Understanding The Failure Modes of Out-of-Distribution Generalization

Vaishnavh Nagarajan[1] Anders Andreassen[2] Behnam Neyshabur[2]

[1]Computer Science Department, Carnegie Mellon University [2]Blueshift, Alphabet

High Level Message

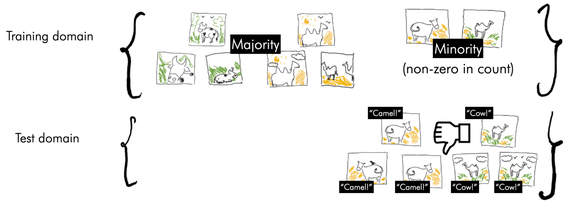
Why do classifiers rely on spurious correlations?

1. Existing theory does not capture the fundamental reasons!
2. We theoretically study GD-trained linear classifiers and discover two fundamental failure modes.
3. We empirically verify these failure modes in deep learning.

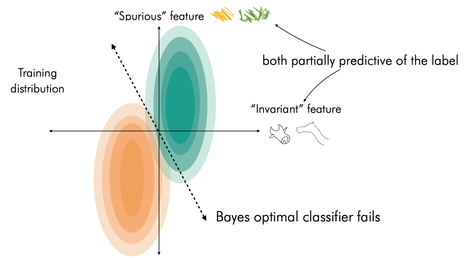
Spurious Correlations

Cow/camel classification

[Arjovsky, Bottou, Gulrajani, Lopez-Paz '19 Beery, Horn, Perona '18]



Gap in Existing Theory vs Reality

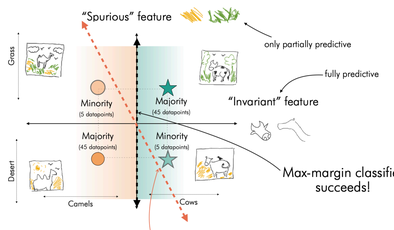


This model forms the basis of a lot of research in this area

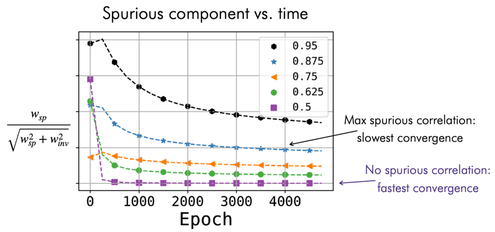
Algorithms for OoD generalization	Synthetic datasets	Theoretical guarantees	Reasoning about spurious correlations
IRM (ARGC19) IRM-CDM (GZLK'21)	ColoredMNIST (ARGC19) CMNIST+ (GZLK'21) Linear Unit Tests (ASABE21)	Rosenfeld, Ravikumar & Risteski '20, Kamath, Tangella, Sutherland & Srebro '21.	Sagawa, Raghunathan, Koh & Liang '20, Tsipras, Santurkar, Engstrom, Turner & Madry '19

In practice, classifiers use the *partially* predictive spurious feature, even when the invariant feature is *fully* predictive!

Failure Mode 1: Statistical Skews



But finite-time-trained GD fails even though it should converge to max-margin [SHNGS'17, JT'18].



Thm 1 (informal): For a large class of linearly separable datasets, under logistic loss,

$$\frac{|w_{sp}(t)|}{\|\vec{w}_{inv}(t)\|} = \Theta\left(\frac{\text{level of spurious correlation}}{\log t}\right)$$

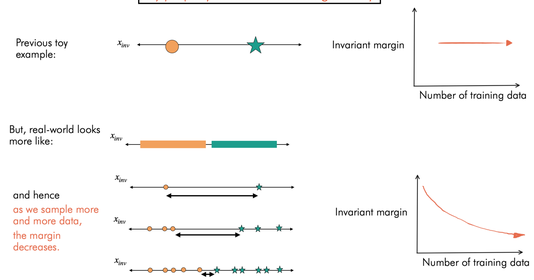
Takeaway: Failure happens due to
(a) bias in finite-time GD &
(b) skew in the "statistics" of the minority vs. majority group.

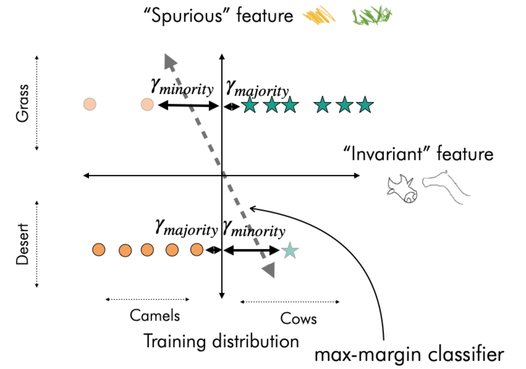
Does max-margin always succeed?

No! We show that when data has non-degenerate geometry, even max-margin classifier can fail…

Failure Mode 2: Geometric Skews

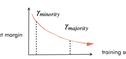
Key property of real-world data geometry





Thm2 (informal): For the max-margin classifier (over all the features),

$$1 - 2\sqrt{\frac{\gamma_{\text{majority}}}{\gamma_{\text{minority}}}} \le w_{sp}$$



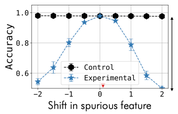
Takeaway: Failure happens because of
(a) the skew in the geometry of the minority and majority group &
(b) margin-maximizing bias.

Empirical Verification: Statistical Skews

A "control" dataset (no spurious correlation) ~~geometric skew~~ ~~statistical skew~~

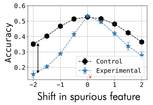
merely duplicate some points

An "experimental" dataset (*with* spurious correlation) ~~geometric skew~~ statistical skew



Colored Binary-MNIST + FNN

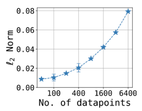
This accuracy gap demonstrates the effect of statistical skews.



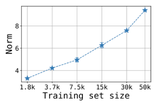
Colored CIFAR10 + ResNet

Empirical Verification: Geometric Skews

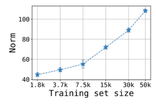
Norms increase with training set size (≈ margin decreases with training set size).



MNIST + max-margin on ReLU random features



Binary-MNIST + FNN



CIFAR10 + ResNet

Future Work

No one unique way by which classifiers fail!

- OoD failure modes unique to how deep representations are learned?
- OoD algorithms specialized towards the two skews?
- Characterize real-world "spurious features" vs. "invariant features"?

[KBM'21]

References

- Arjovsky, Bottou, Gulrajani and Lopez-Paz, "Invariant Risk Minimization", 2019.
- Guo, Zhang, Lio, Kiciman, "Out-of-distribution prediction with IRM: The limitation and an effective fix", 2021
- Aubin, Slowik, Arjovsky, Bottou, Lopez-Paz, "Linear unit-tests for invariance discovery", CDCI ML workshop NeurIPS 2021
- Soudry, Hoffer, Nacson, Gunasekar, and Srebro. "The implicit bias of gradient descent on separable data". J. Mach. Learn. Res., 19, 2018
- Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression
- Sagawa, Raghunathan, Koh, and Liang. An investigation of why overparameterization exacerbates spurious correlations. 2020