

A Minimalist Example of Edge-of-Stability and Progressive Sharpening

Zixuan Zhang

Georgia Tech ISyE

Sep 2025

Joint Work with



Liming Liu
Georgia Tech



Simon Shaolei Du
Univ. of Washington

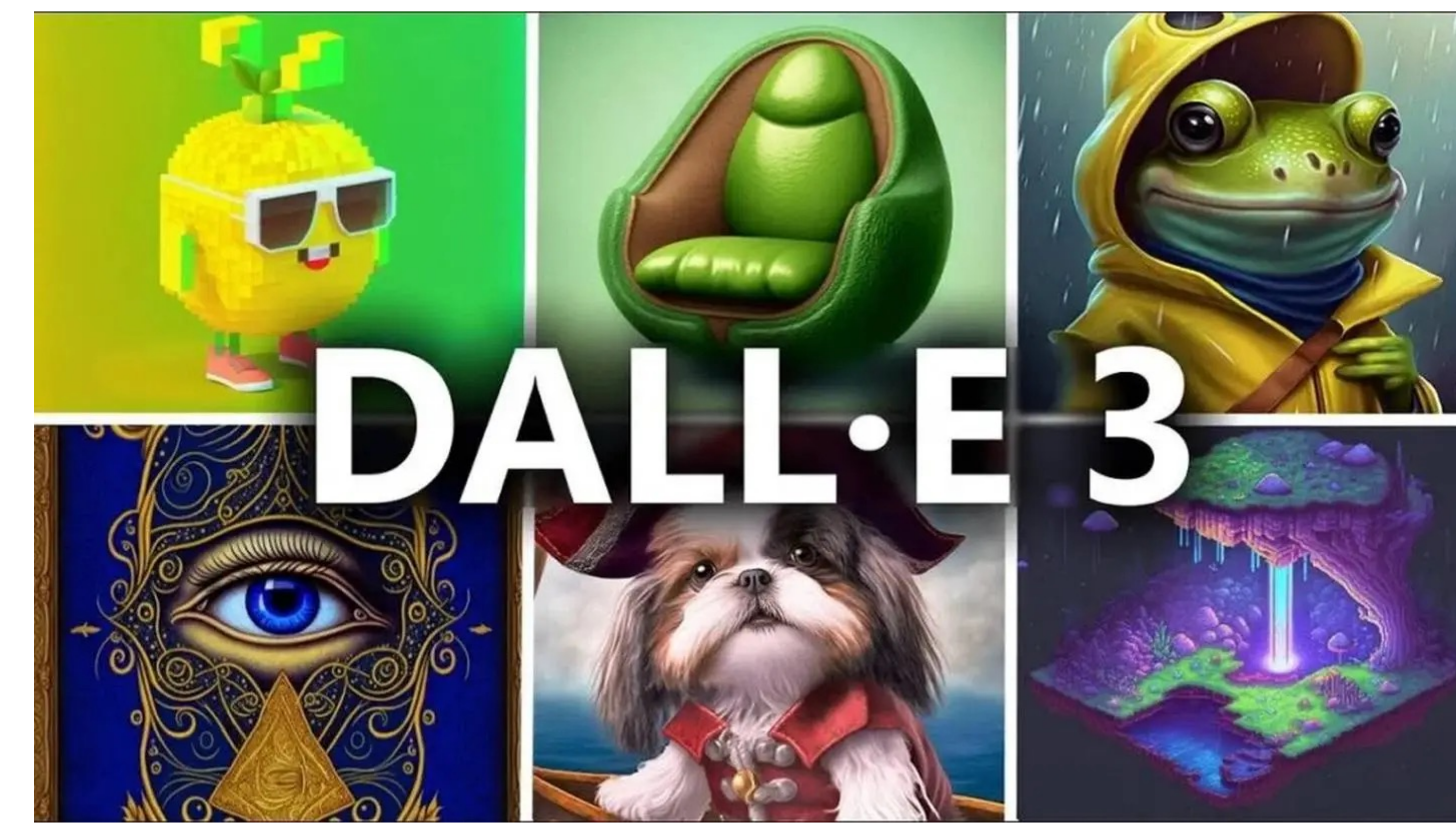


Tuo Zhao
Georgia Tech

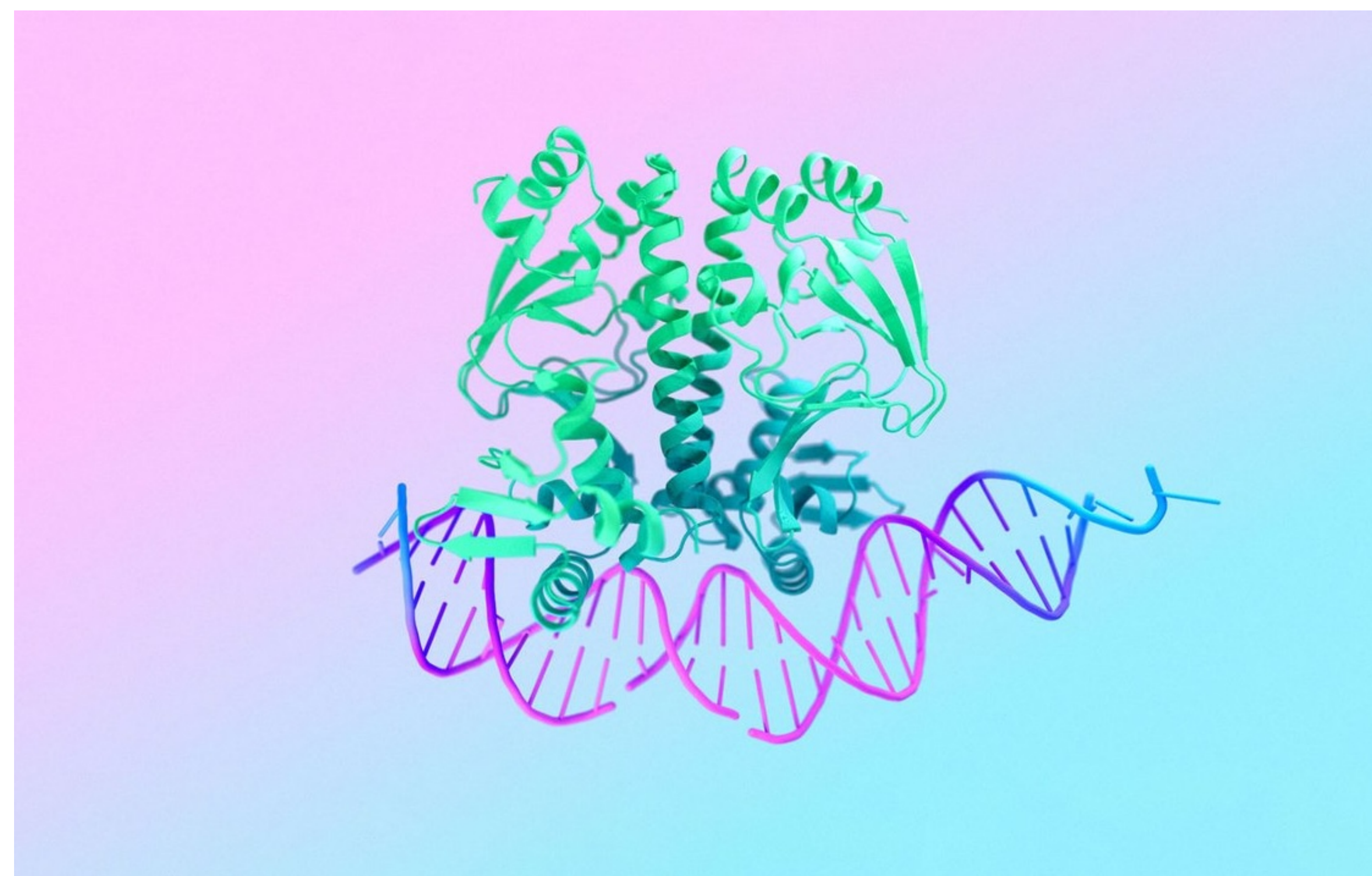
Practical Success of Deep Learning



ChatGPT by OpenAI
NLP



DALL-E 3 by OpenAI
Vision

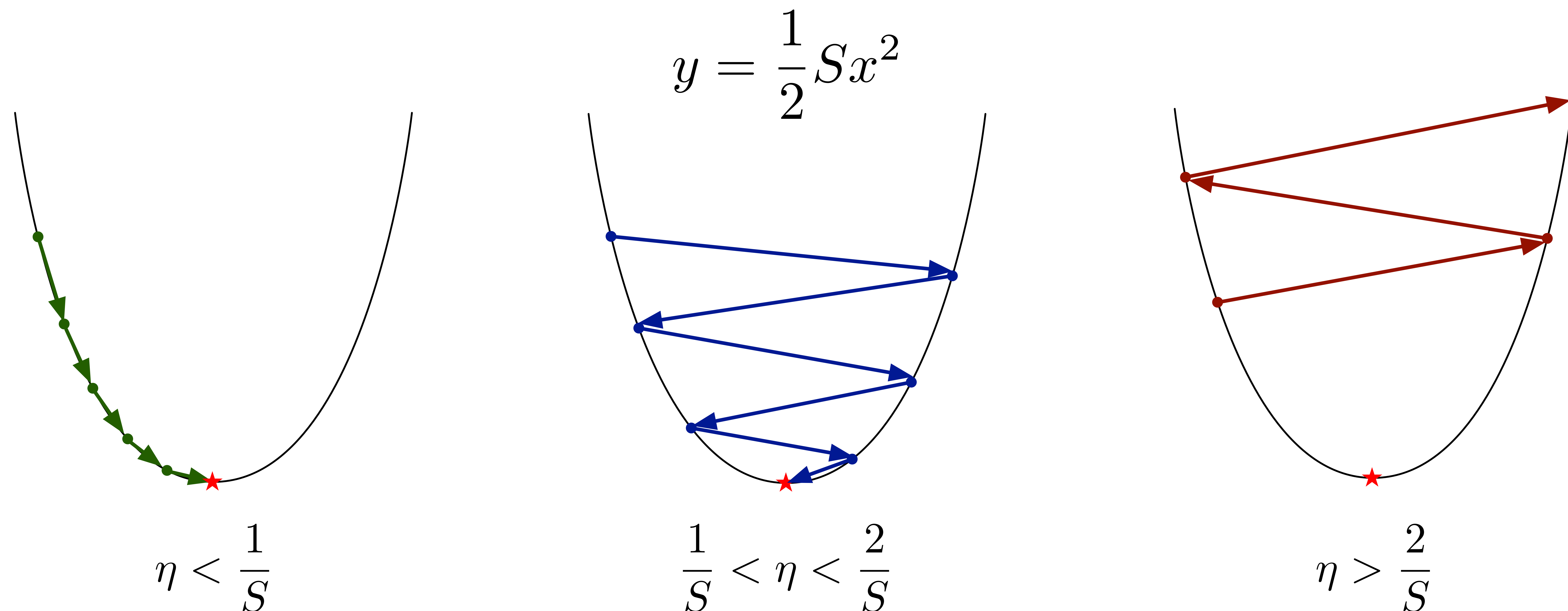


AlphaFold by DeepMind
Science



FDS Visualization by Tesla
Robotics

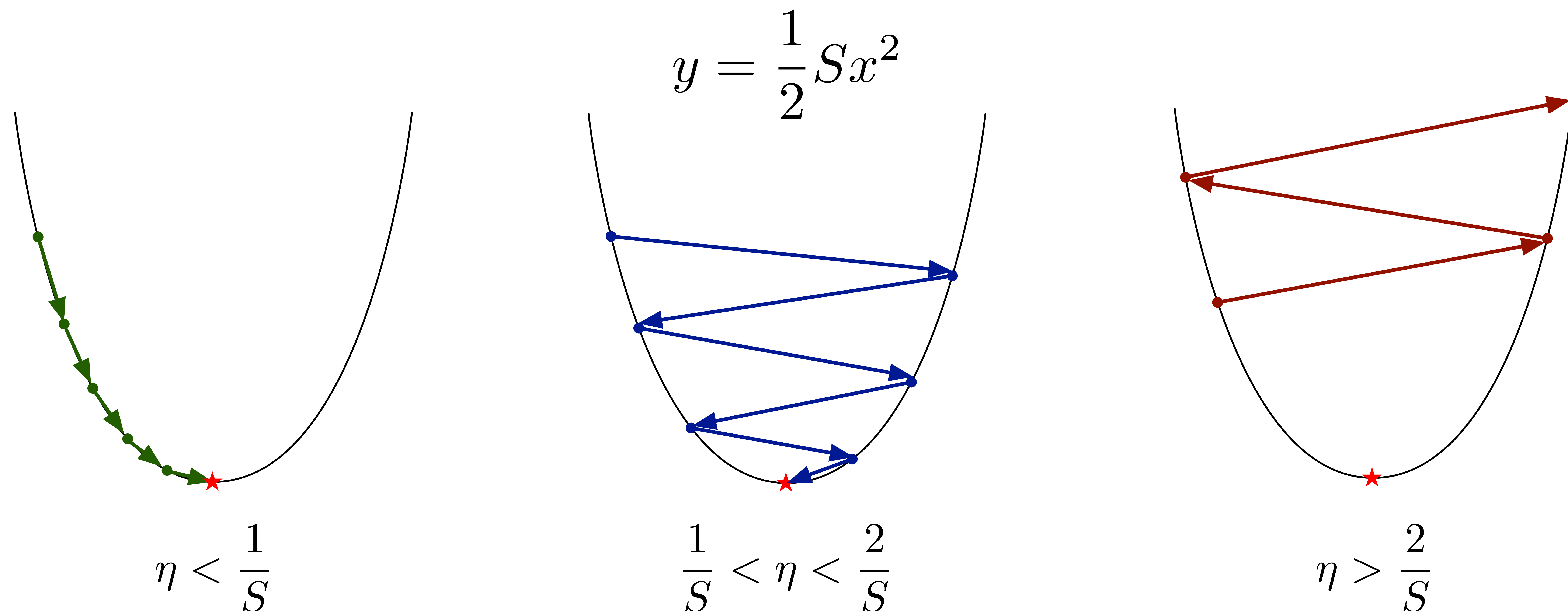
Classical Gradient Descent Analysis



Lemma (Descent Lemma)

If the GD stepsize $\eta < 2/S$, the loss L decreases at each GD iteration, where $S = \lambda_{\max}(\nabla^2 L)$.

