Impact of Label Noise on Learning Complex Features

Rahul Vashisht*, P Krishna Kumar*, Harsha Vardhan Govind[1], Harish G. Ramaswamy

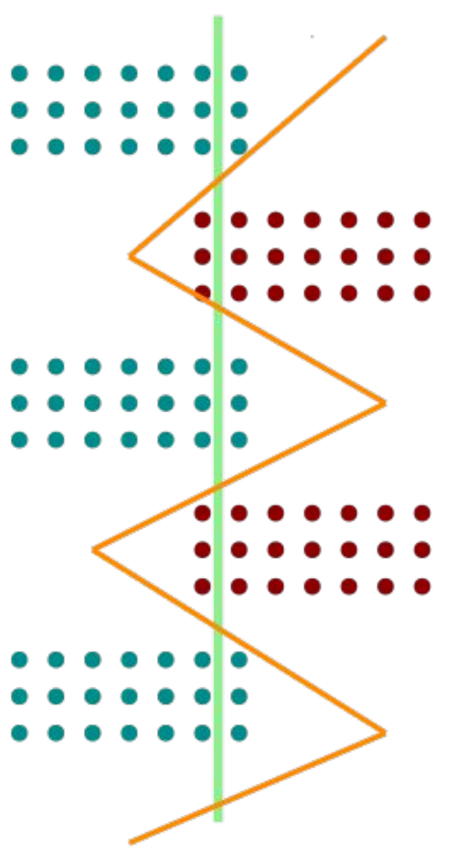
Indian Institute of Technology Madras, [1]IIITDM Kancheepuram

Introduction

Neural networks trained with stochastic gradient descent exhibit an inductive bias towards simpler decision boundaries, typically converging to a narrow family of functions

- We investigate the impact of pre-training models with noisy labels on the dynamics of SGD across various architectures and dataset.
- We show that pre-training with noisy labels encourages gradient descent to find alternate minima that do not solely depend upon simple feature.
- Model begins to leverage a broader range of features and improved ***out-of-distribution generalization***

Ill-effects of Extreme Simplicity Bias



- **Susceptible to perturbation attack**: Neural networks that learn simple functions lack robustness
- **Suboptimal generalization:** performance because more powerful discriminative features are ignored
- **Out-of-distribution performance**: poor due to excessively simple decision boundaries
- Can we **mitigate** SB?

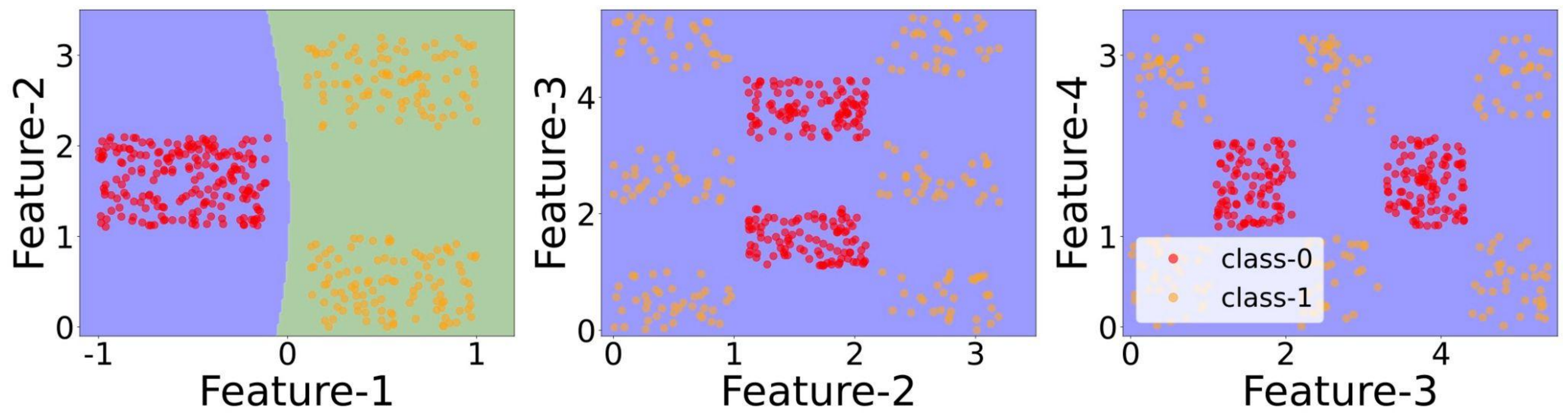
In the **initial epochs**, the model learned by SGD can be explained by a **linear classifier**, and later as the epochs progress, SGD learns functions of increasing complexity.

