Understanding Gradient Descent on Edge of Stability in Deep Learning

Sanjeev Arora, Zhiyuan Li, Abhishek Panigrahi
Department of Computer Science, Princeton University
{arora,zhiyuanli,ap34}@cs.princeton.edu
@prfsanjeevarora,@zhiyuanli_,@Abhishek_034

The authors are supported by NSF, ONR, Simons Foundation, DARPA, and SRC.
Zhiyuan Li is also supported by Microsoft Research Ph.D. Fellowship.

Descent Lemma for Gradient Descent

Underpins most convergence proofs in Deep Learning

Descent Lemma for Gradient Descent

Underpins most convergence proofs in Deep Learning

If $\eta < \dfrac{2}{\lambda_{max}(\nabla^2 L)}$ then loss drops each iteration

Learning Rate (LR)

Sharpness
(aka smoothness)

Hessian of loss L

`

Descent Lemma for Gradient Descent

Underpins most convergence proofs in Deep Learning

If $\eta < \dfrac{2}{\lambda_{max}(\nabla^2 L)}$ then loss drops each iteration

Learning Rate (LR)

Sharpness
(aka smoothness)

Hessian of loss L

Usual interpretation: $\lambda_{max}(\nabla^2 L)$ is globally bounded; trial and error is used to discover η that satisfies descent lemma.

Edge of Stability (EoS)

Cohen et al. [2021]

Edge of Stability (EoS)

Cohen et al. [2021]

Finding: GD in popular architectures violates descent lemma.

Edge of Stability (EoS)

Cohen et al. [2021]

Finding: GD in popular architectures violates descent lemma.

EoS phase

Edge of Stability (EoS)

Cohen et al. [2021]

Finding: GD in popular architectures violates descent lemma.

EoS phase

- $\lambda_{max}(\nabla^2 L)$ along trajectory increases above $2/\eta$, then levels off.

Edge of Stability (EoS)

Cohen et al. [2021]

Finding: GD in popular architectures violates descent lemma.

EoS phase

- $\lambda_{max}(\nabla^2 L)$ along trajectory increases above $2/\eta$, then levels off.
- Loss oscillates across iterations, with overall downward trend.

Edge of Stability (EoS)

Cohen et al. [2021]

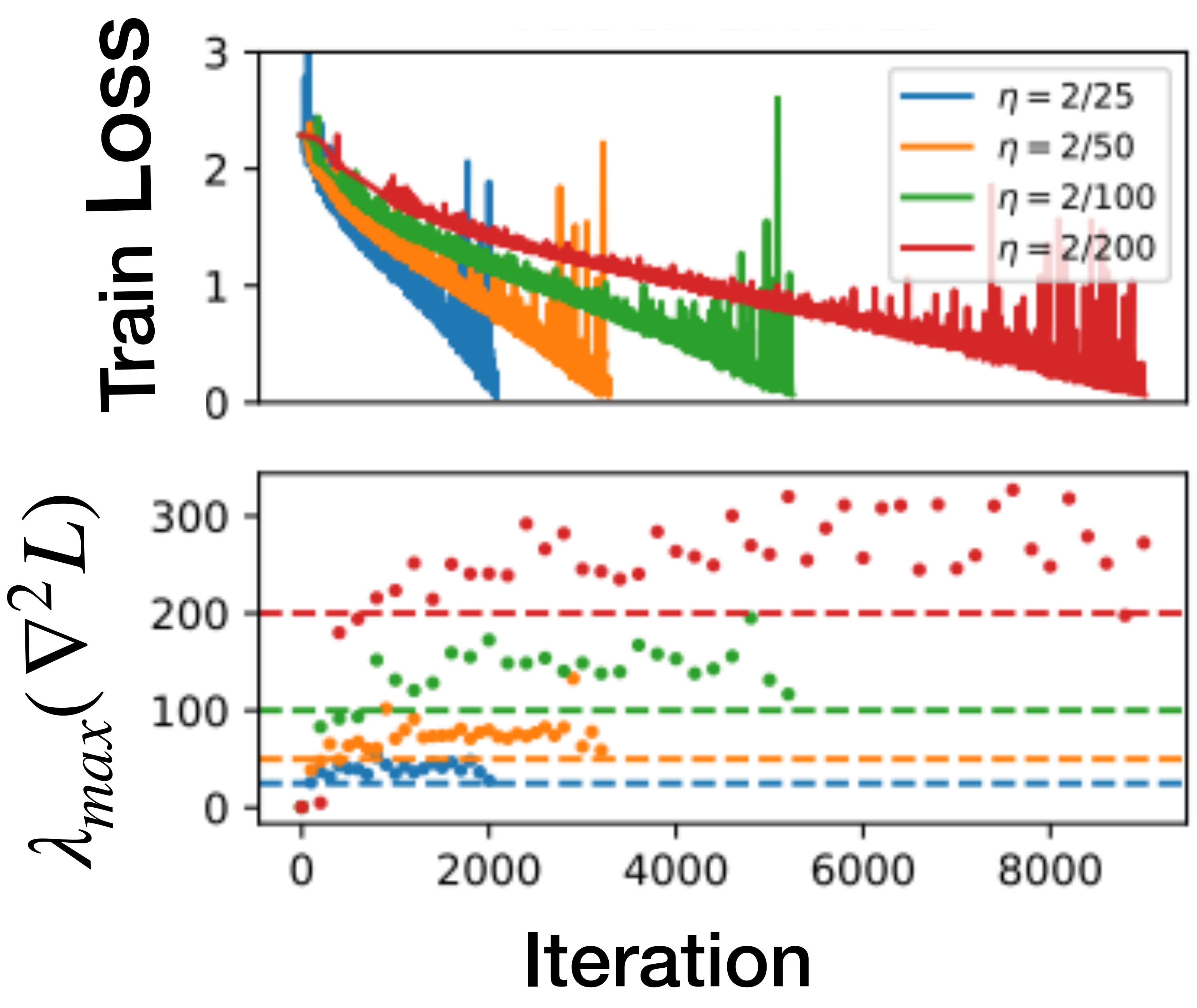
Finding: GD in popular architectures violates descent lemma.