Classical Gradient Descent Analysis

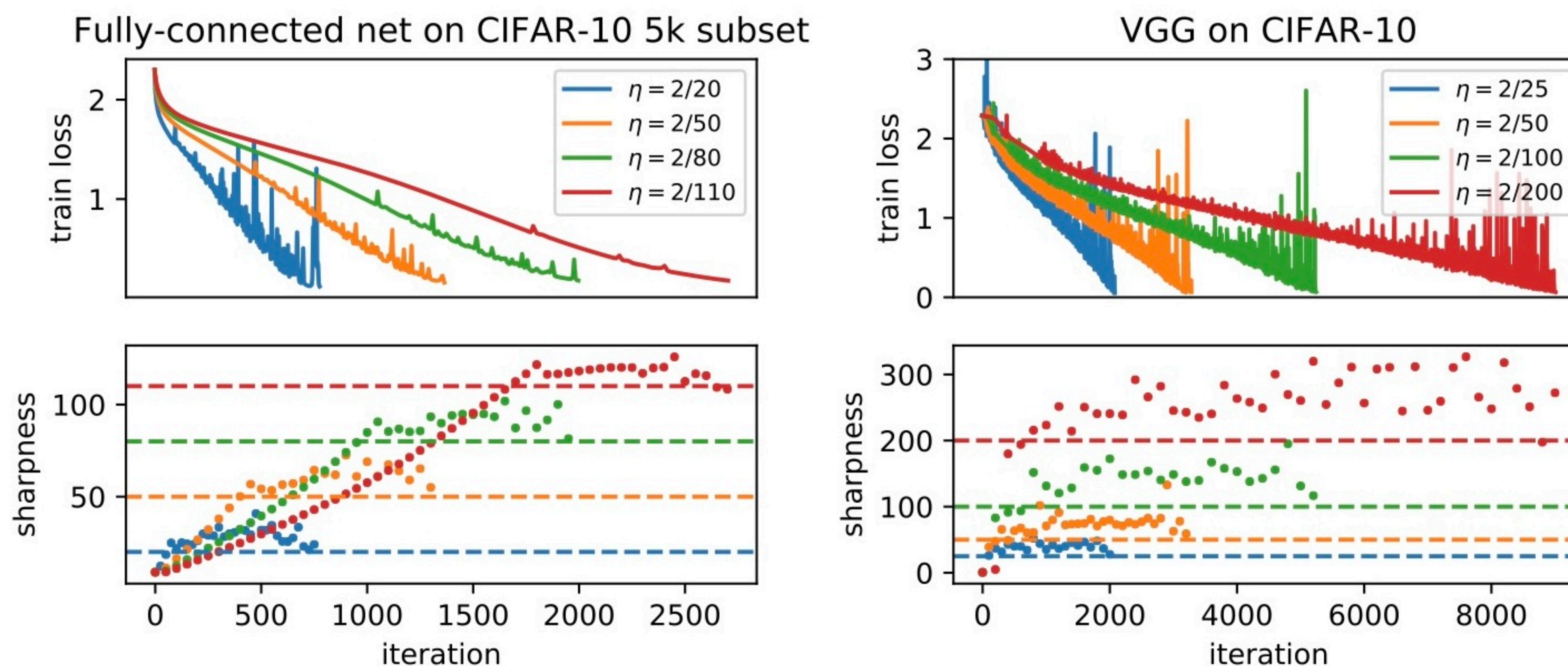


Lemma (Descent Lemma)

If the GD stepsize $\eta < 2/S$, the loss L decreases at each GD iteration, where $S = \lambda_{\max}(\nabla^2 L)$.

Empirical Departures from Descent Lemma

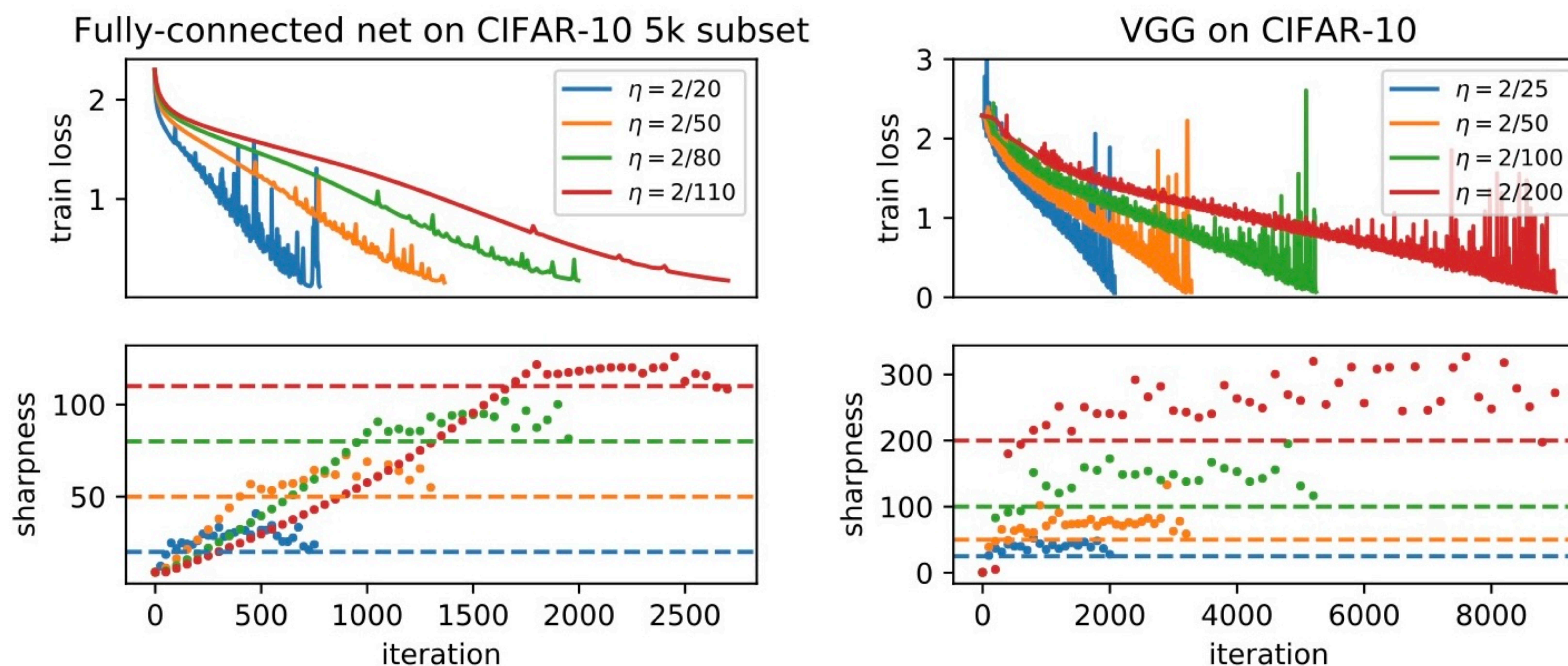
Neural network training with GD violates the descent lemma
[Cohen et al. ICLR 2021].



- The sharpness $S = \lambda_{\max}(\nabla^2 L)$ increases until it reaches $2/\eta$.
- Loss decreases overall, even if $S > 2/\eta$.

Empirical Departures from Descent Lemma

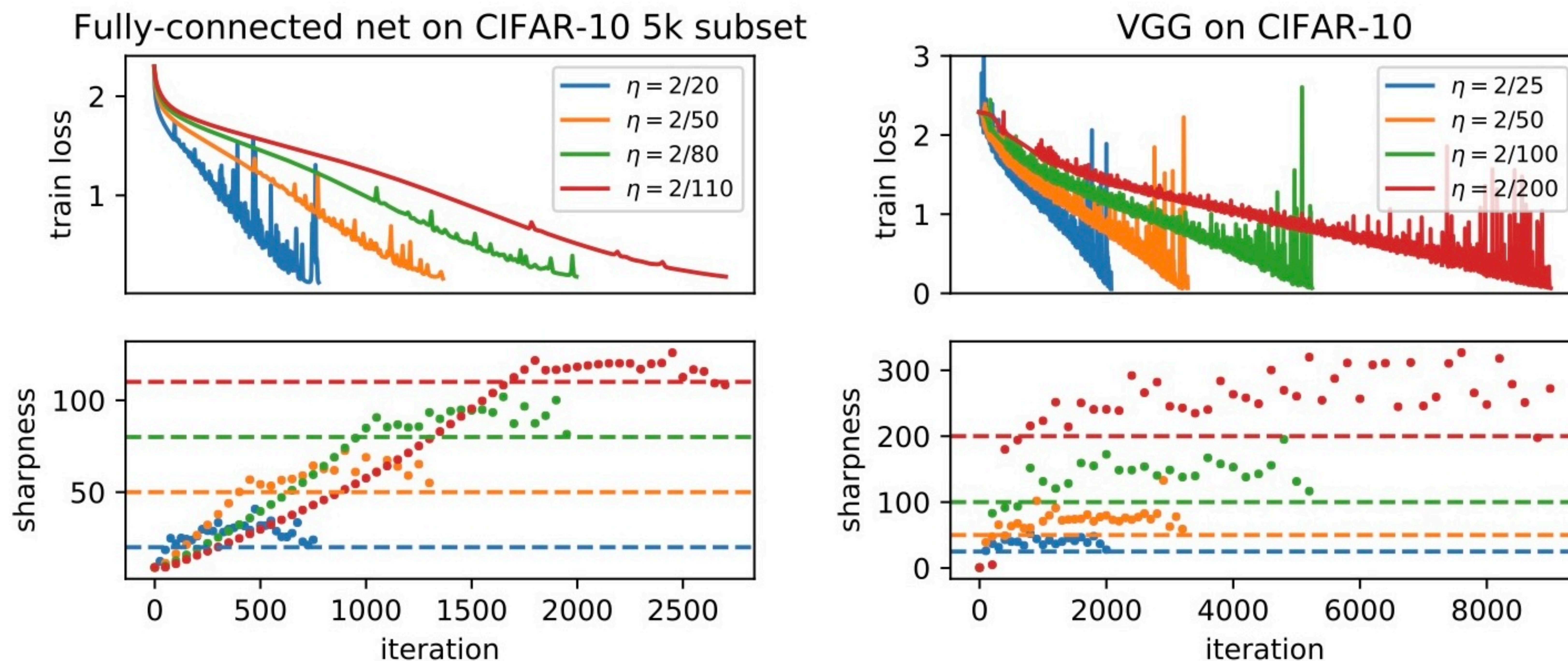
Neural network training with GD violates the descent lemma
[Cohen et al. ICLR 2021].



- The sharpness $S = \lambda_{\max}(\nabla^2 L)$ increases until it reaches $2/\eta$.
- Loss decreases overall, even if $S > 2/\eta$.

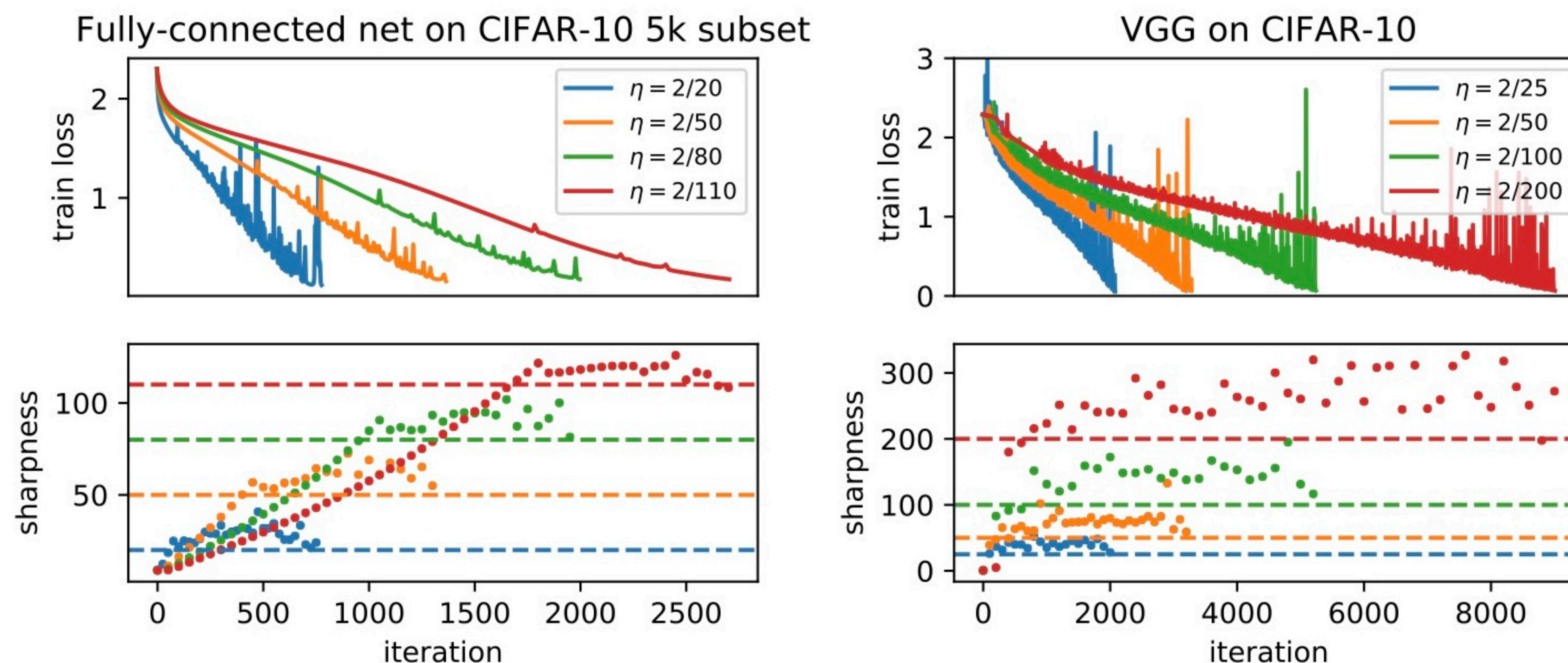
Empirical Departures from Descent Lemma

Neural network training with GD violates the descent lemma
[Cohen et al. ICLR 2021].



- The sharpness $S = \lambda_{\max}(\nabla^2 L)$ increases until it reaches $2/\eta$.
- Loss decreases overall, even if $S > 2/\eta$.

