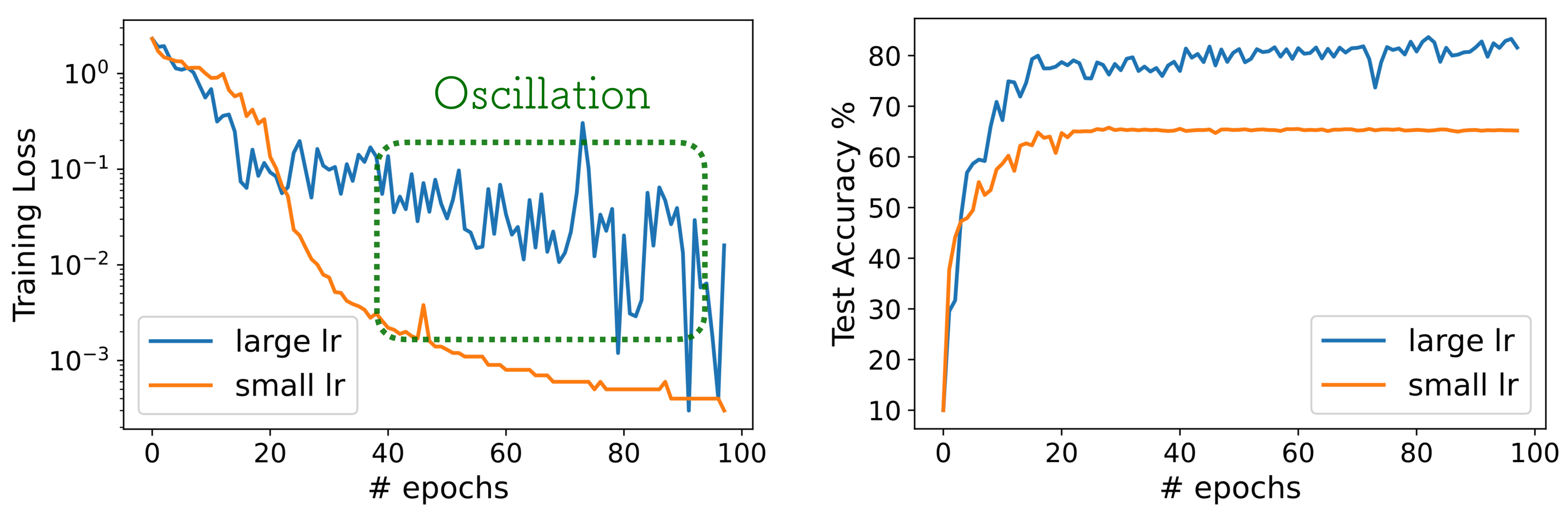
Benign Oscillation of Stochastic Gradient Descent with Large Learning Rate

Miao Lu[1], Beining Wu[2], Xiaodong Yang[3], and Difan Zou[4]

Motivation and Observations

Large learning rate NN Training can result in **better generalization**. But why?

- Small LR training: curve is much smoother, converges more rapidly.
- **Large LR** training: an "**oscillating**" training curve, but better generalization!



ResNet-18 on CIFAR-10: Small LR v.s. Large LR

Oscillation during training can be closely tied to better generalization of SGD with Large LR!

What is the mechanism behind?

Our Key Message

The oscillation prevents the over-greedy convergence and serves as the engine that drives the learning of less-prominent data pattern.

- Allow for **all** useful data patterns to be discovered and learned :)
- These data patterns are beneficial for the NN to generalize to unseen data!
- We refer to such a phenomenon as "**benign oscillation**".

Our Main Contributions:

- Dynamic analysis framework for SGD with large learning rates.
- A new theoretical argument for feature learning driven by oscillation.
- Division for generalization by different learning rates.

