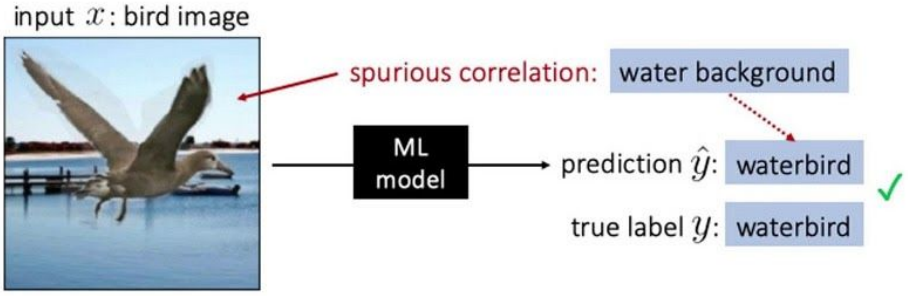
An investigation of why overparameterization exacerbates spurious correlation

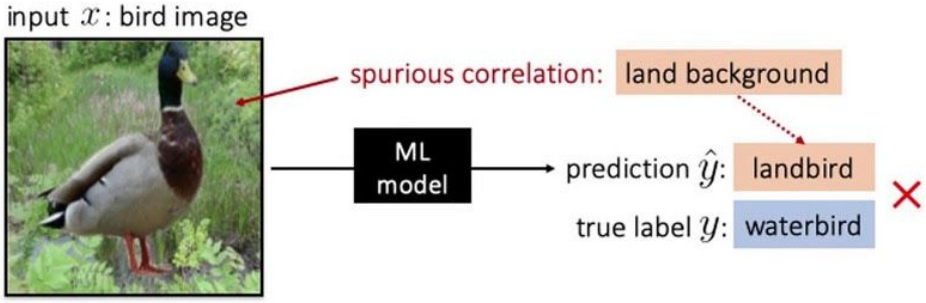
Yang Guo, Ashish Singh

I. Motivation for Spurious Correlation Mitigation

- **Spurious Correlation**: mathematical relationship in which two or more events or variables are associated but not causally related, due to either coincidence or the presence of a certain third, unseen factor, aka confounding factor.
    - In typical ML research, people usually restrict the discussion of spurious correlation to the scenario that correlated features lead to similar output. Based on the knowledge from the actual world, we decide which one is the **spurious feature.**
    - Classic waterbird example:



-



        - The ML model predicts the waterbird/landbird based on the background feature water/land, because high correlation between the background type and bird type
- Previous mitigation strategies:
    - **Upweighting** the minority samples (aka. **Importance weighting**): assign a weight to the minor samples. for example, the weight can be #majority samples/#minority samples
    - Group Distribution Robust Optimization (Group DRO)[Sagawa et al. 2020]:

- Main idea: Train the DRO objective and the group weight simultaneously while using a heavy l_2 regularization.
- This research actually motivates the development of the paper in the discussion by observing that Group DRO fails under the overparameterized setting.

II. Background: Bias in Machine Learning.

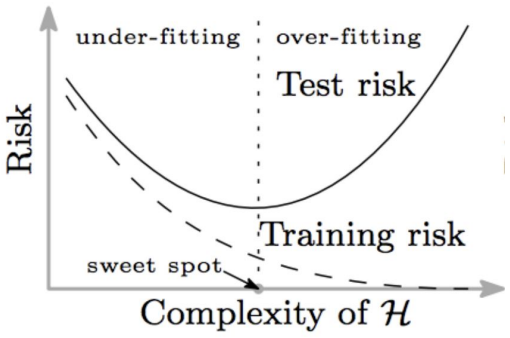
- Skewed Samples
  - People may collect more samples from group A, and less sample from group B, but the collected samples are not representative of the entire population
- Tainted examples
  - Data contamination
- Sample size disparity
  - For example, fewer examples in the minority can lead to a potentially worse predictive model than the one trained in the majority group
- Proxies
  - Features can be very **correlated**, and certain features can serve as a proxy for other features.
- Limited features
  - Features may be less informative or less reliably collected for certain parts of the population
  - For example, a feature set that supports accurate predictions for the majority group may not for a minority group