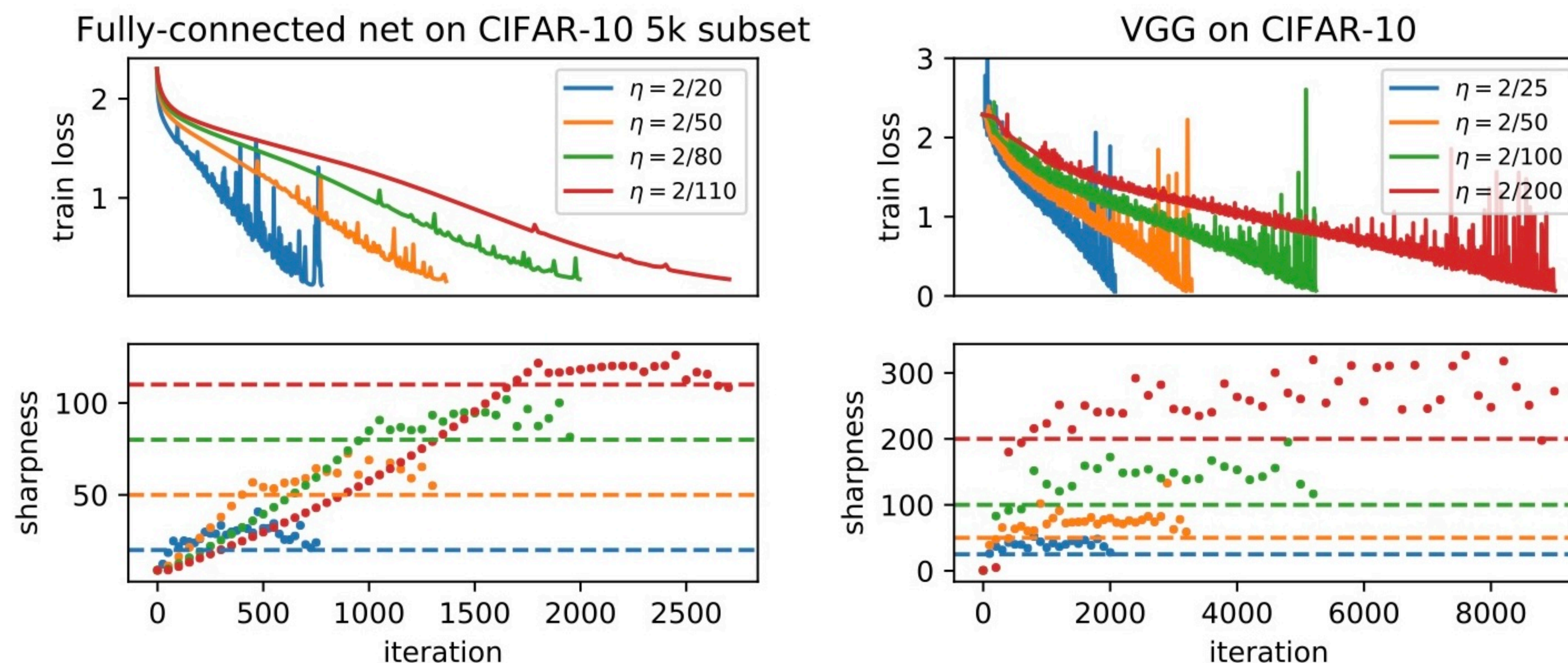
Two Phenomena in GD training



- **Progressive Sharpening (PS):** the sharpness increases until it reaches $2/\eta$.
- **Edge of Stability (EOS):** the sharpness hovers at $2/\eta$, but the loss decreases non-monotonically.

Why should we care about EoS? – Guiding learning-rate schedules, implicit regularization & generalization quality...

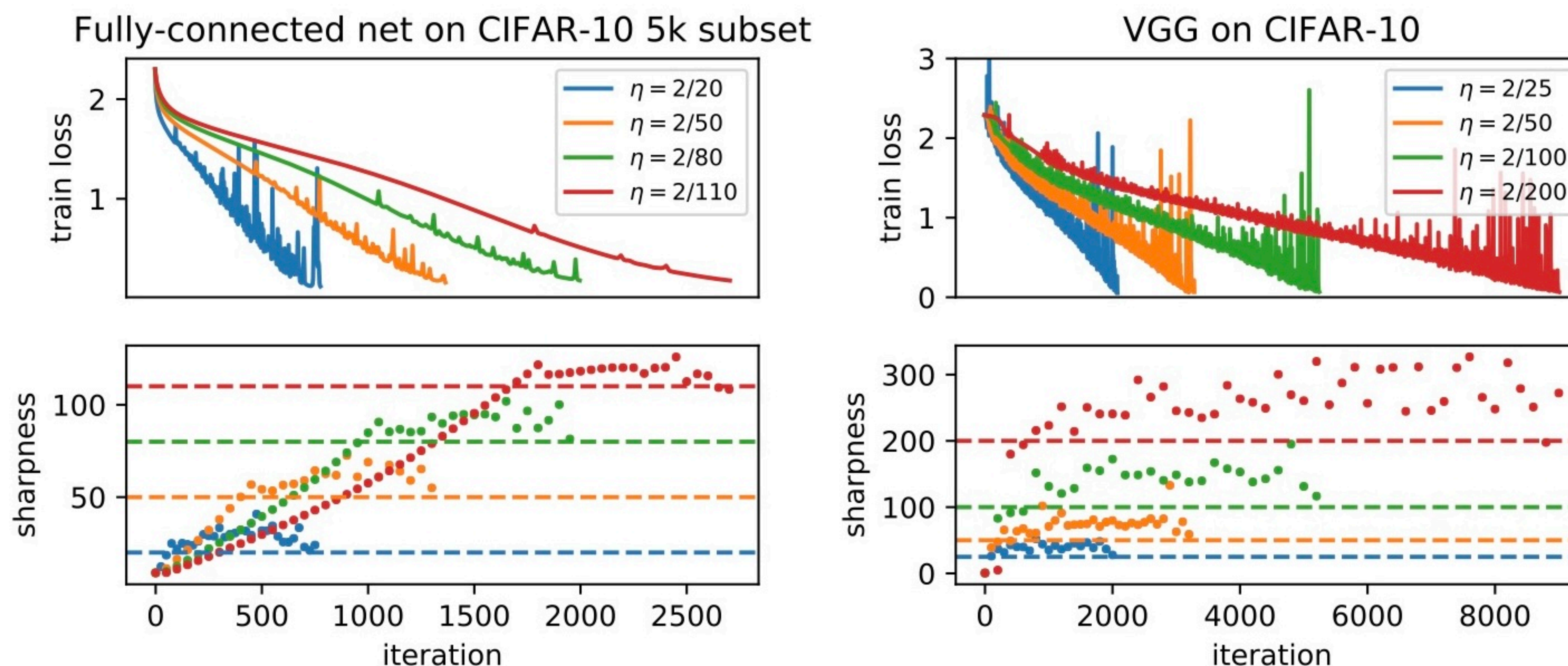
Two Phenomena in GD training



- **Progressive Sharpening (PS):** the sharpness increases until it reaches $2/\eta$.
- **Edge of Stability (EOS):** the sharpness hovers at $2/\eta$, but the loss decreases non-monotonically.

Why should we care about EoS? – Guiding learning-rate schedules, implicit regularization & generalization quality...

Two Phenomena in GD training



- **Progressive Sharpening (PS)**: the sharpness increases until it reaches $2/\eta$.
- **Edge of Stability (EOS)**: the sharpness hovers at $2/\eta$, but the loss decreases non-monotonically.

Why should we care about EoS? – Guiding learning-rate schedules, implicit regularization & generalization quality...

Two Lines of Current Research

1 Generalist analysis:

- Li et al. (2022); Damian et al. (2022)...
- Generic, covering wide settings.
- Limitations: “hard-to-verify” assumptions, such as the existence of progressive sharpening.

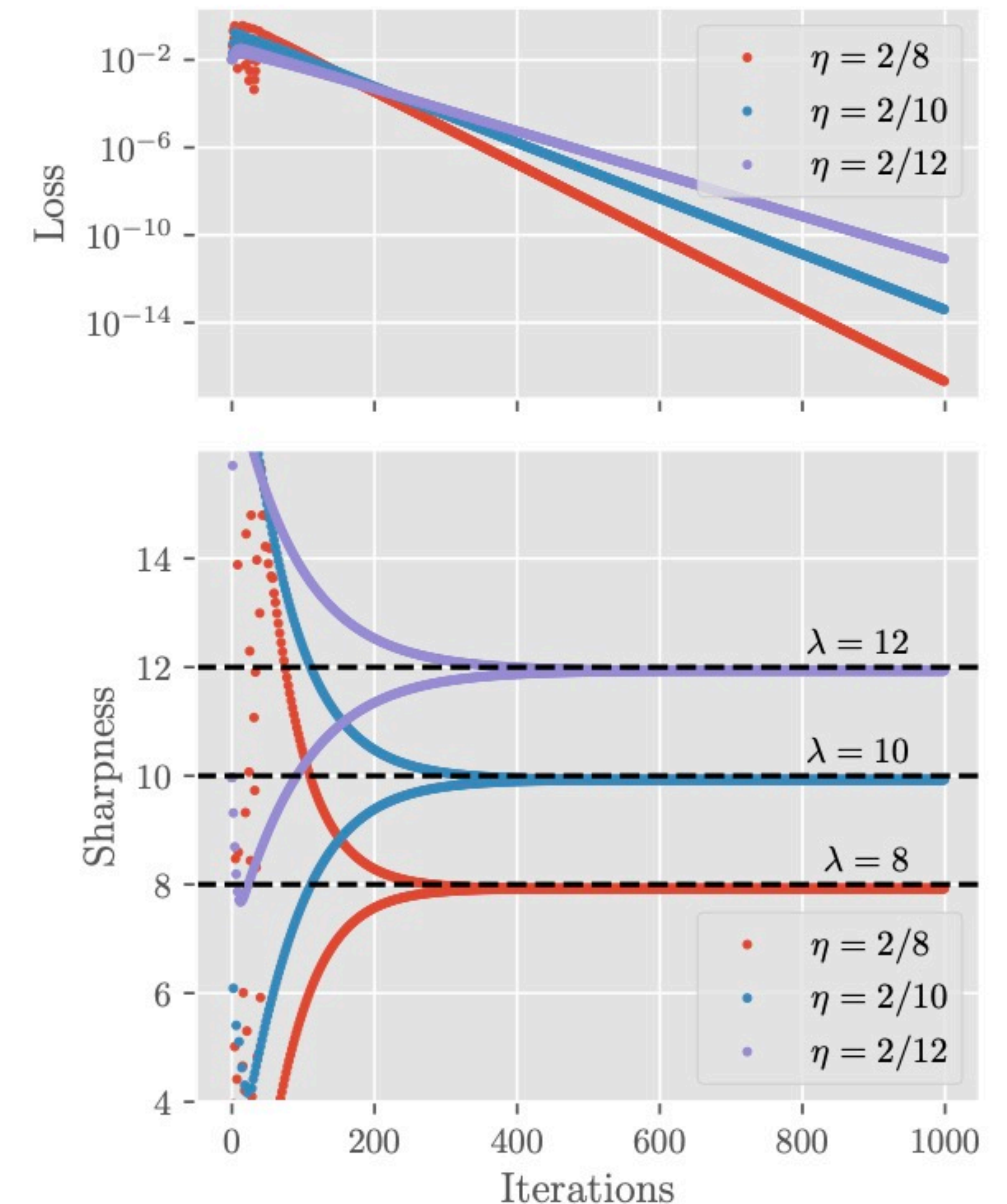
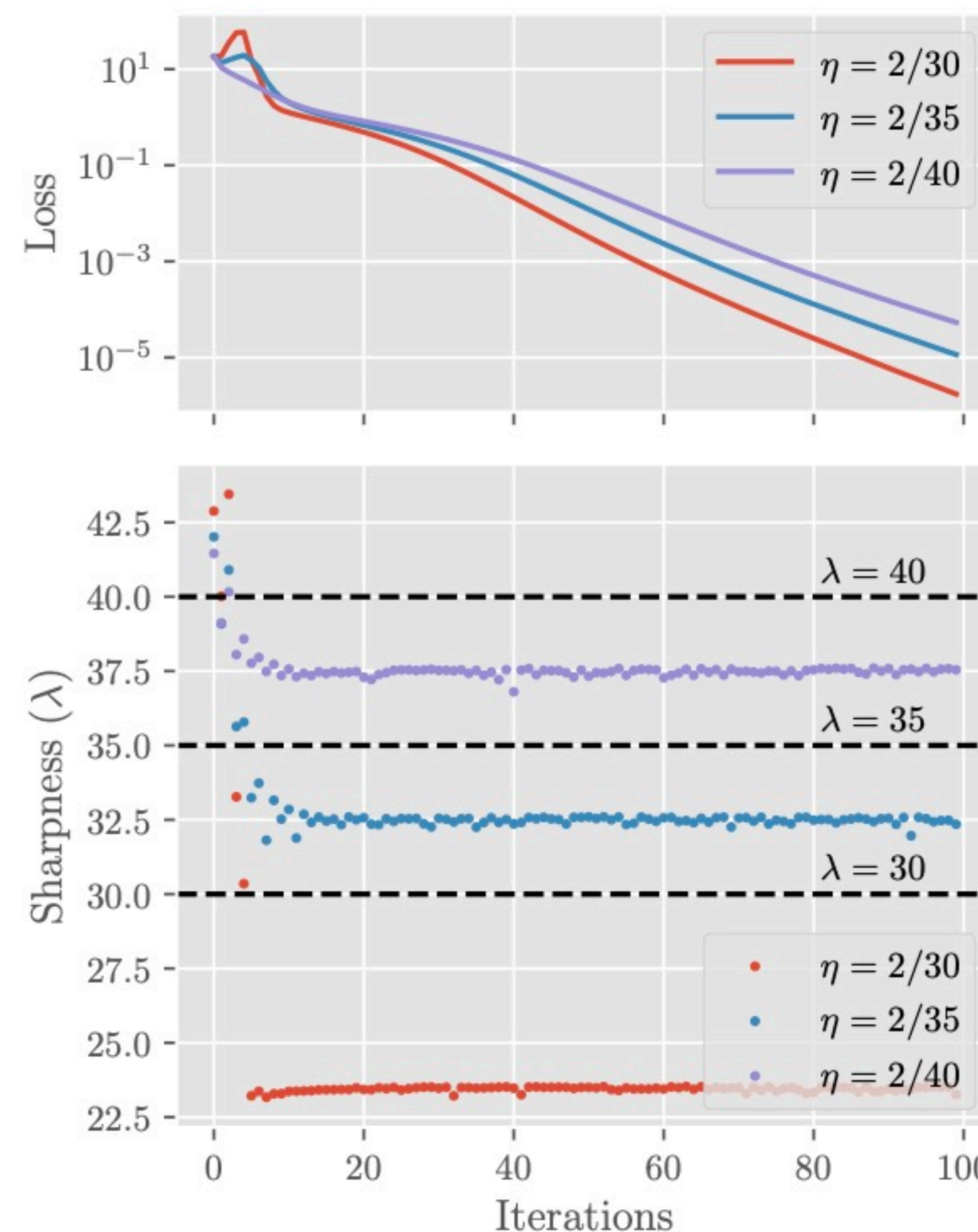
2 Minimalist examples:

- Zhu et al. (2022); Wang et al. (2023); Kreisler et al. (2023)...
- Concrete, verifiable insights.
- Limitations: only asymptotic results, over-simplified setting of scalar networks, cannot explain input effects.