EoS phase

- $\lambda_{max}(\nabla^2 L)$ along trajectory increases above $2/\eta$, then levels off.
- Loss oscillates across iterations, with overall downward trend.

- Phenomenon appears for all finite η.

VGG-16 on CIFAR-10



(Also shown for other architectures)

Edge of Stability (EoS)

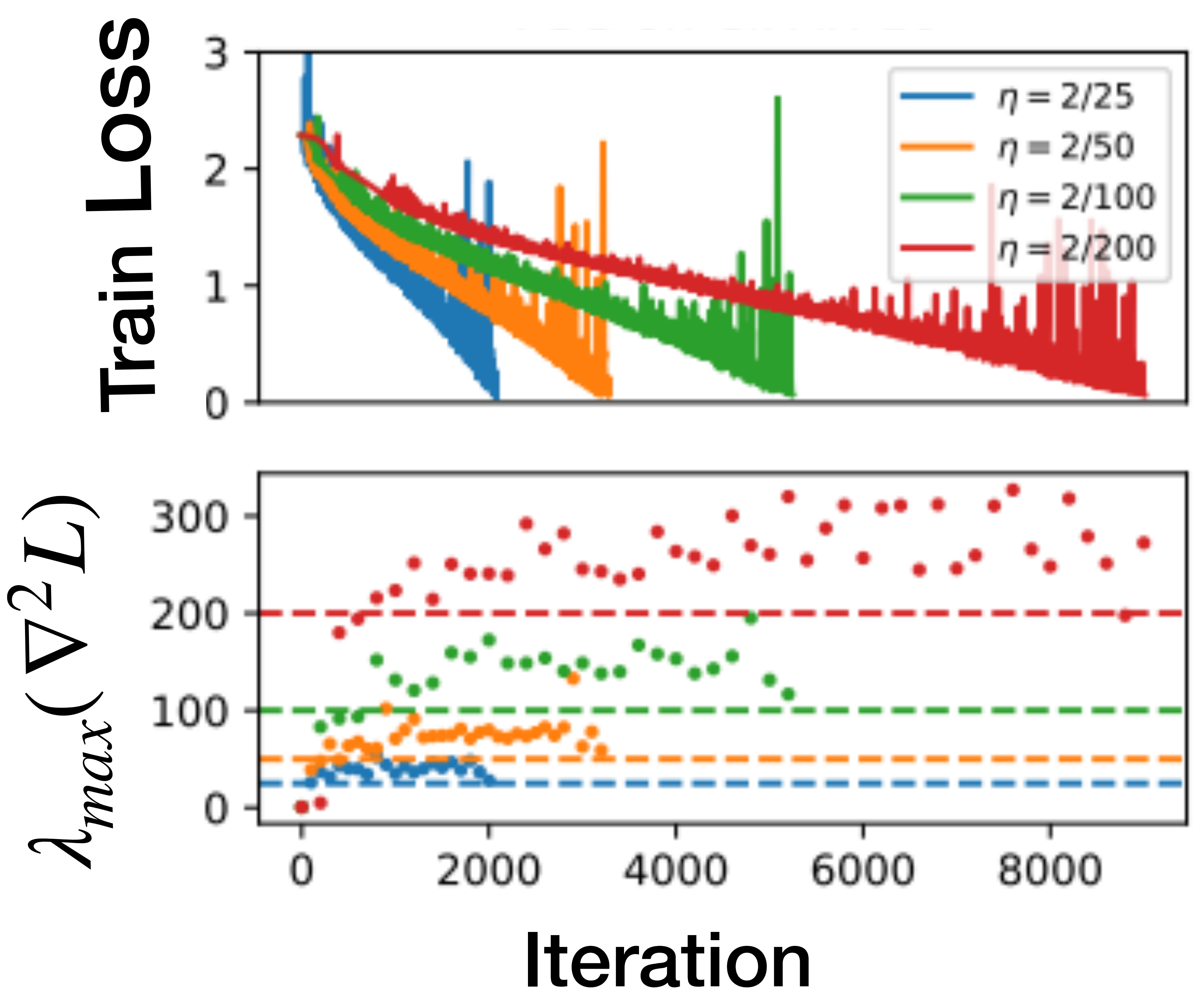
Cohen et al. [2021]