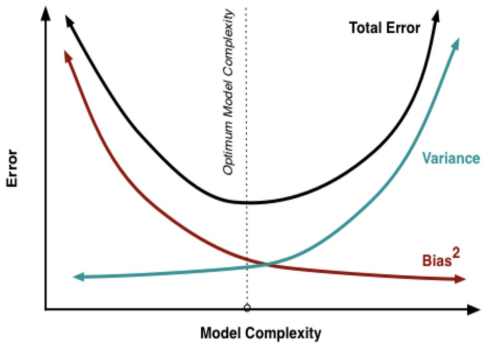
Suggested Reference:
NIPS 2017 Fairness in Machine Learning by Solon Barocas, Moritz Hardt
https://nips.cc/Conferences/2017/Schedule?showEvent=8734

- Spurious correlation is just one particular type of bias that will be addressed in the paper. There are huge amount of literature in ML fairness in recent years, and there is also a recently established conference: ACM FAccT

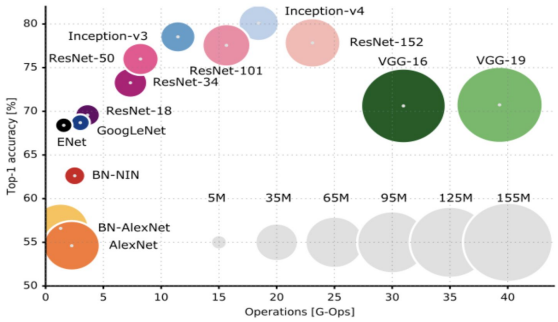
III. Background: Why the need for Overparameterization?

- **Overparameterization:** A model has more parameters than that can be estimated from the data.
  - Examples: Typical neural networks with # parameters >> # samples; Linear regression with more features than the number of data.
- Model size vs Test Risk
  - Traditional wisdom: Test risk cannot be increased further by increasing the model complexity beyond the sweetspot of underfitting/overfitting due to the bias variance tradeoff

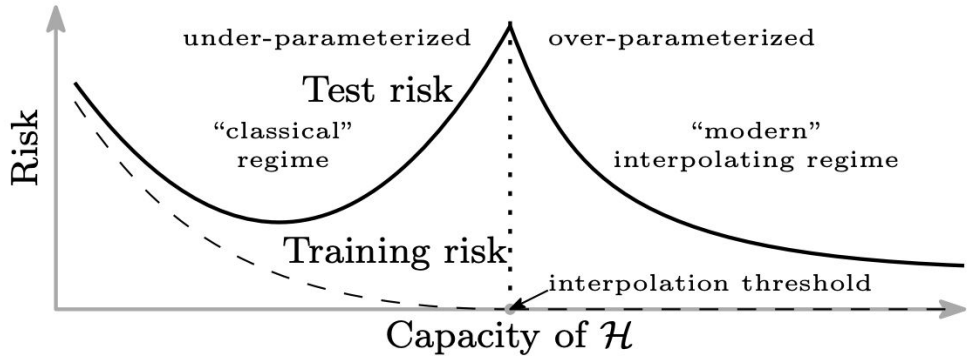




- Recent development of DNN shows that test risk can increase with the model size.



- [Belkin et al 2018] proposes the following double descent curve, and argue
  - 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.



- Many researchers in recent years are able to show that
  - **Inductive bias of SGD-type algorithm leads to the success of over-parameterized model** under Two layer Relu[Zhu et al 2017], Neural tangent kernel [Zhu et al. 2019], Matrix sensing [Yuanzhi et al 2018]
  - Overparameterized formulation becomes an important setting that people will analyze under

Suggested Reading: An empirical studies of the double descent curve: https://openai.com/blog/deep-double-descent/

IV. Problem and Experiment Setup.

- Problem Setup
  - Label (core attribute): y \in {1,-1}
    - Waterbird vs landbird
  - Spurious attribute: a \in {1, -1}
    - Water background vs land background
  - (a,y) forms 4 different groups
  - Goal: Minimize the **worst group error**

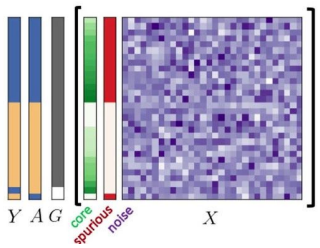
$$\mathrm{Err}_{\mathrm{wg}}(w) := \max_{g \in \mathcal{G}} \mathbb{E}_{x,y|g} \left[\ell_{0-1}(w; (x, y)) \right]$$

  - Baseline method: reweighting (importance weighting)

$$\hat{\mathcal{R}}_{\mathsf{reweight}}(w) = \hat{\mathbb{E}}_{(x,y,g)} \left[\frac{1}{\hat{p}_g} \ell(w; (x, y)) \right]$$