Two Lines of Current

1 Generalist analysis:

- Li et al. (2022);
- Generic, covering
- Limitations: “ha
existence of prog



2 Minimalist examples:

- Zhu et al. (2022); Wang et al. (2023); Kreisler et al. (2023)...
- Concrete, verifiable insights.
- Limitations: only asymptotic results, over-simplified setting of scalar networks, cannot explain input effects.

Two Lines of Current Research

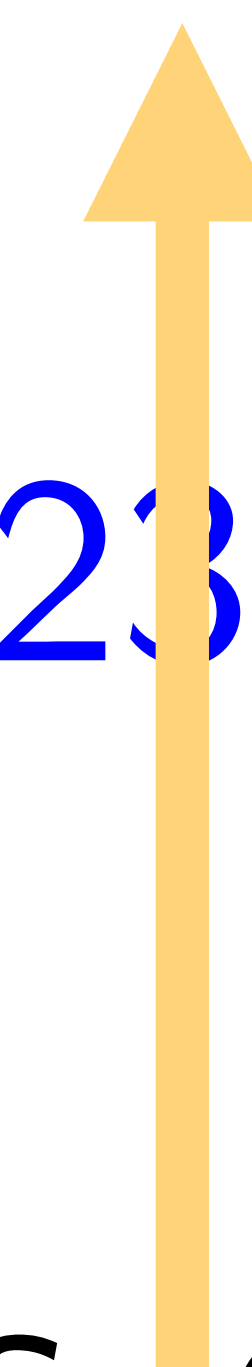


Rosenfeld et. al. 2023



2 Minimalist examples:

- Zhu et al. (2022); Wang et al. (2023); Kreisler et al. (2023)...
- Concrete, verifiable insights.
- Limitations: only asymptotic results, over-simplified setting of scalar networks, cannot explain **input effects**.



What's missing?

Can we replicate PS & EoS with a two-feature network?

- (i) What drives the non-monotonic but overall decreasing loss?
- (ii) How do different input features affect this behavior?

What's missing?

Can we replicate PS & EoS with a two-feature network?

- (i) What drives the non-monotonic but overall decreasing loss?
- (ii) How do different input features affect this behavior?

Our Contributions

We study the GD dynamics of a two layer, width-one neural network with a 2-D input of different relevance.

- **Non-asymptotic analysis:** Entire GD trajectory characterization, proving the existence of PS and EoS.
- **Feature-aware insights:** How relevant vs. irrelevant features drive dynamics.

Our Contributions

We study the GD dynamics of a two layer, width-one neural network with a 2-D input of different relevance.

- **Non-asymptotic analysis:** Entire GD trajectory characterization, proving the existence of PS and EoS.
- **Feature-aware insights:** How relevant vs. irrelevant features drive dynamics.

Our Contributions

We study the GD dynamics of a two layer, width-one neural network with a 2-D input of different relevance.

- **Non-asymptotic analysis:** Entire GD trajectory characterization, proving the existence of PS and EoS.
- **Feature-aware insights:** How relevant vs. irrelevant features drive dynamics.

Problem Setup

- Input $x = (x_1, x_2)^\top \in \mathbb{R}^2$. Output $y = f^*(x) \in \mathbb{R}$.

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N \left(0, \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \right), \quad \text{and } f^*(x) = x_2.$$

Here we suppose $\lambda_1 \gg \lambda_2$.

- To learn the target function f^* , we use the network with weights $\theta = (\alpha, \beta_1, \beta_2)^T \in \mathbb{R}^3$:

$$f(x; \theta) = \alpha\beta_1x_1 + \alpha\beta_2x_2.$$

- Consider the square loss

$$L(\theta) = \frac{1}{2} \mathbb{E}_{x,y} \left[(y - f(x; \theta))^2 \right].$$

Problem Setup

- Input $x = (x_1, x_2)^\top \in \mathbb{R}^2$. Output $y = f^*(x) \in \mathbb{R}$.

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N \left(0, \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \right), \quad \text{and } f^*(x) = x_2.$$

Here we suppose $\lambda_1 \gg \lambda_2$.

- To learn the target function f^* , we use the network with weights $\theta = (\alpha, \beta_1, \beta_2)^T \in \mathbb{R}^3$:

$$f(x; \theta) = \alpha\beta_1x_1 + \alpha\beta_2x_2.$$

- Consider the square loss

$$L(\theta) = \frac{1}{2} \mathbb{E}_{x,y} \left[(y - f(x; \theta))^2 \right].$$

Problem Setup

- Input $x = (x_1, x_2)^\top \in \mathbb{R}^2$. Output $y = f^*(x) \in \mathbb{R}$.

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N \left(0, \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \right), \quad \text{and } f^*(x) = x_2.$$

Here we suppose $\lambda_1 \gg \lambda_2$.

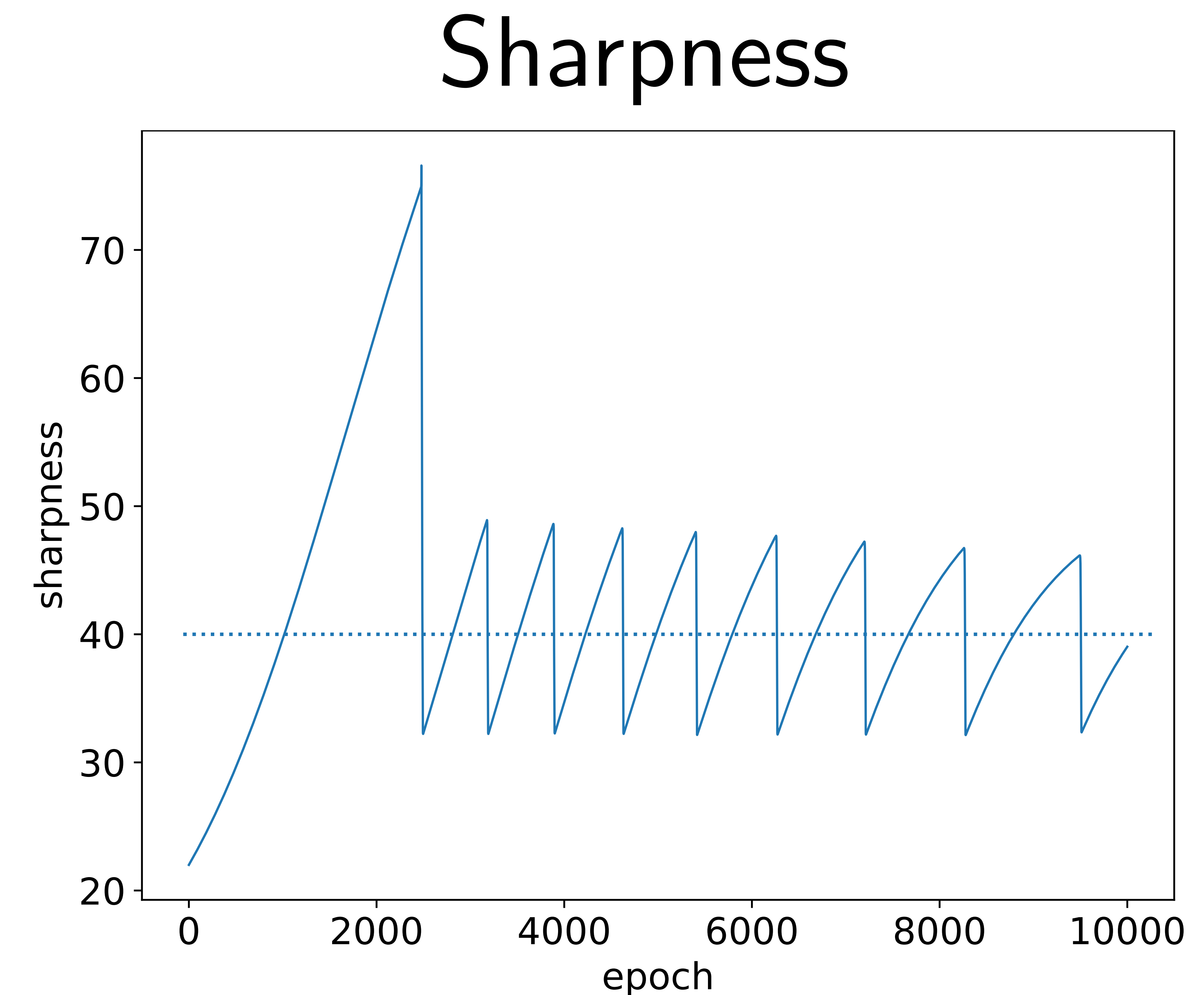
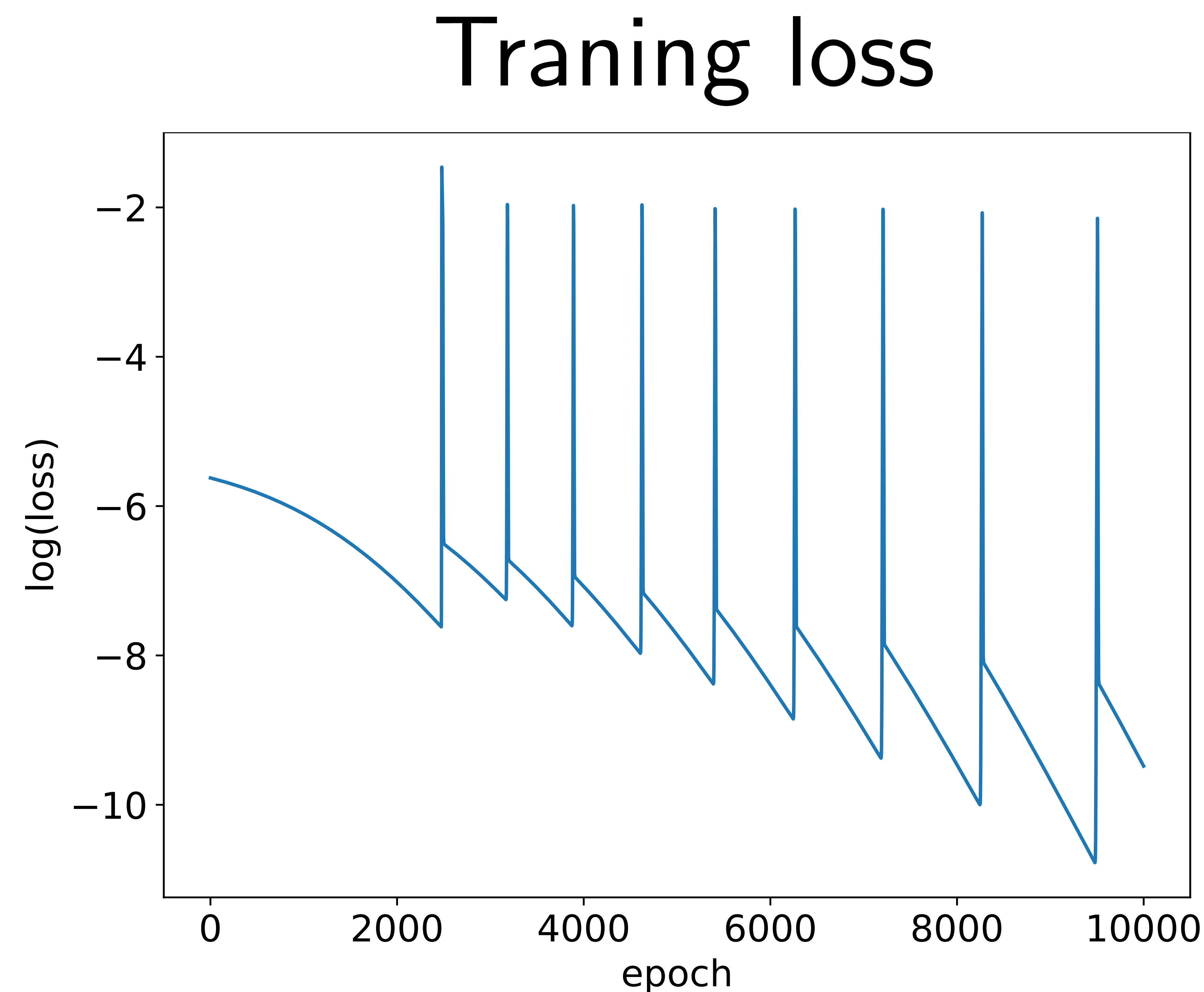
- To learn the target function f^* , we use the network with weights $\theta = (\alpha, \beta_1, \beta_2)^T \in \mathbb{R}^3$:

$$f(x; \theta) = \alpha\beta_1x_1 + \alpha\beta_2x_2.$$

- Consider the square loss

$$L(\theta) = \frac{1}{2} \mathbb{E}_{x,y} \left[(y - f(x; \theta))^2 \right].$$

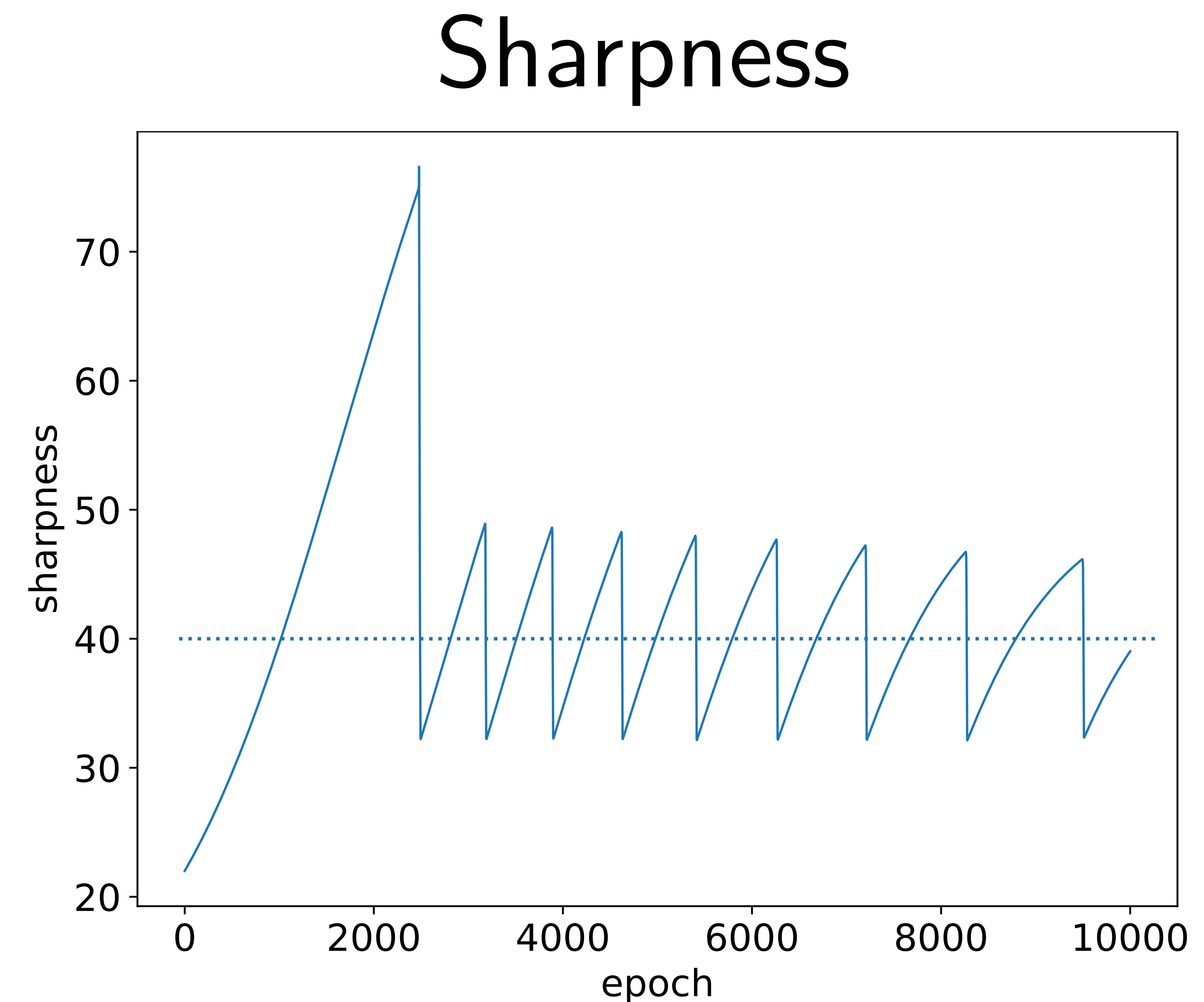
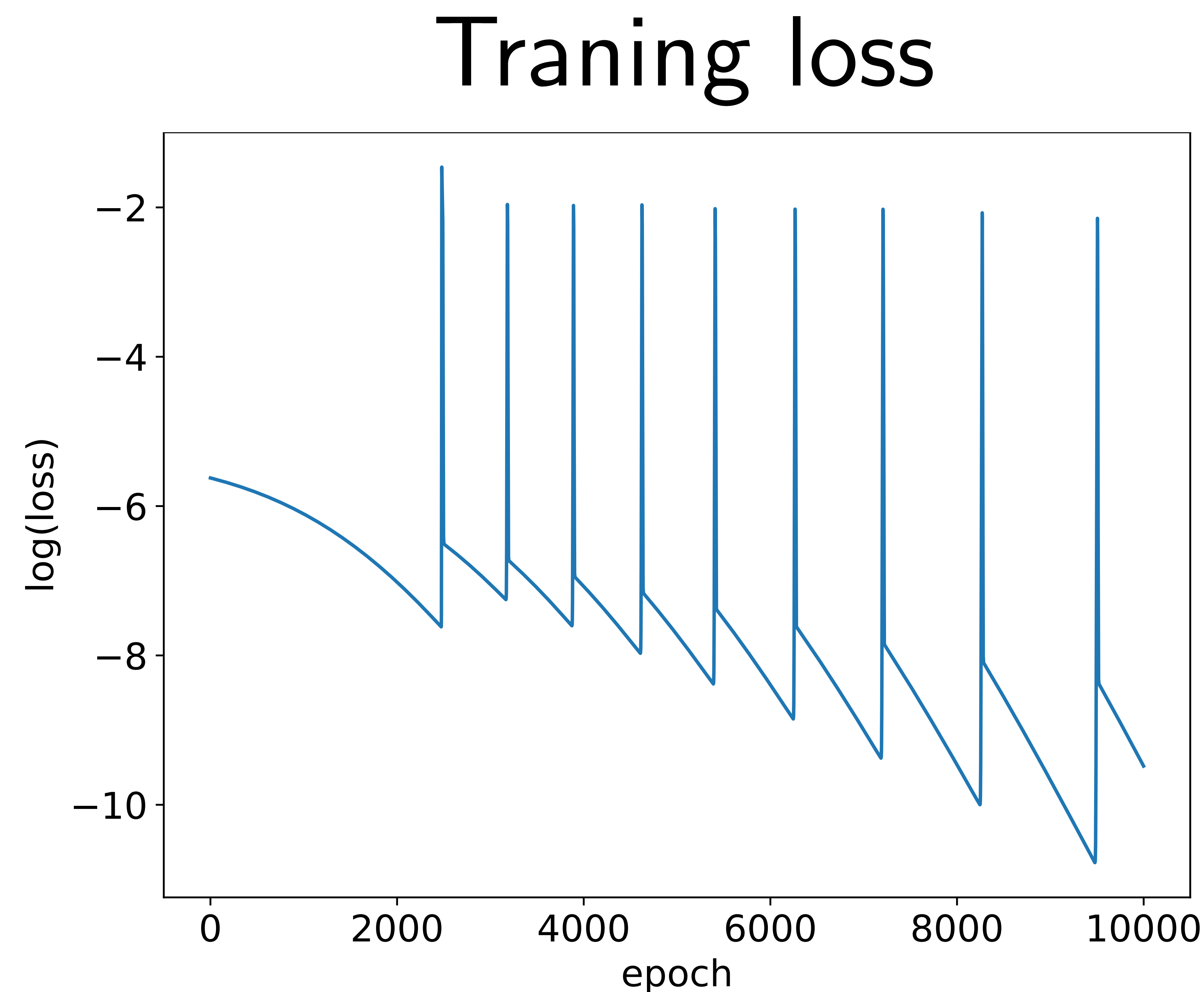
Experimental Observation



Three phases identified in our GD dynamics:

- Phase I: **Progressive sharpening** before EoS.
- Phase II.1: **Progressive sharpening** during EoS.
- Phase II.2: **Self-stabilization** during EoS.

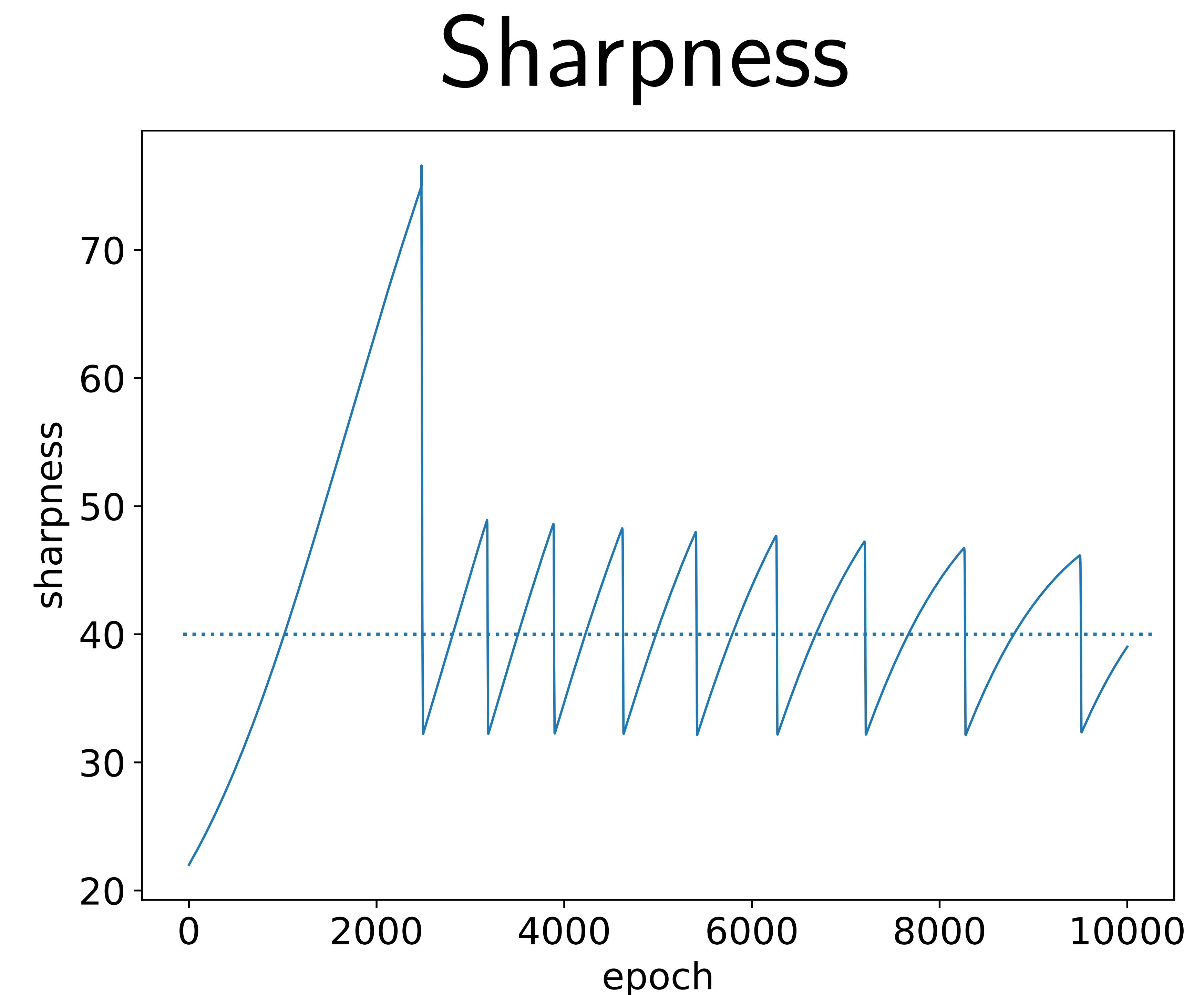
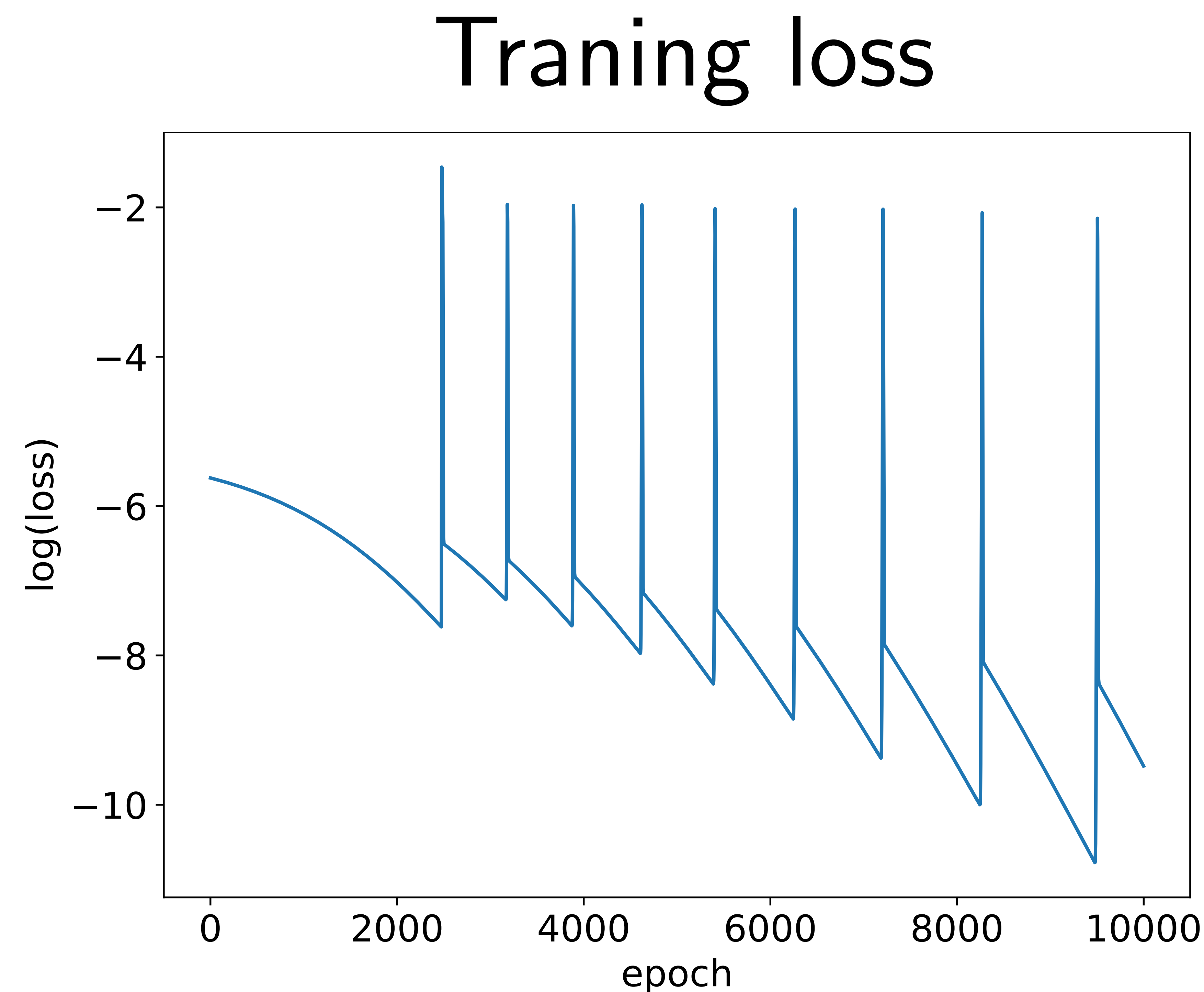
Experimental Observation



Three phases identified in our GD dynamics:

- Phase I: **Progressive sharpening** before EoS.
- Phase II.1: **Progressive sharpening** during EoS.
- Phase II.2: **Self-stabilization** during EoS.