- Complex features are often overshadowed by the amplification and replication of simpler features
- *Ensembling* and *Adversarial training* fail to effectively address the limitations imposed by this bias

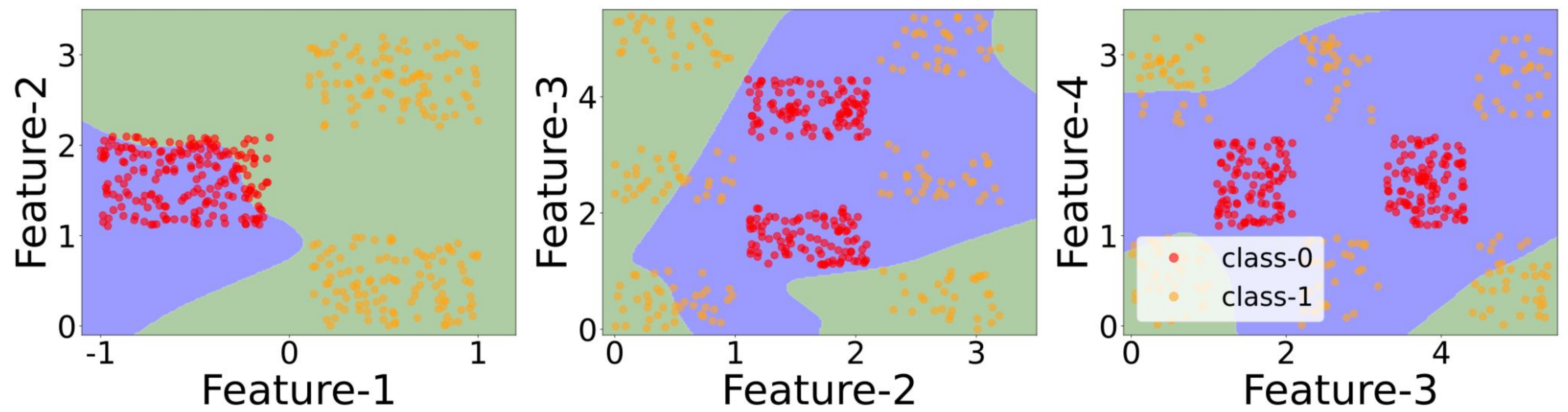
Effect of Noisy Perturbations in Data

Hypothesis: Training neural networks with SGD under noisy data can partially mitigate simplicity bias.

- We consider a 4-dimensional slab data, with each dimension having increasing complexity to classify the data.