Finding: GD in popular architectures violates descent lemma.

EoS phase

- $\lambda_{max}(\nabla^2 L)$ along trajectory increases above $2/\eta$, then levels off.
- Loss oscillates across iterations, with overall downward trend.

VGG-16 on CIFAR-10



(Also shown for other architectures)

- Phenomenon appears for all finite η.

1. How can we analyze optimization in EoS setting? (Given that descent lemma fails)

Edge of Stability (EoS)

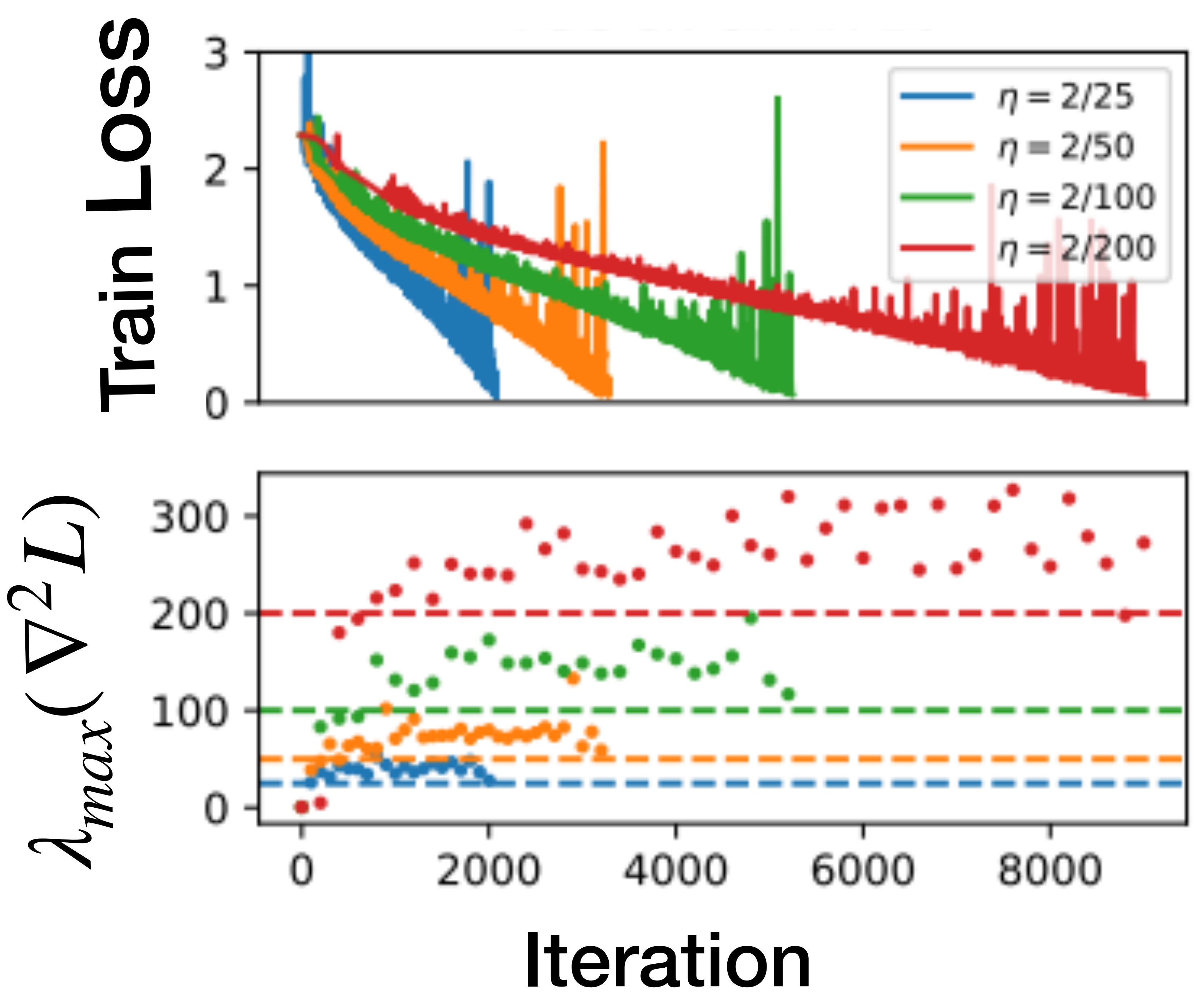
Cohen et al. [2021]