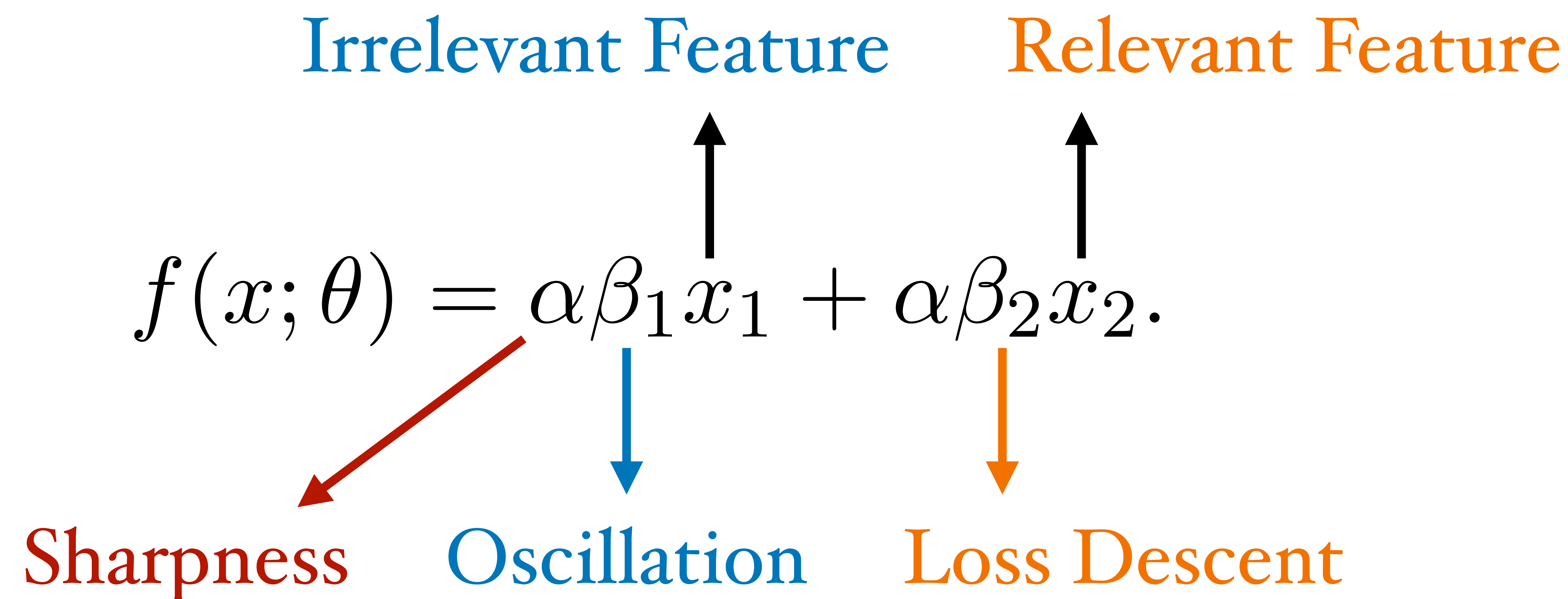
Experimental Observation



Three phases identified in our GD dynamics:

- Phase I: **Progressive sharpening** before EoS.
- Phase II.1: **Progressive sharpening** during EoS.
- Phase II.2: **Self-stabilization** during EoS.

Model Interpretation – Parameter Roles



Lemma

Assume bounds on λ_1, λ_2 and initialization conditions. Along the whole GD trajectory, we have

- 1) **Sharpness dominated by α :** $\lambda_1\alpha^2(t) \leq S(\theta) \leq 1.12\lambda_1\alpha^2(t)$.
- 2) **Instability of β_1 :** The top eigenvector of the loss Hessian aligns with β_1 direction.
- 3) **Monotonic Increase of β_2 :** $\beta_2(t+1) > \beta_2(t)$.

Main Theoretical Results

Theorem

Under some initialization conditions, we have

- 1 Global Convergence** *to flat minima $S(\theta) \leq \frac{2+\delta}{\eta}$.*
- 2 Progressive Sharpening:** $\alpha > 0$ *increases monotonically until $\lambda_1 \alpha^2 \geq 1.5/\eta$. The sharpness will not exceed $4.71/\eta$ along the whole GD trajectory.*
- 3 Edge of Stability:** α *first increases and drives $S(\theta) > 2.37/\eta$, then α decreases, bringing $S(\theta) < 2/\eta + \eta$.*

Main Theoretical Results

Theorem

Under some initialization conditions, we have

- 1 Global Convergence** *to flat minima $S(\theta) \leq \frac{2+\delta}{\eta}$.*
- 2 Progressive Sharpening:** $\alpha > 0$ *increases monotonically until $\lambda_1 \alpha^2 \geq 1.5/\eta$. The sharpness will not exceed $4.71/\eta$ along the whole GD trajectory.*
- 3 Edge of Stability:** α *first increases and drives $S(\theta) > 2.37/\eta$, then α decreases, bringing $S(\theta) < 2/\eta + \eta$.*

Main Theoretical Results

Theorem

Under some initialization conditions, we have

- 1 Global Convergence** *to flat minima $S(\theta) \leq \frac{2+\delta}{\eta}$.*
- 2 Progressive Sharpening:** $\alpha > 0$ *increases monotonically until $\lambda_1 \alpha^2 \geq 1.5/\eta$. The sharpness will not exceed $4.71/\eta$ along the whole GD trajectory.*
- 3 Edge of Stability:** α *first increases and drives $S(\theta) > 2.37/\eta$, then α decreases, bringing $S(\theta) < 2/\eta + \eta$.*

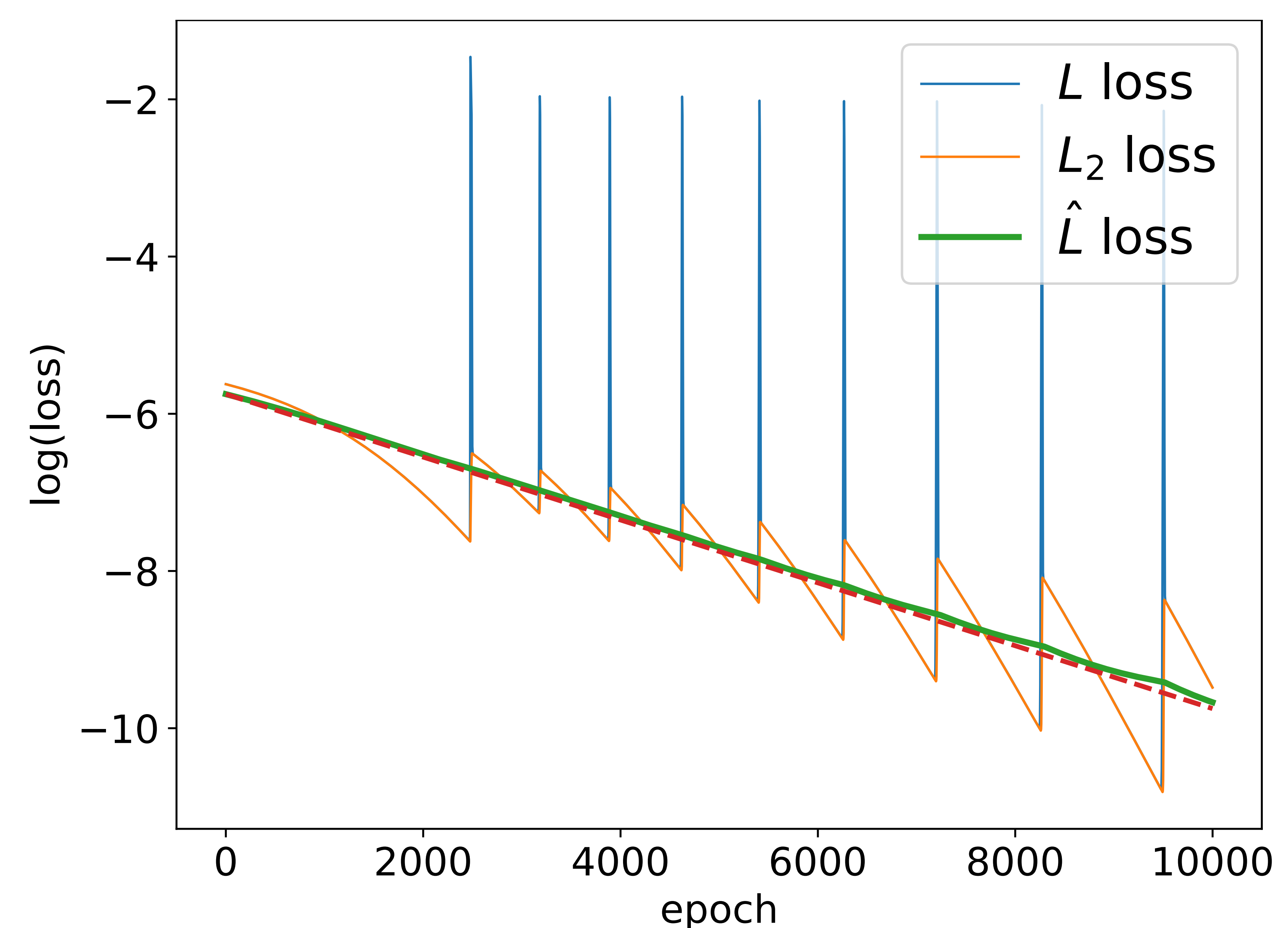
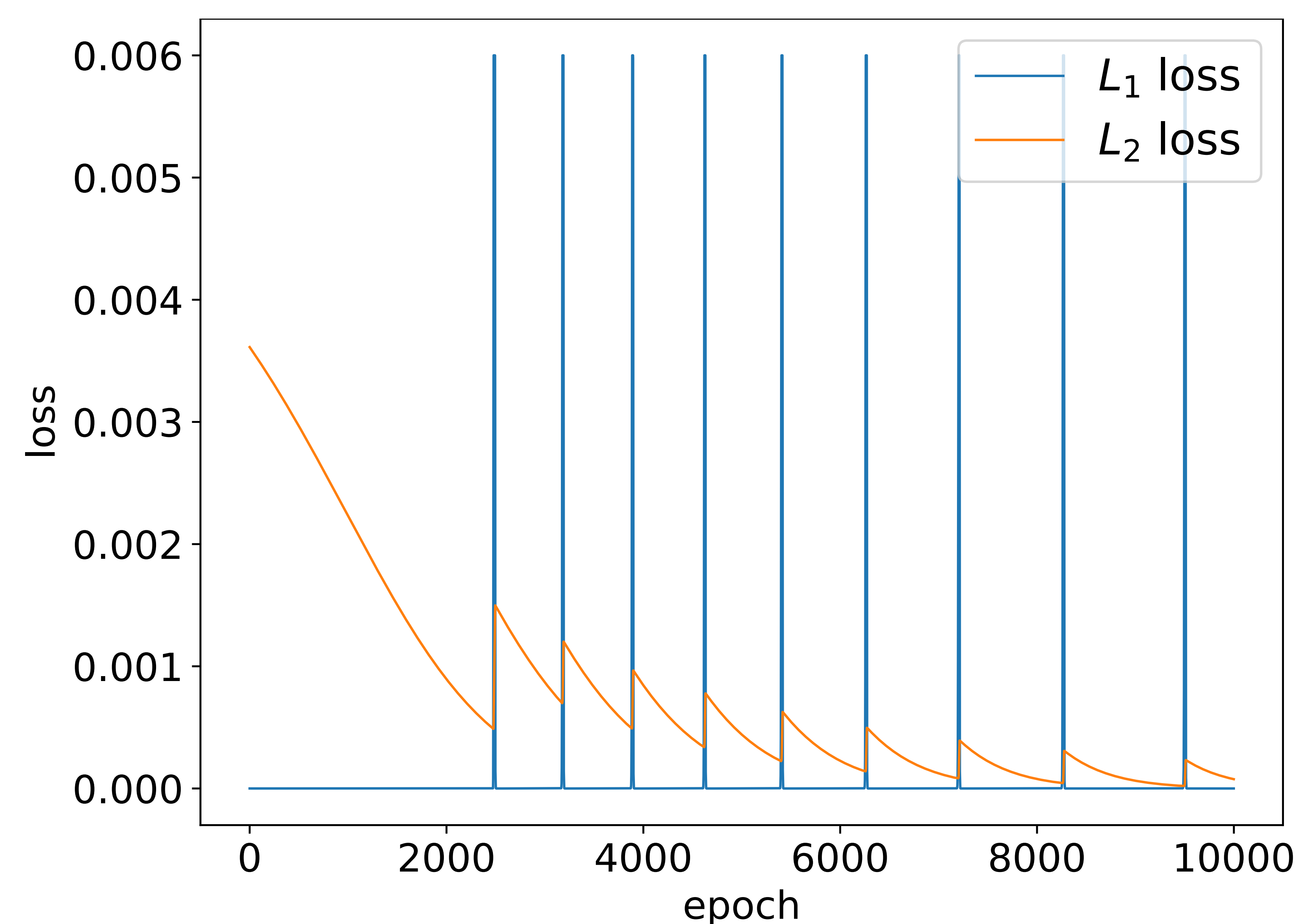
Loss Decomposition

$$L(\theta) = \underbrace{\frac{1}{2}\lambda_1(\alpha\beta_1)^2}_{L_1} + \underbrace{\frac{1}{2}\lambda_2(\alpha\beta_2 - 1)^2}_{L_2}.$$

Oscillation

Convergence

- L_1 learns the irrelevant x_1 and exhibits drastic fluctuations.
- L_2 dominates the loss descent while still non-monotonic.