    - Where $\hat{p}_g$ is the fraction of the training samples in group g.
- Empirical setup:
  - For CelebA dataset {hair color, gender}, ResNet10 model and model size is varied by increasing the network width from 1 to 96.
  - 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.
- Analytical setup:

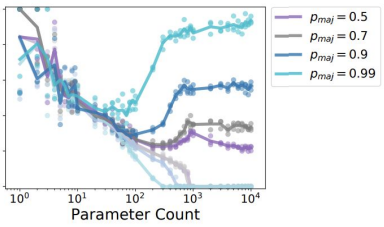
$$x = [x_{\mathsf{core}}, x_{\mathsf{spu}}, x_{\mathsf{noise}}]$$
$$x_{\mathsf{core}} \in \mathbb{R}$$
$$x_{\mathsf{core}} \mid y \sim \mathcal{N}(y, \sigma^2_{\mathsf{core}})$$
$$x_{\mathsf{spu}} \in \mathbb{R}$$
$$x_{\mathsf{spu}} \mid a \sim \mathcal{N}(a, \sigma^2_{\mathsf{spu}})$$
$$x_{\mathsf{noise}} \in \mathbb{R}^N$$
$$x_{\mathsf{noise}} \sim \mathcal{N}\left(0, \frac{\sigma^2_{\mathsf{noise}}}{N} I_N\right)$$



  - The vertical dimension represents the number of samples, and the horizontal dimension represents the number of features.
  - X is the random noise. By controlling the length of X, we can control the degree of over-parameterization

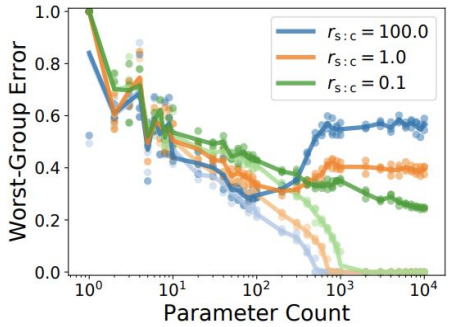
V. Messages from the Empirical and Theoretical Results.

- Message 1: Increasing the major fraction makes overparameterization hurt the worst group error more



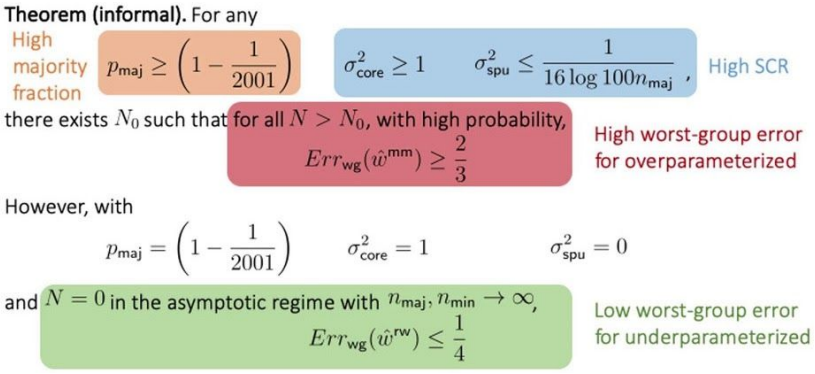
- In the extreme case when the groups are perfectly balanced i.e. p_maj = 0.5, the overparameterization does not hurt at all, since direct ERM works well

- Message 2: Increasing the spurious-core information ratio makes overparameterization hurt the worst group error more



- **Spurious core ratio (SCR)**: $r_{s:c} = \sigma^2_{core}/\sigma^2_{spu}$ (i.e. std of core feature/ std of spurious feature) measures the relative informativeness of the spurious features vs. core features.

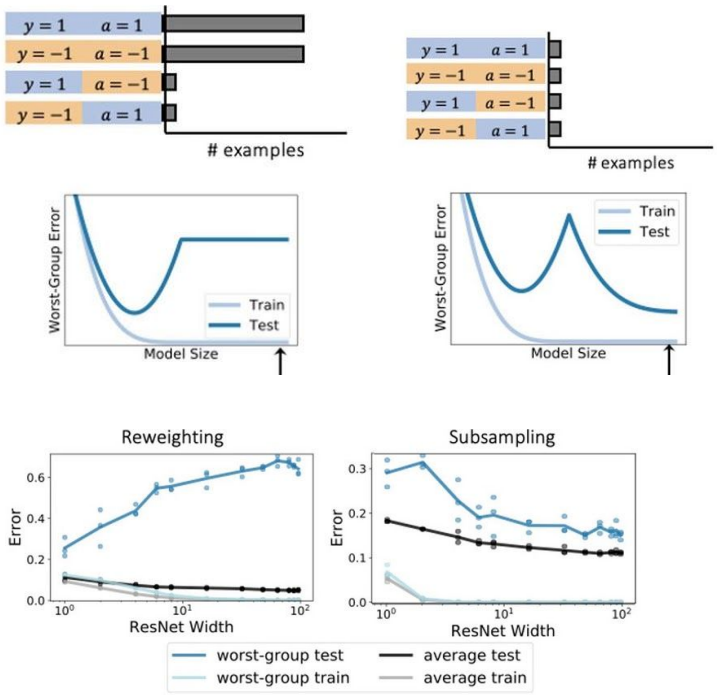
- Combining Mesage 1 and 2, the paper presents the main theorem:



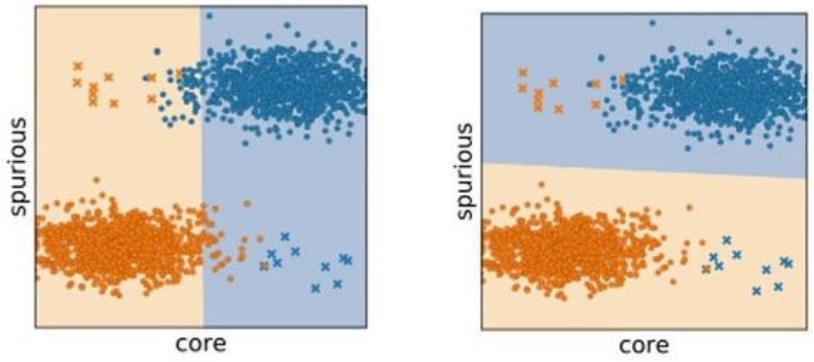
- I.e. Given high majority fraction, and high SCR, we will have high worst-group error for overparameterized but low worst-group error for the under parameterized model
  - Note: the error is evaluated for the minimum norm estimator (or l_2 regularized reweighted estimator) for logistic regression.

- Message 3: Overparameterized model memorizes as few examples as possible under the min-norm inductive bias
  - Under this setup, the paper theoretically show
    - || estimator using the spurious feature ||_2 < || estimator using the core feature||_2 in Prop 1,2

- Message 4: Resampling outperforms Reweighting in worst group error under the over-parameterized setting.

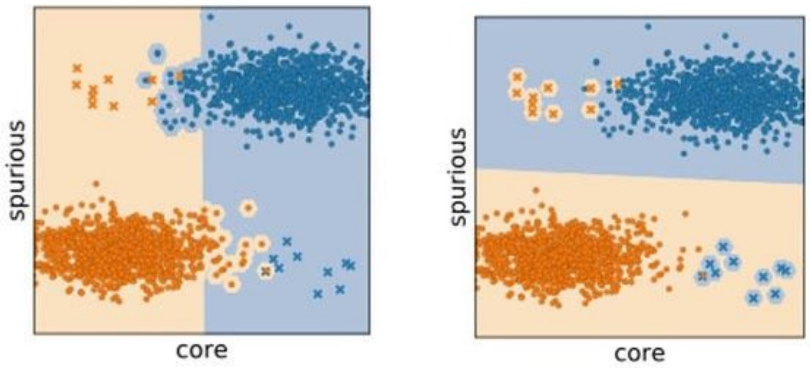


- Resampling reduces the Majority fraction by lowering the memorization cost of learning the core features. The idea is best explained with the following diagram:



-

(With reweighting) (Naive ERM)

- 

  (With Resampling) (Naive ERM or Reweighting)
- In this diagram, let's assume the decision boundary is
  - (Underparameterized setting) linear
  - (Overparameterized setting) nonlinear and can memorize samples
- In the overparameterized setting, naive ERM will remember the few minority examples and assigning large weight to these samples won't change the decision boundary. However, if we perform subsampling on the majority samples, memorizing the outliers for majority samples becomes cheap.
- Caveat of subsampling: We could suffer a drop in average test error. Recent work from [Goel et al 2020] attempts to handle this issue by applying the data augmentation.

Reference

- Reconciling modern machine learning practice and the bias-variance trade-off [Belkin et al. 2018]
- Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization [Sagawa et al. 2020]
- An investigation of why overparameterization exacerbates spurious correlations [Sagawa et al. 2020]
- Towards Understanding the Role of Over-Parameterization in Generalization of Neural Networks [Neyshabur et al. 2018]
- Model Patching: Closing the Subgroup Performance Gap with Data Augmentation [Goel et al 2020]