Finding: GD in popular architectures violates descent lemma.

EoS phase

- $\lambda_{max}(\nabla^2 L)$ along trajectory increases above $2/\eta$, then levels off.
- Loss oscillates across iterations, with overall downward trend.

VGG-16 on CIFAR-10



(Also shown for other architectures)

- Phenomenon appears for all finite η.

1. How can we analyze optimization in EoS setting? (Given that descent lemma fails)

2. What mechanism controls $\lambda_{max}(\nabla^2 L)$ in the EoS phase? 🤔

This paper (* setting 1): GD on $\sqrt{L}$

($\min_x L(x) = 0$, **with smooth** L)

(*Setting 2: Normalized GD; see paper)

This paper (* setting 1): GD on $\sqrt{L}$

($\min_x L(x) = 0$, **with smooth** L)

Note: $\lambda_{max}(\nabla^2 \sqrt{L})$ diverges when $L \to 0$ and $\nabla^2 L$ has rank at least 2.

(*Setting 2: Normalized GD; see paper)

This paper (* setting 1): GD on $\sqrt{L}$

($\min_x L(x) = 0$, with smooth L)

Note: $\lambda_{max}(\nabla^2 \sqrt{L})$ diverges when $L \to 0$ and $\nabla^2 L$ has rank at least 2.

Theorem: GD on loss $\sqrt{L}$ for small η has two phases.

(*Setting 2: Normalized GD; see paper)