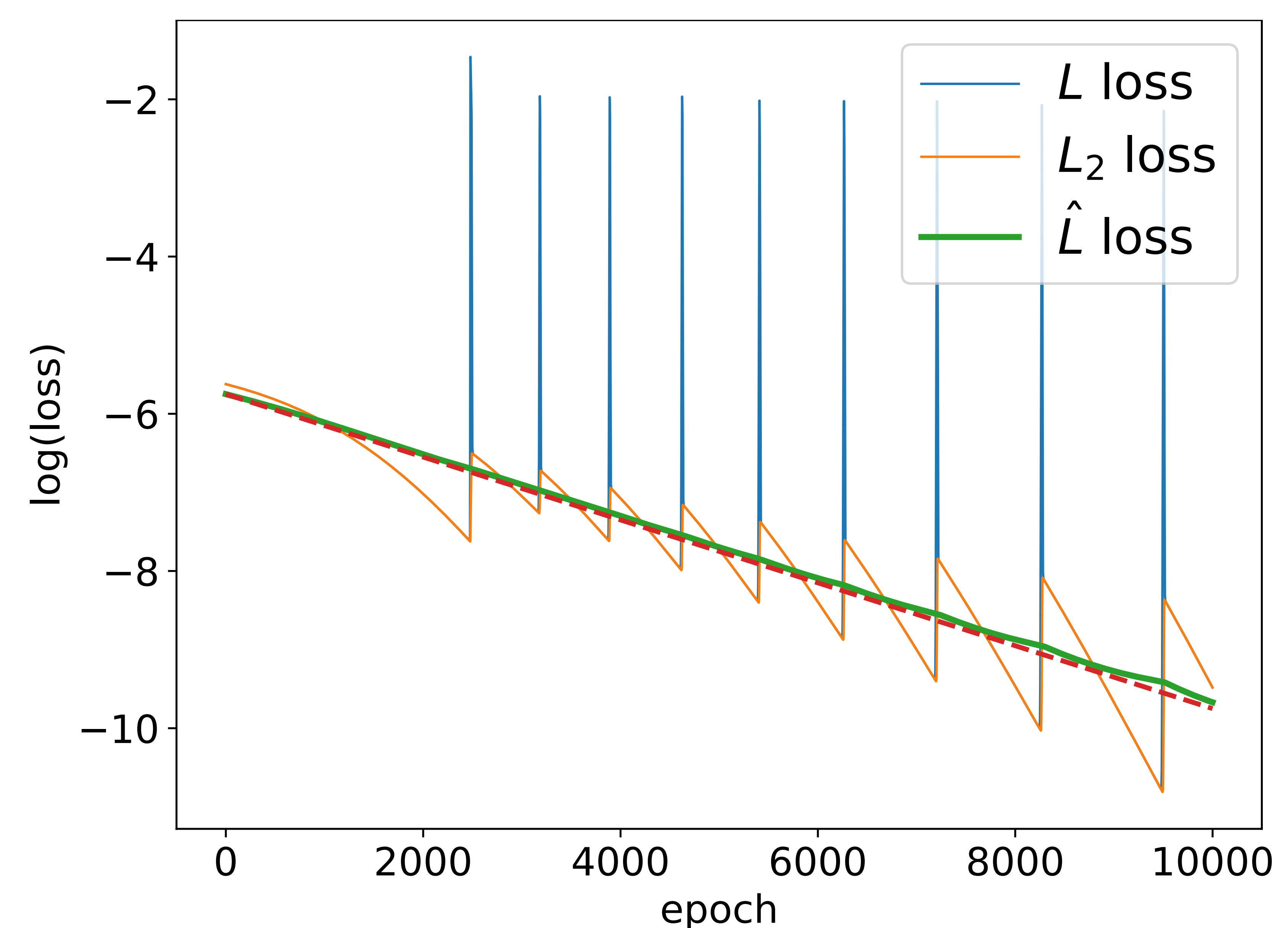
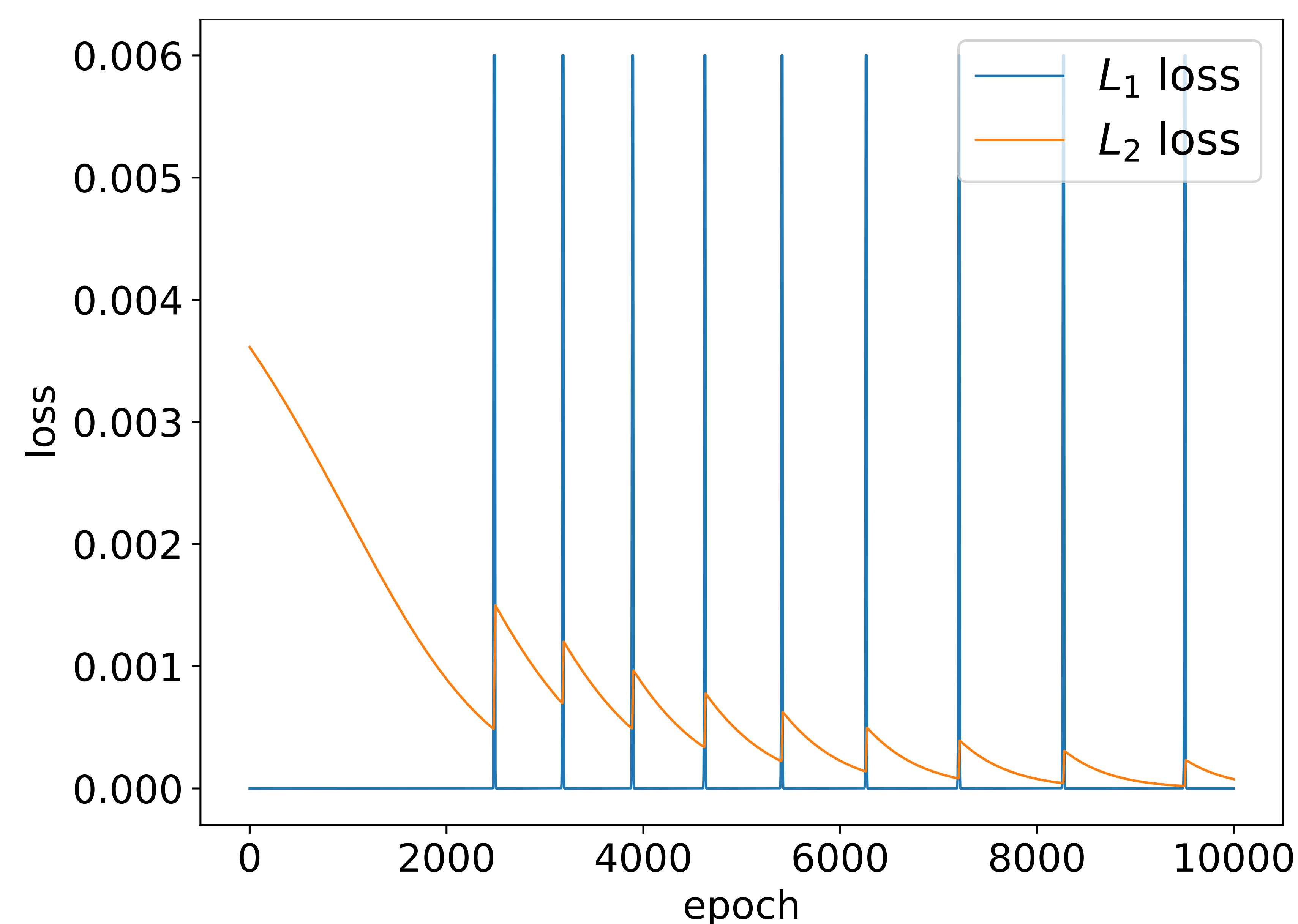
Loss Decomposition

$$L(\theta) = \underbrace{\frac{1}{2}\lambda_1(\alpha\beta_1)^2}_{L_1} + \underbrace{\frac{1}{2}\lambda_2(\alpha\beta_2 - 1)^2}_{L_2}.$$

Oscillation

Convergence

- L_1 learns the irrelevant x_1 and exhibits drastic fluctuations.
- L_2 dominates the loss descent while still non-monotonic.



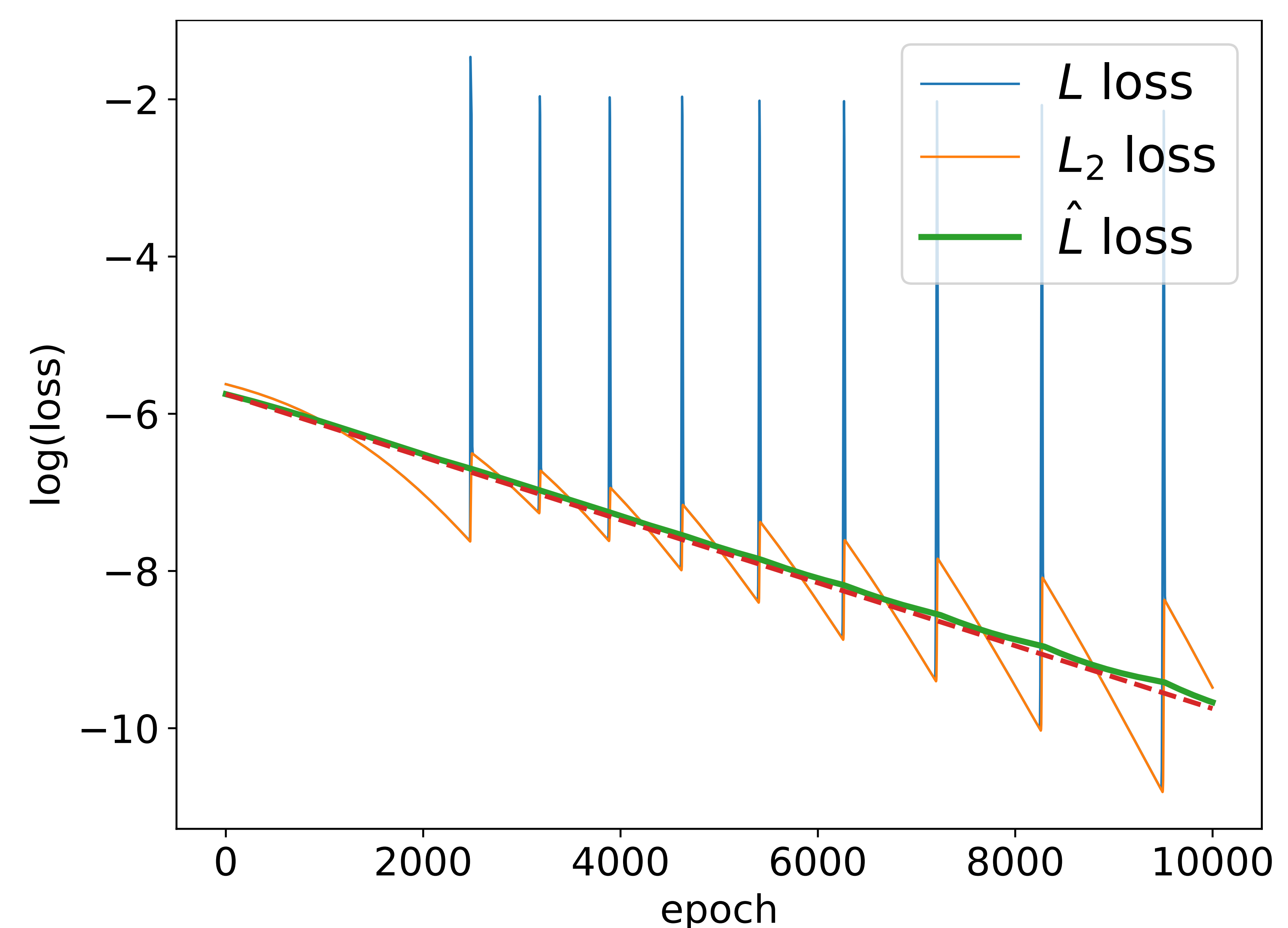
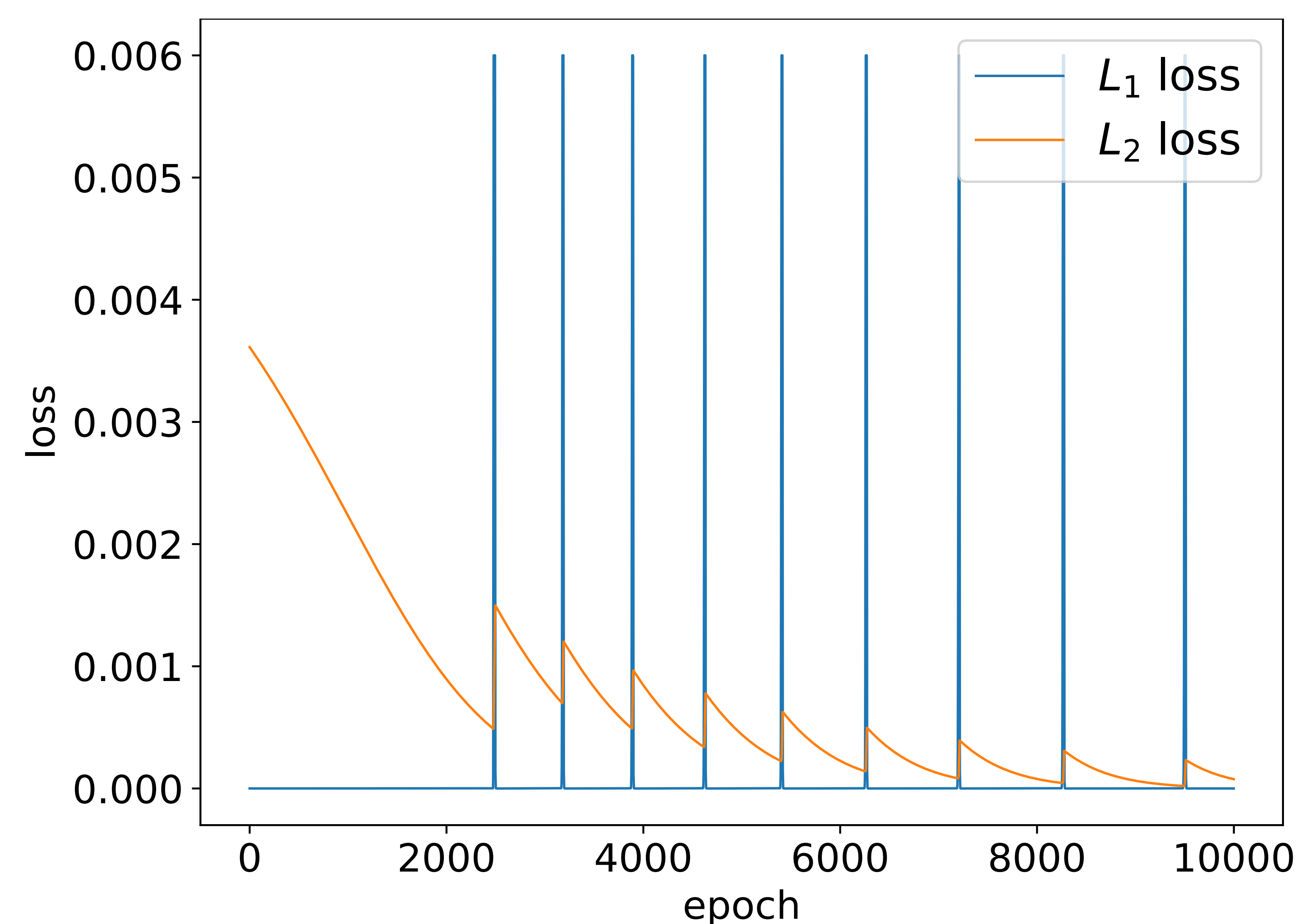
Loss Decomposition

$$L(\theta) = \underbrace{\frac{1}{2}\lambda_1(\alpha\beta_1)^2}_{L_1} + \underbrace{\frac{1}{2}\lambda_2(\alpha\beta_2 - 1)^2}_{L_2}.$$

Oscillation

Convergence

- L_1 learns the irrelevant x_1 and exhibits drastic fluctuations.
- L_2 dominates the loss descent while still non-monotonic.

