SGD on Neural Nets Learns Functions of Increasing Complexity

Preetum Nakkiran, Gal Kaplun, Dimitris Kalimeris, Tristan Yang,
Benjamin L. Edelman, Fred Zhang and Boaz Barak

Harvard University

NeurIPS 2019
Poster #170

Motivation: Why do Neural Nets generalize?

- Optimization algorithm matters:
  Not sufficient to “minimize train loss” arbitrarily [Zhang et al. 2017]

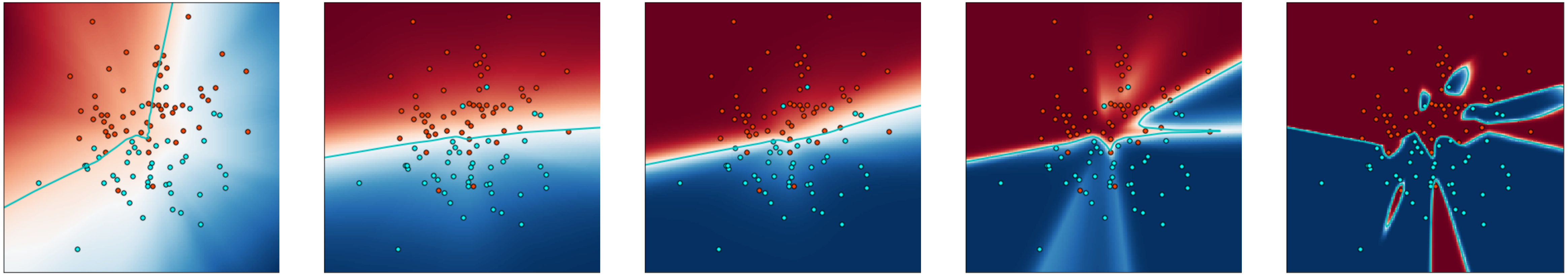
- Informal conjecture: SGD outputs “low complexity” classifiers

This work: Formalizing this conjecture

Main Claims (informal)

Claim 1: SGD starts by learning an "**essentially linear**" classifier

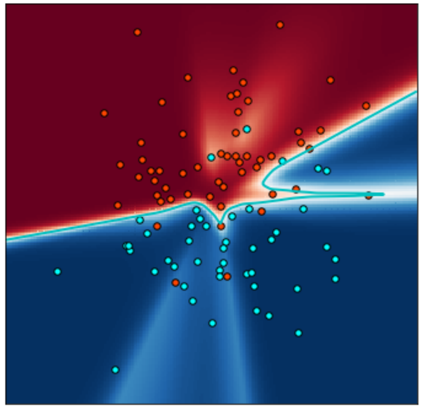
Claim 2: In later stages, SGD learns models of increasing complexity.

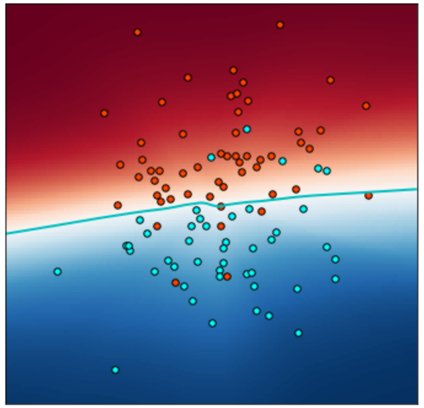


Increasing # epochs →

Performance Correlation

"How well **performance** of complex model
is explained by a simple model "





Performance Correlation

Input distribution : $x \sim D$

Joint distribution $\{ \underset{\text{True label}}{\underset{\uparrow}{Y(x)}}, \quad \underset{\text{NN output}}{\underset{\uparrow}{F(x)}}, \quad \underset{\text{Linear model}}{\underset{\uparrow}{L(x)}} \}_{x \sim D}$

(1) $I(F; Y)$: Accuracy of Neural Network

(2) $I(F; Y \mid L)$: "Unexplained accuracy"

(How much more F reveals about Y, after knowing L)