This paper (* setting 1): GD on $\sqrt{L}$

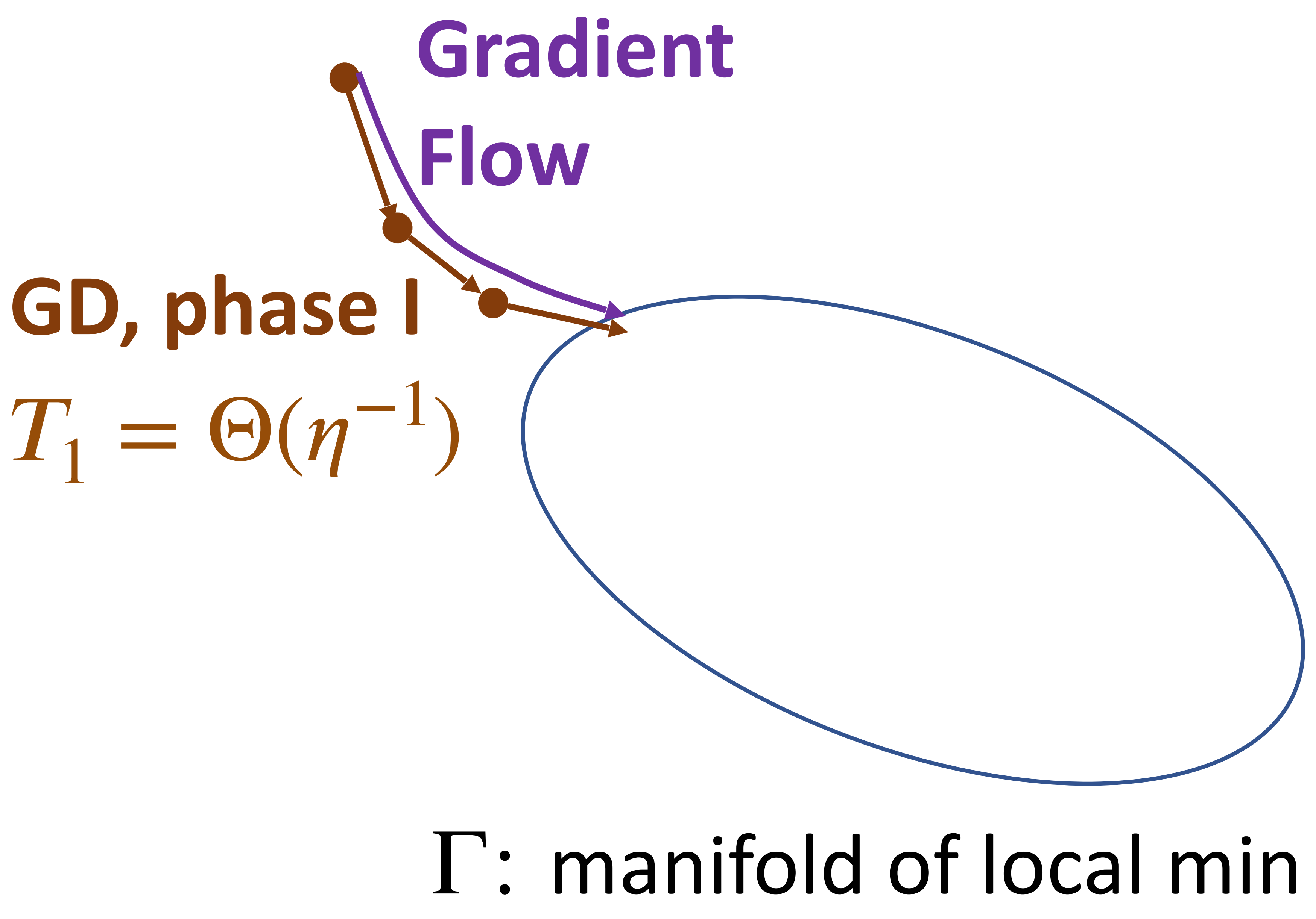
($\min_x L(x) = 0$, with smooth L)

Note: $\lambda_{max}(\nabla^2 \sqrt{L})$ diverges when $L \to 0$ and $\nabla^2 L$ has rank at least 2.

Theorem: GD on loss $\sqrt{L}$ for small η has two phases.

Phase 1:

Loss monotonically decreases till it becomes $\mathcal{O}(\eta)$ in $\Theta(1/\eta)$ steps.



(*Setting 2: Normalized GD; see paper)

This paper (* setting 1): GD on $\sqrt{L}$

$(\min_x L(x) = 0$, **with smooth** $L)$

Note: $\lambda_{max}(\nabla^2 \sqrt{L})$ diverges when $L \to 0$ and $\nabla^2 L$ has rank at least 2.

Theorem: GD on loss $\sqrt{L}$ for small η has two phases.

This paper (* setting 1): GD on $\sqrt{L}$

$(\min_x L(x) = 0$, **with smooth** $L)$

Note: $\lambda_{max}(\nabla^2\sqrt{L})$ diverges when $L \to 0$ and $\nabla^2 L$ has rank at least 2.

Theorem: GD on loss $\sqrt{L}$ for small η has two phases.

Phase 2:

This paper (* setting 1): GD on $\sqrt{L}$

$(\min_x L(x) = 0$, **with smooth** $L)$

Note: $\lambda_{max}(\nabla^2 \sqrt{L})$ diverges when $L \to 0$ and $\nabla^2 L$ has rank at least 2.

Theorem: GD on loss $\sqrt{L}$ for small η has two phases.