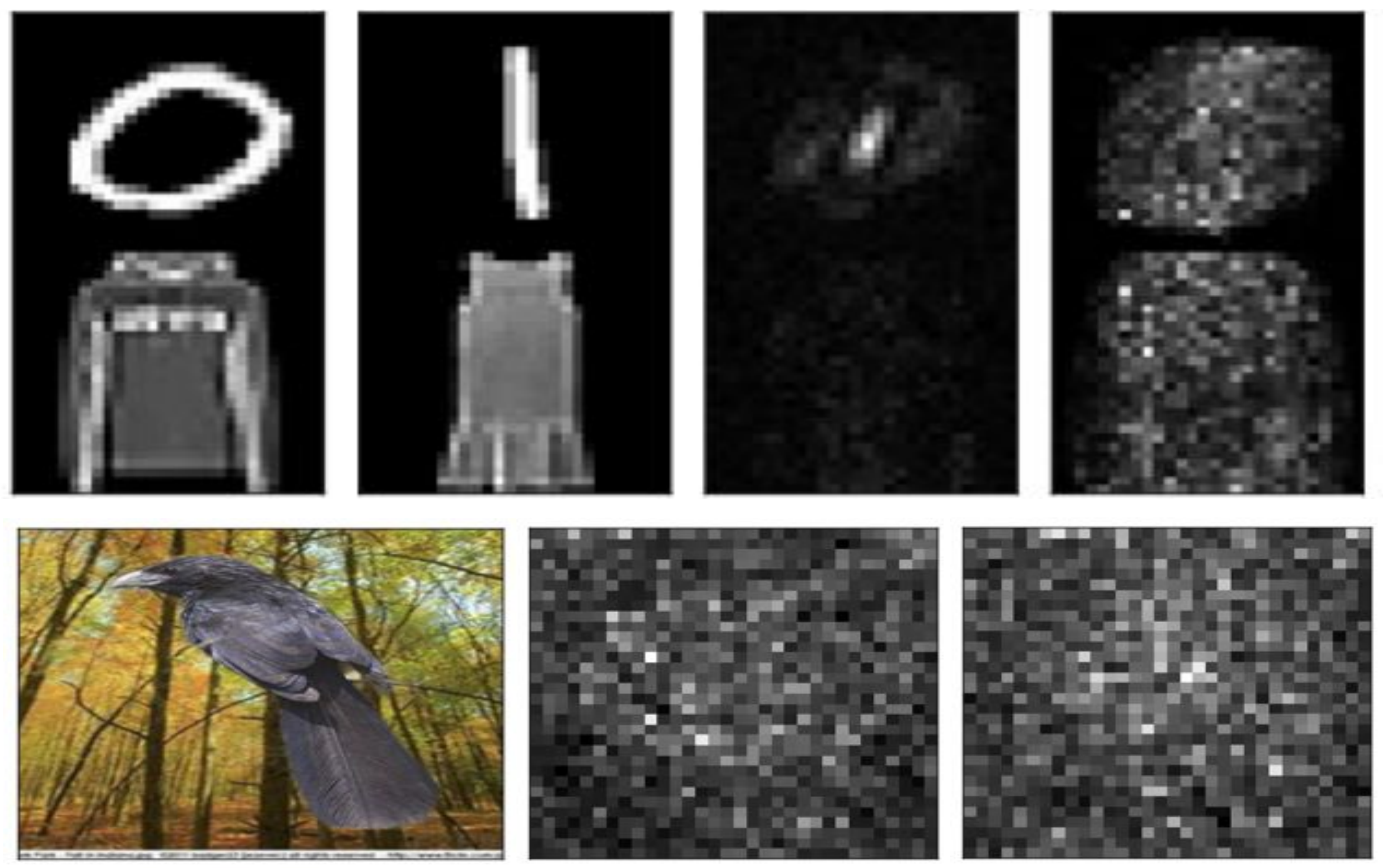


Learned decision depends only on the first two dimensions.



Training on noisy data leads to dependence on the other dimensions.

Datasets for Measuring Feature Dependence



WaterBirds & Dominoes Dataset with Gram matrix

Randomized Performance

Data	Standard Training		Noisy Pre-training	
	MNIST Rnd.	F-MNIST Rnd.	MNIST Rnd.	F-MNIST Rnd.
$\mathcal{D}$	52.5 ± 0.33	98.3 ± 0.05	53.6 ± 1.56	88.6 ± 0.76
$\mathcal{D}'$	93.1 ± 0.33	56.5 ± 0.42	81.2 ± 1.02	57.2 ± 1.50

All Models 100% Training Accuracy

Data	Standard Training		Noisy Pre-training	
	In-group	Out-group	In-group	Out-group
$\mathcal{D}$	85.2 ± 0.43	38.5 ± 0.88	78.1 ± 1.02	44.1 ± 1.60
$\mathcal{D}'$	84.1 ± 0.48	44.4 ± 0.67	77.8 ± 1.15	46.9 ± 0.92

WaterBirds Dataset

Parallels of Label Smoothing

Ground truth labels are a mixture of one-hot-vectors and uniform distribution that acts as addition of noise to ground truth.

Data	LS coeff.	Standard Training		Noisy Pre-training	
		In-group	Out-group	In-group	Out-group
$\mathcal{D}$	0.0	85.2 ± 0.43	38.5 ± 0.88	78.1 ± 1.02	44.1 ± 1.60
	0.2	84.4 ± 0.42	43.5 ± 3.62	77.3 ± 0.75	51.1 ± 2.17
$\mathcal{D}'$	0.0	84.1 ± 0.48	44.4 ± 0.67	77.8 ± 1.15	46.9 ± 0.92
	0.2	83.4 ± 0.49	49.1 ± 3.63	78.5 ± 0.28	52.2 ± 1.29

WaterBirds with 100% (D) and 95% (D') bg correlation

Discussion & Conclusions

- Although SGD has strong implicit regularization, we show that noisy-label pre-training can successfully trap models in complex local minimas.
- Overparameterized neural networks **can learn more complex and diverse features** with the right initialization.
- Deep neural networks learn broader set of features when pre-trained on noisy labels