Performance Correlation

Input distribution : $x \sim D$

Joint distribution $\{ Y(x), \;\; F(x), \;\; L(x) \}_{x \sim D}$

True label NN output Linear model

(1) $I(F;Y)$: Accuracy of Neural Network

(2) $I(F;Y \mid L)$: "Unexplained accuracy"

(How much more F reveals about Y, after knowing L)

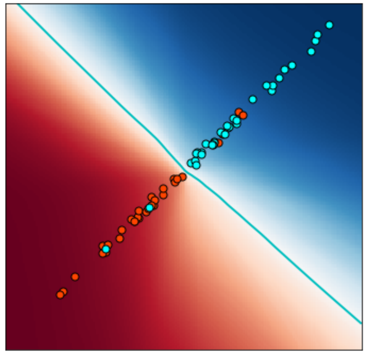
<u>Defn: Performance Correlation</u>

$$\mu_Y(F;L) \coloneqq I(F;Y) - I(F;Y \mid L)$$

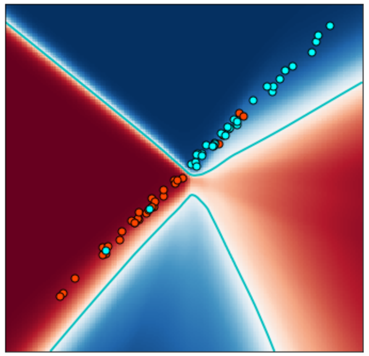
Accuracy of F explained by the linear classifier L.