Phase 2:

For $\Theta(1/\eta^2)$ steps,

This paper (* setting 1): GD on $\sqrt{L}$

$(\min_x L(x) = 0$, **with smooth** $L)$

Note: $\lambda_{max}(\nabla^2 \sqrt{L})$ diverges when $L \to 0$ and $\nabla^2 L$ has rank at least 2.

Theorem: GD on loss $\sqrt{L}$ for small η has two phases.

Phase 2:

For $\Theta(1/\eta^2)$ steps,

1. $\sqrt{L}(x(t)) + \sqrt{L}(x(t+1))$
$$= \eta \lambda_{\max}(\nabla^2 L(x(t))) + \mathcal{O}(\eta^2)$$

This paper (* setting 1): GD on $\sqrt{L}$

$(\min_x L(x) = 0$, **with smooth** $L)$

Note: $\lambda_{max}(\nabla^2 \sqrt{L})$ diverges when $L \to 0$ and $\nabla^2 L$ has rank at least 2.

Theorem: GD on loss $\sqrt{L}$ for small η has two phases.

Phase 2:

For $\Theta(1/\eta^2)$ steps,

1. $\sqrt{L}(x(t)) + \sqrt{L}(x(t+1))$

$$= \eta \lambda_{\max}(\nabla^2 L(x(t))) + \mathcal{O}(\eta^2)$$

2. $\lambda_{\max}(\nabla^2 L)$ decreases at a rate $\Theta(\eta^2)$.

This paper (* setting 1): GD on $\sqrt{L}$

($\min_x L(x) = 0$, **with smooth** L)

Note: $\lambda_{max}(\nabla^2 \sqrt{L})$ diverges when $L \to 0$ and $\nabla^2 L$ has rank at least 2.

Theorem: GD on loss $\sqrt{L}$ for small η has two phases.

Phase 2:

For $\Theta(1/\eta^2)$ steps,

1. $\sqrt{L}(x(t)) + \sqrt{L}(x(t+1))$
$$= \eta \lambda_{\max}(\nabla^2 L(x(t))) + \mathcal{O}(\eta^2)$$

2. $\lambda_{\max}(\nabla^2 L)$ decreases at a rate $\Theta(\eta^2)$.

$\sqrt{L}$ oscillates in consecutive steps, implying EoS

This paper (* setting 1): GD on $\sqrt{L}$

($\min_x L(x) = 0$, with smooth L)

Note: $\lambda_{max}(\nabla^2\sqrt{L})$ diverges when $L \to 0$ and $\nabla^2 L$ has rank at least 2.

Theorem: GD on loss $\sqrt{L}$ for small η has two phases.

Phase 2:

For $\Theta(1/\eta^2)$ steps,

1. $\sqrt{L}(x(t)) + \sqrt{L}(x(t+1))$
$$= \eta\lambda_{\max}(\nabla^2 L(x(t))) + \mathcal{O}(\eta^2)$$

2. $\lambda_{\max}(\nabla^2 L)$ decreases at a rate $\Theta(\eta^2)$.

$\sqrt{L}$ oscillates in consecutive steps, implying EoS

$\sqrt{L}$ decreases over time overall.

This paper (* setting 1): GD on $\sqrt{L}$

$(\min_x L(x) = 0$, **with smooth** $L)$

Note: $\lambda_{max}(\nabla^2\sqrt{L})$ diverges when $L \to 0$ and $\nabla^2 L$ has rank at least 2.

Theorem: GD on loss $\sqrt{L}$ for small η has two phases.

Phase 2:

For $\Theta(1/\eta^2)$ steps,

1. $\sqrt{L}(x(t)) + \sqrt{L}(x(t+1))$

$$= \eta\lambda_{\max}(\nabla^2 L(x(t))) + \mathcal{O}(\eta^2)$$

2. $\lambda_{\max}(\nabla^2 L)$ decreases at a rate $\Theta(\eta^2)$.

This paper (* setting 1): GD on $\sqrt{L}$

($\min_x L(x) = 0$, with smooth L)

Note: $\lambda_{max}(\nabla^2\sqrt{L})$ diverges when $L \to 0$ and $\nabla^2 L$ has rank at least 2.

Theorem: GD on loss $\sqrt{L}$ for small η has two phases.