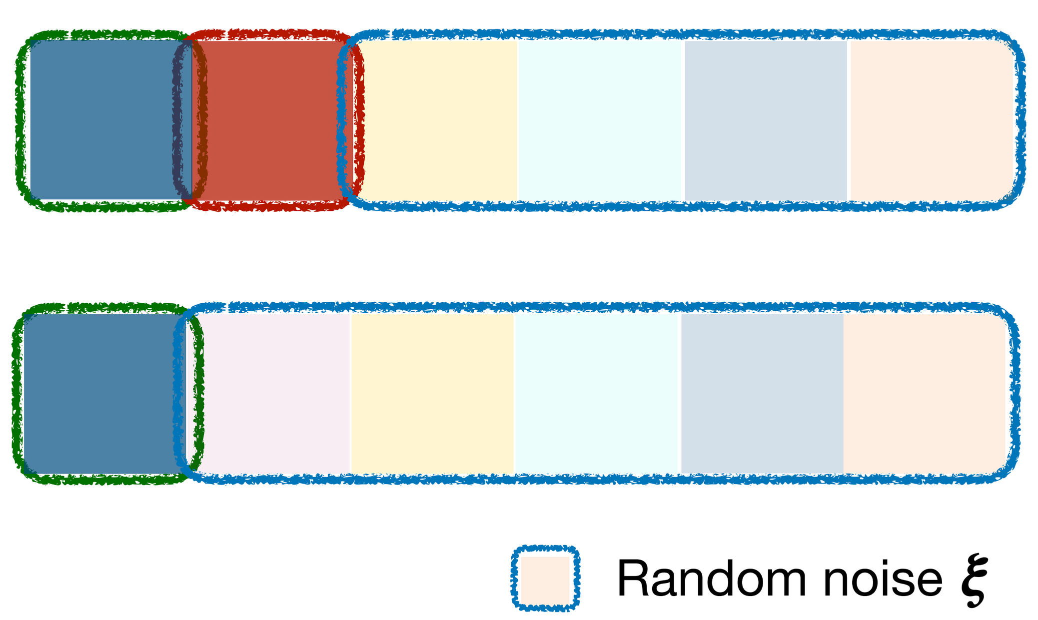
Theoretical Demonstration and Finding

Strong feature v.s. Weak feature Model

- **Strong feature patch:** with probability $1-\rho$, this patch is the strong feature $y \cdot \mathbf{u}$, otherwise this patch is a random noise $\boldsymbol{\xi}$.
- **Weak feature patch:** always taken by the weak feature $y \cdot \mathbf{v}$.
- **Random noise patch:** random Gaussian noise $\boldsymbol{\xi}$ is randomly sampled for the remaining patches

Strong data:
- Has strong/weak feature
- Only appear with probability $1-\rho$

Weak data:
- Only has weak feature
- Every data has the weak feature



To generalize to **all** new data points, the NN must effectively learn the weak feature patch in face of the strong feature!

Network on strong data with label 1:

$$f(\mathbf{w};\mathbf{x}) = \sum_{r=1}^{m} \sigma(\langle \mathbf{w}_r, \mathbf{u}\rangle) + \sigma(\langle \mathbf{w}_r, \mathbf{v}\rangle) + \text{noise}$$

Loss function: $L(\mathbf{w}) = \big(f(\mathbf{w};\mathbf{x}) - 1\big)^2$

Why oscillation helps? A dynamic analysis

- **Strong signal oscillation:**

$$-\sum_{s\in\mathcal{S}^-} \big(1 - f(\mathbf{w}^{(s)};\mathbf{x})\big) \cdot \langle \mathbf{w}_r^{(s)}, \mathbf{u}\rangle \approx \sum_{s\in\mathcal{S}^+} \big(1 - f(\mathbf{w}^{(s)};\mathbf{x})\big) \cdot \langle \mathbf{w}_r^{(s)}, \mathbf{u}\rangle$$

Accumulation of negative updates — Accumulation of positive updates

Since larger $\langle \mathbf{w}^{(s)}, \mathbf{u}\rangle$ implies larger $f(\mathbf{w}^{(s)};\mathbf{x})$, we have:

$$\sum_{s=t_0}^{t_1} \big(1 - f(\mathbf{w}^{(s)});\mathbf{x}\big) = \Omega(\delta \cdot (t_1 - t_0))$$

Average oscillation magnitude (δ)

Oscillation accumulates!

Boosting

- **Weak signal learning:**

$$\sum_{s=t_0}^{t_1} \big(1 - f(\mathbf{w}^{(s)};\mathbf{x})\big) \cdot \langle \mathbf{w}_r^{(s)}, \mathbf{v}\rangle \approx \sum_{s=t_0}^{t_1} \big(1 - f(\mathbf{w}^{(s)};\mathbf{x})\big) \cdot \langle \mathbf{w}_r^{(t_0)}, \mathbf{v}\rangle$$

$$\geq \Omega(\delta \cdot (t_1 - t_0)) \cdot \langle \mathbf{w}_r^{(t_0)}, \mathbf{v}\rangle$$