Performance Correlation: Properties

1. Depends only on predictions of F **on distribution D**



F: Linear everywhere

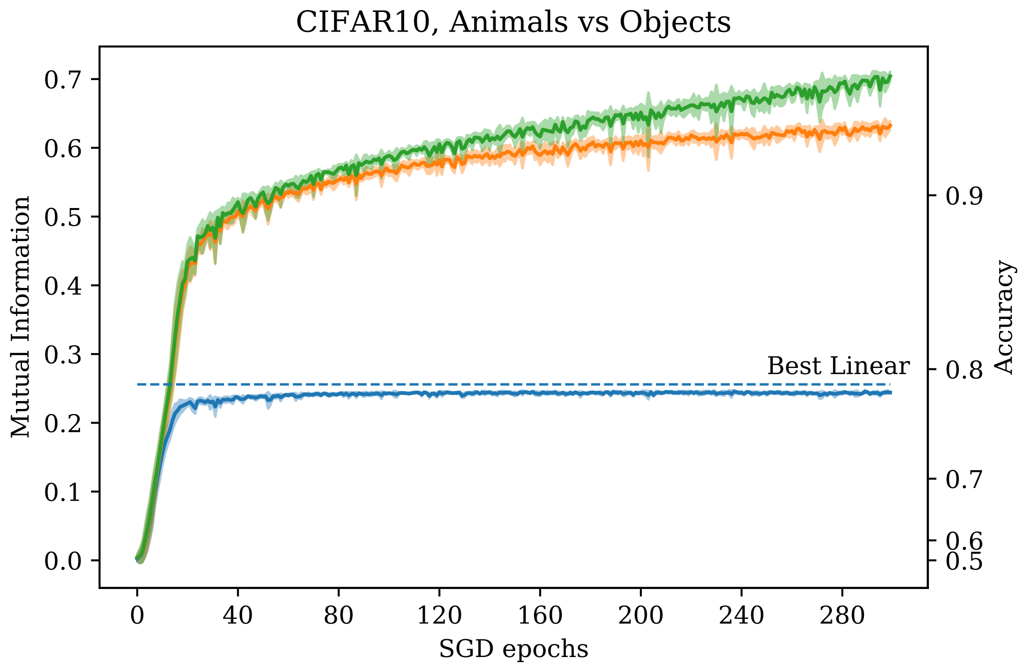


F: Linear **on distribution**

2. Ignore component of F that is **nonlinear**, but not **useful** to predict Y

Eg: $F = L + noise$ is still fully explained by L

Experiments: Linear Learning

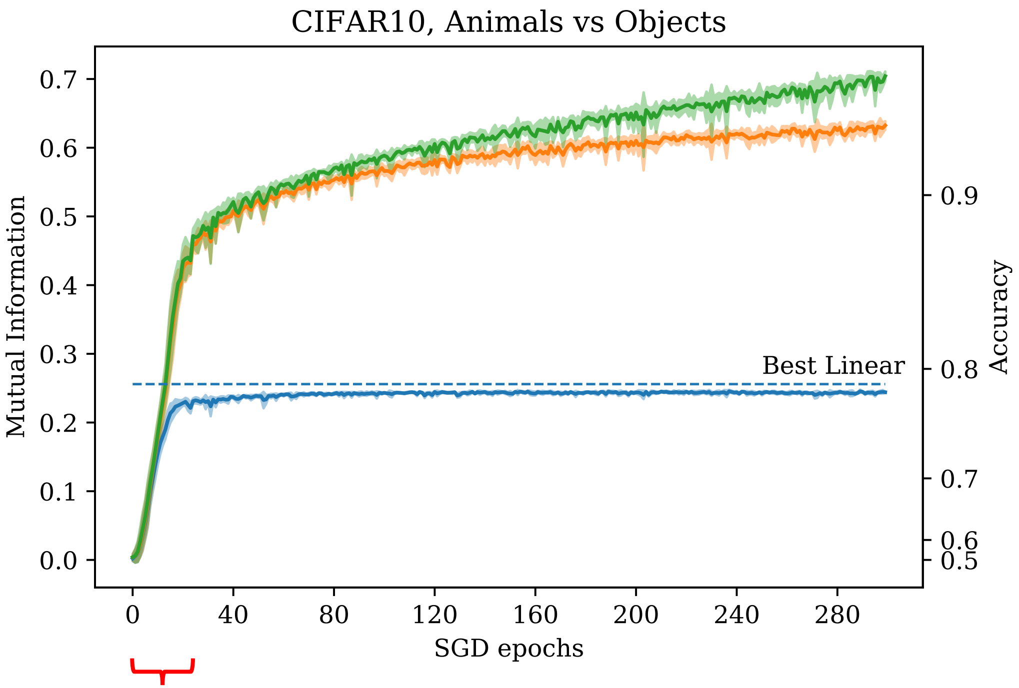


Train Accuracy F_t

Test Accuracy F_t

Performance Correlation $\mu_Y(F_t; L)$

Experiments: Linear Learning



Linear learning

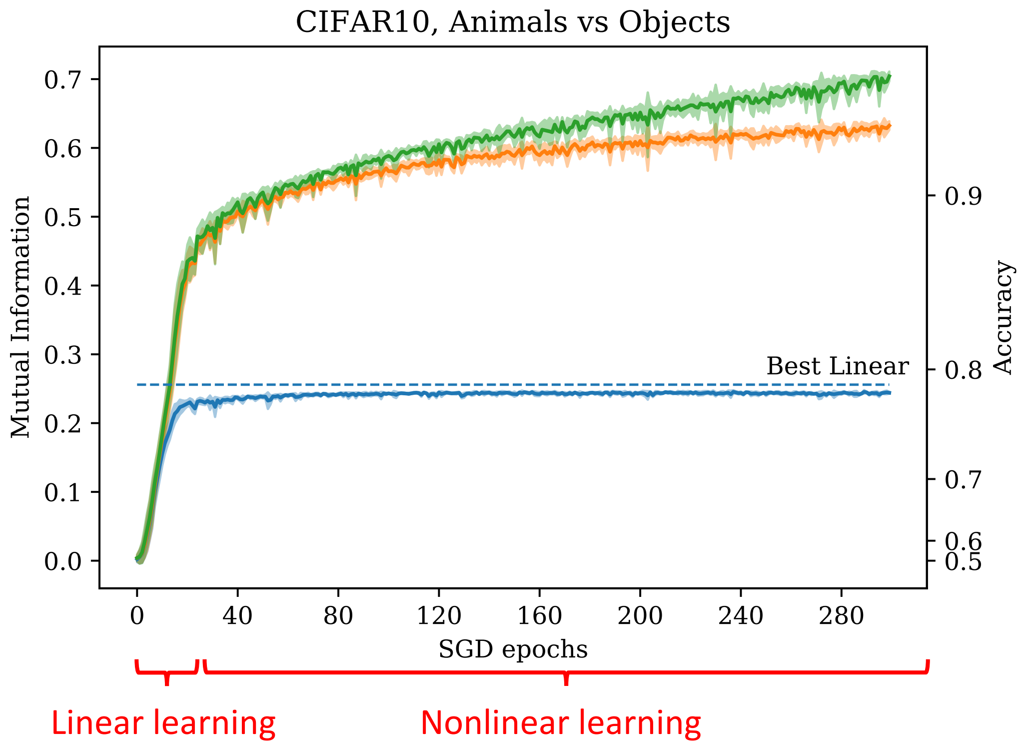
Train Accuracy F_t

Test Accuracy F_t

Performance Correlation $\mu_Y(F_t; L)$

1. Linear learning phase:
   - NN explained by linear model
   - Lasts until NN matches the best linear model

Experiments: Linear Learning



Train Accuracy F_t

Test Accuracy F_t