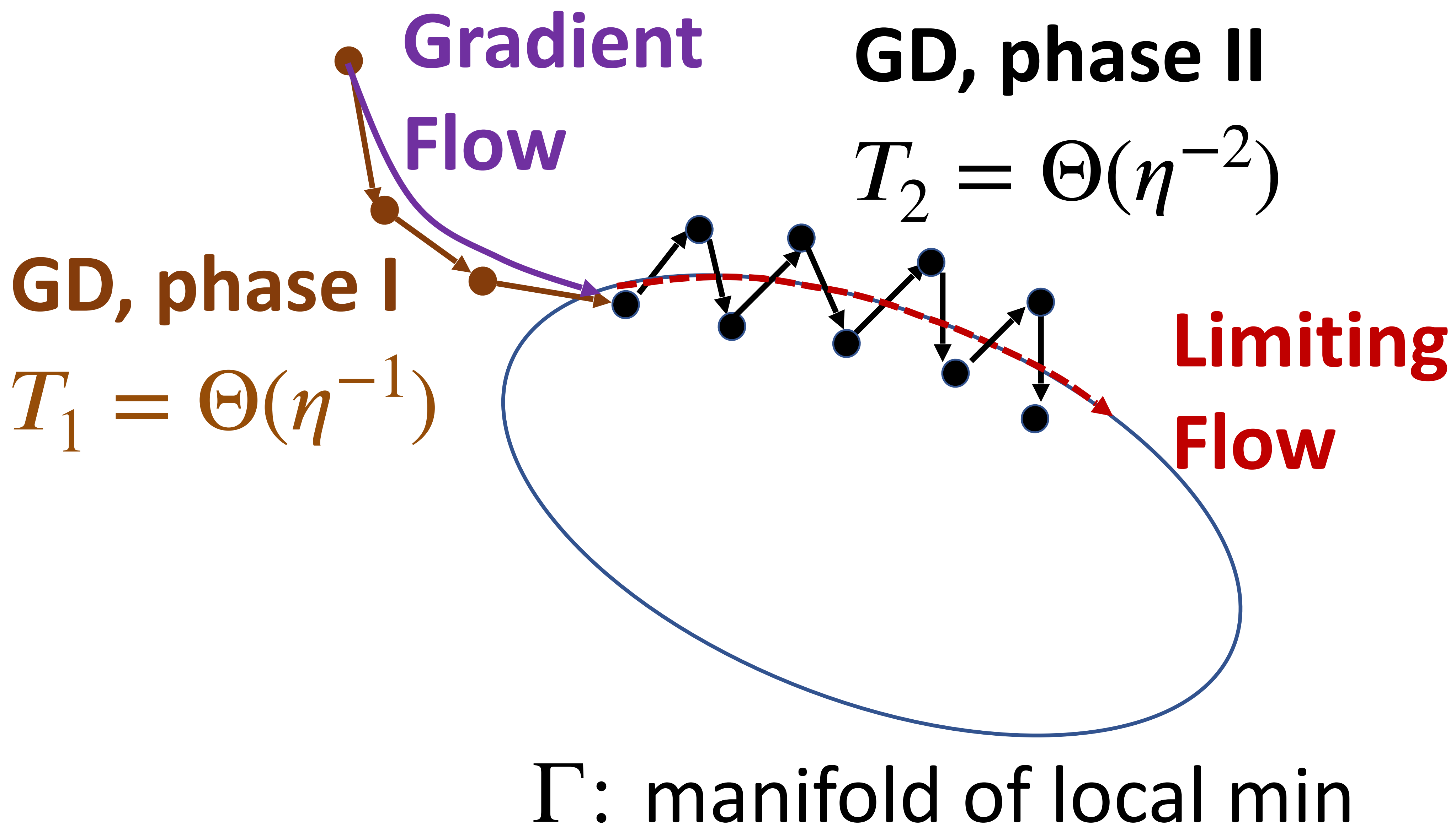
Phase 2:

For $\Theta(1/\eta^2)$ steps,

1. $\sqrt{L}(x(t)) + \sqrt{L}(x(t+1))$

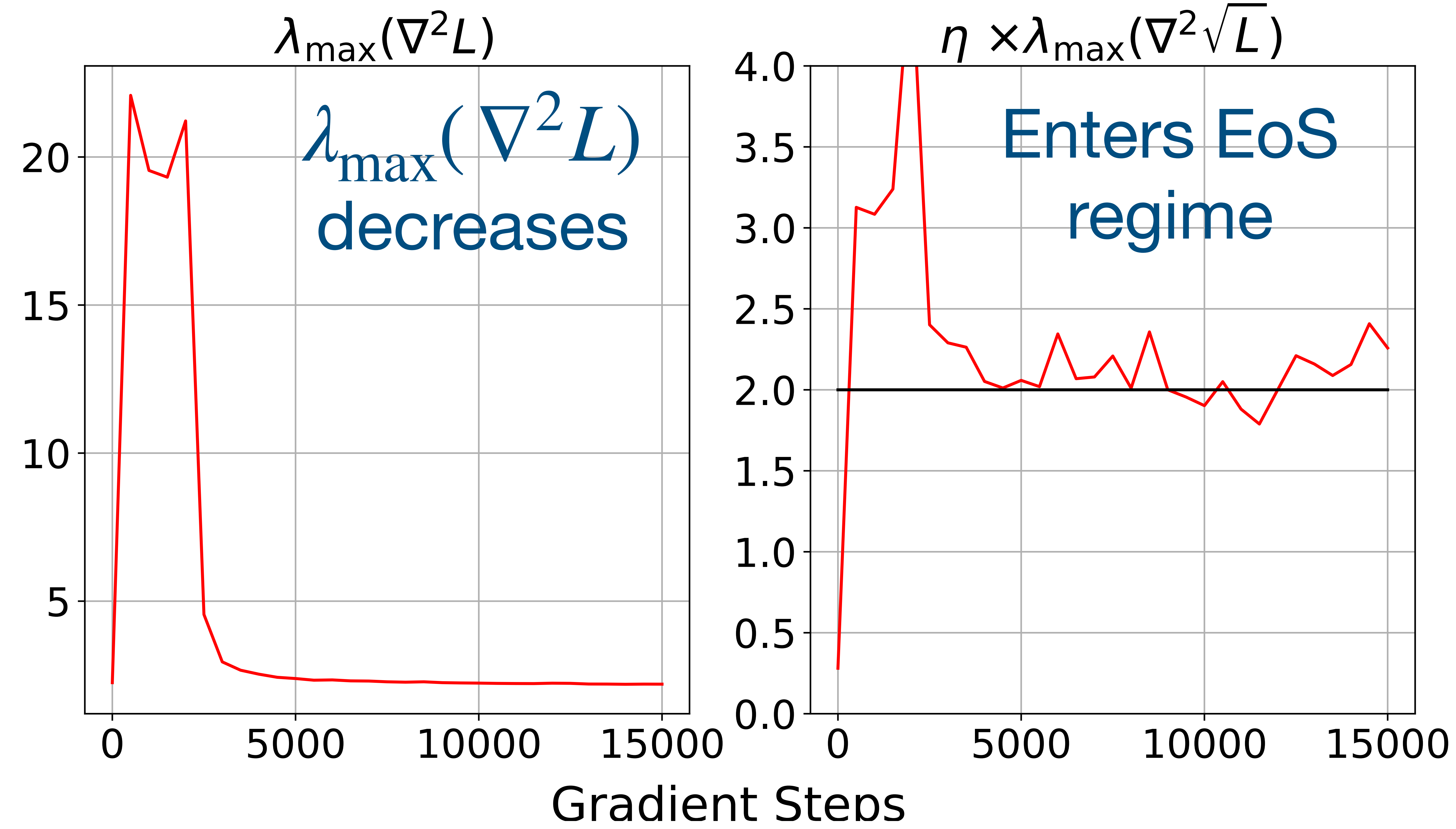
$$= \eta\lambda_{\max}(\nabla^2 L(x(t))) + \mathcal{O}(\eta^2)$$

2. $\lambda_{\max}(\nabla^2 L)$ decreases at a rate $\Theta(\eta^2)$.



"Implicit bias for Sharpness Minimization": $\lambda_{max}(\nabla^2 L)$ decreases over time.

Experiments: GD trajectory consistent with theory



VGG-16 on CIFAR-10 dataset with Mean Square Loss

Future work

Future work

Future work

- Analyse EoS for loss L? (Hurdle: Real-life losses have complicated mathematical structure.)

Future work

- Analyse EoS for loss L? (Hurdle: Real-life losses have complicated mathematical structure.)

- Analyse Edge of Stability far from manifold of zero loss. (Our analysis only applies close to manifold.)

Future work

- Analyse EoS for loss L? (Hurdle: Real-life losses have complicated mathematical structure.)

- Analyse Edge of Stability far from manifold of zero loss. (Our analysis only applies close to manifold.)

- Explore EoS in SGD setting.