Longer oscillation implies larger weak signal learning

Division of generalization properties

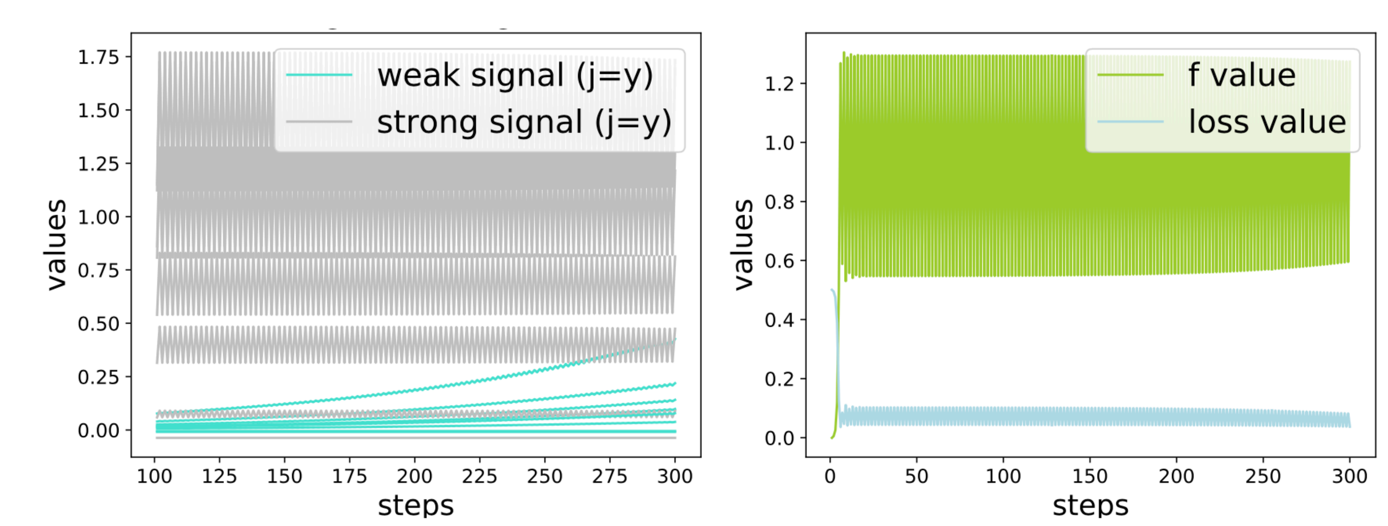
- **SGD, large LR:** learns both strong/weak features.
- **SGD, small LR:** only learns the strong feature.

This explains why SGD with large LR can generalize better!

Experimental verification

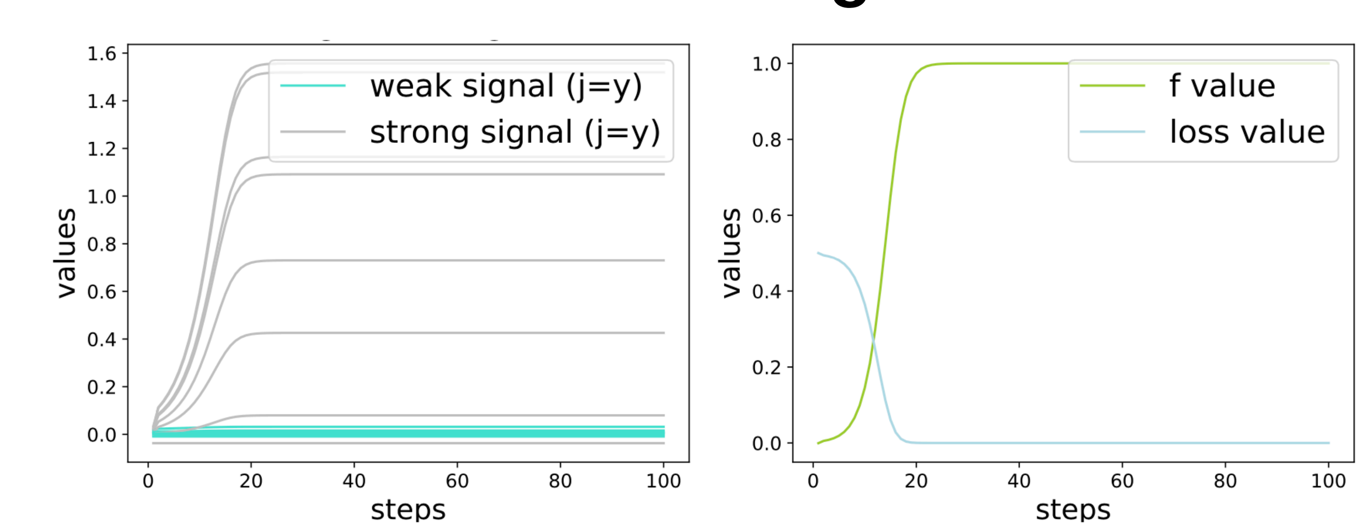
Illustrative setup: single data with two patches

- **SGD with large learning rate:**



- Oscillation occurs.
- **Both** strong and weak features are learned.

- **SGD with small learning rate:**



- Convergence is smooth.
- **Only** strong feature is learned.