Performance Correlation $\mu_Y(F_t; L)$

1. Linear learning phase:
   - NN explained by linear model
   - Lasts until NN matches the best linear model

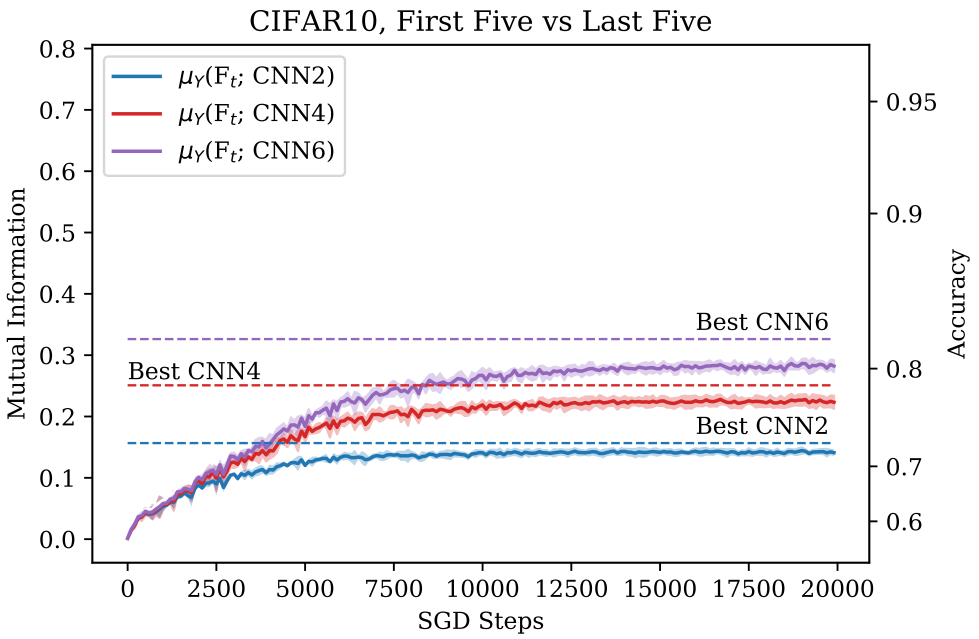
2. Nonlinear learning phase:
   - NN becomes nonlinear
   - Retains linear component

Holds for variety of real & synthetic tasks (CIFAR, MNIST, MLPs, CNNs).

Experiments: Increasing Complexity

<u>Generalized Hypothesis (informal):</u>

SGD learns functions of increasing complexity



Consider $\mu_Y(F_t; CNNk)$:
"How well NN explained by small k-layer CNN"

Conclusion

Our Work:

1. Introduce **performance correlation**
2. SGD initially learns an **essentially linear function**, then more complex ones

Future Work:

- Better understanding of why NNs generalize,
  by studying implicit bias of SGD **throughout training**

Thanks!

Poster #170