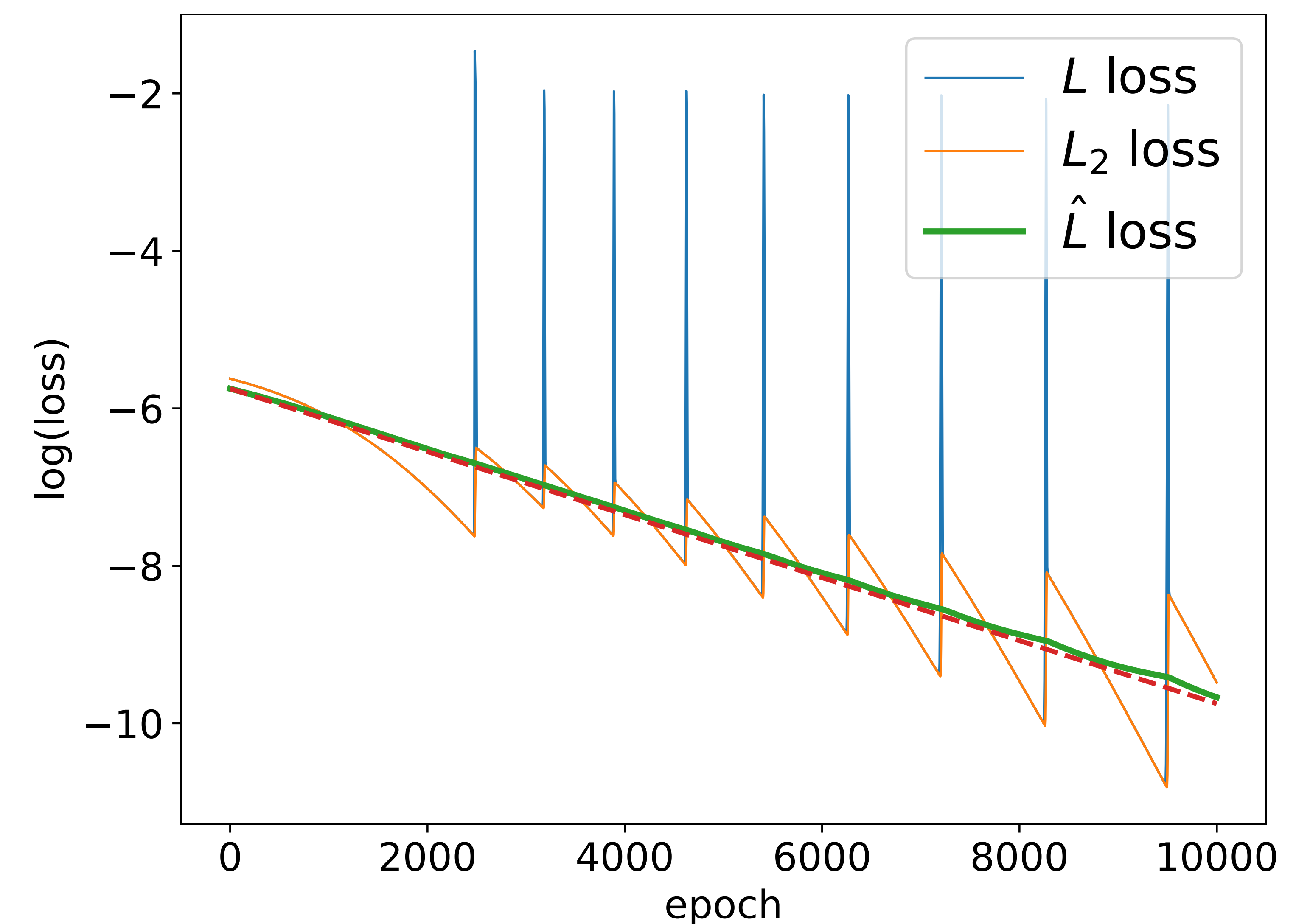


Loss Decay Rate

- Approximate L_2 by

$$\hat{L}(\theta) = \frac{1}{2}(1 - \hat{\alpha}\beta_2(t))^2,$$

where we restrict $\hat{\alpha} = \sqrt{2/(\lambda_1\eta)}$.



Theorem (Informal)

*Under initialization conditions, $\hat{L}$ is a good estimate for L_2 .
Moreover, $\hat{L}$ decays at a linear rate*

$$\left(1 - \frac{4\lambda_2}{\lambda_1}\right)^2 \leq \frac{\hat{L}(\theta(t+1))}{\hat{L}(\theta(t))} \leq \left(1 - \frac{\lambda_2}{\lambda_1}\right)^2.$$

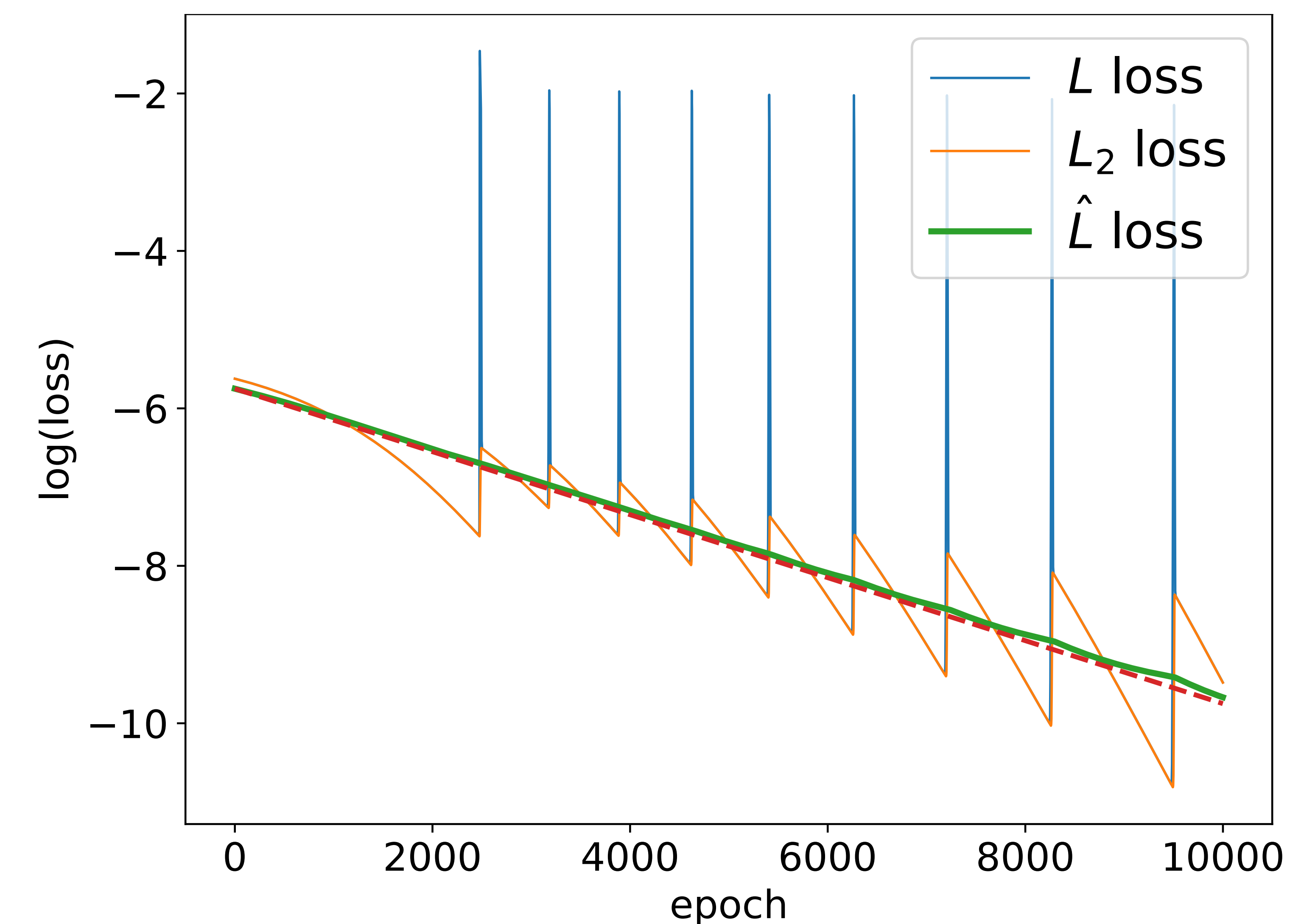
- A larger λ_2/λ_1 suggests greater signal-noise ratio, leading to faster loss descent.
- The decay rate is independent of the stepsize η .

Loss Decay Rate

- Approximate L_2 by

$$\hat{L}(\theta) = \frac{1}{2}(1 - \hat{\alpha}\beta_2(t))^2,$$

where we restrict $\hat{\alpha} = \sqrt{2/(\lambda_1\eta)}$.



Theorem (Informal)

Under initialization conditions, $\hat{L}$ is a good estimate for L_2 . Moreover, $\hat{L}$ decays at a linear rate

$$\left(1 - \frac{4\lambda_2}{\lambda_1}\right)^2 \leq \frac{\hat{L}(\theta(t+1))}{\hat{L}(\theta(t))} \leq \left(1 - \frac{\lambda_2}{\lambda_1}\right)^2.$$

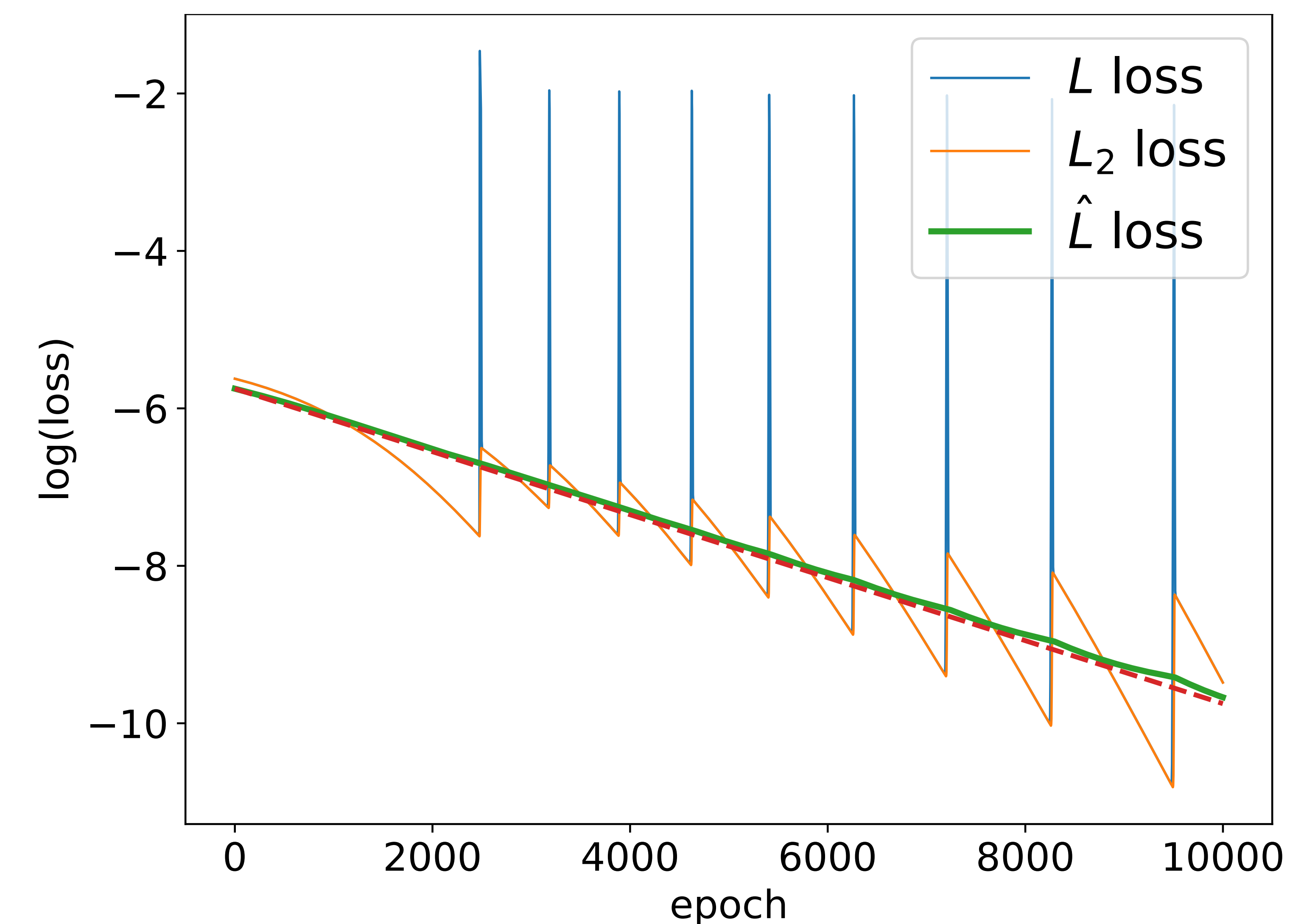
- A larger λ_2/λ_1 suggests greater signal-noise ratio, leading to faster loss descent.
- The decay rate is independent of the stepsize η .

Loss Decay Rate

- Approximate L_2 by

$$\hat{L}(\theta) = \frac{1}{2}(1 - \hat{\alpha}\beta_2(t))^2,$$

where we restrict $\hat{\alpha} = \sqrt{2/(\lambda_1\eta)}$.



Theorem (Informal)

*Under initialization conditions, $\hat{L}$ is a good estimate for L_2 .
Moreover, $\hat{L}$ decays at a linear rate*

$$\left(1 - \frac{4\lambda_2}{\lambda_1}\right)^2 \leq \frac{\hat{L}(\theta(t+1))}{\hat{L}(\theta(t))} \leq \left(1 - \frac{\lambda_2}{\lambda_1}\right)^2.$$

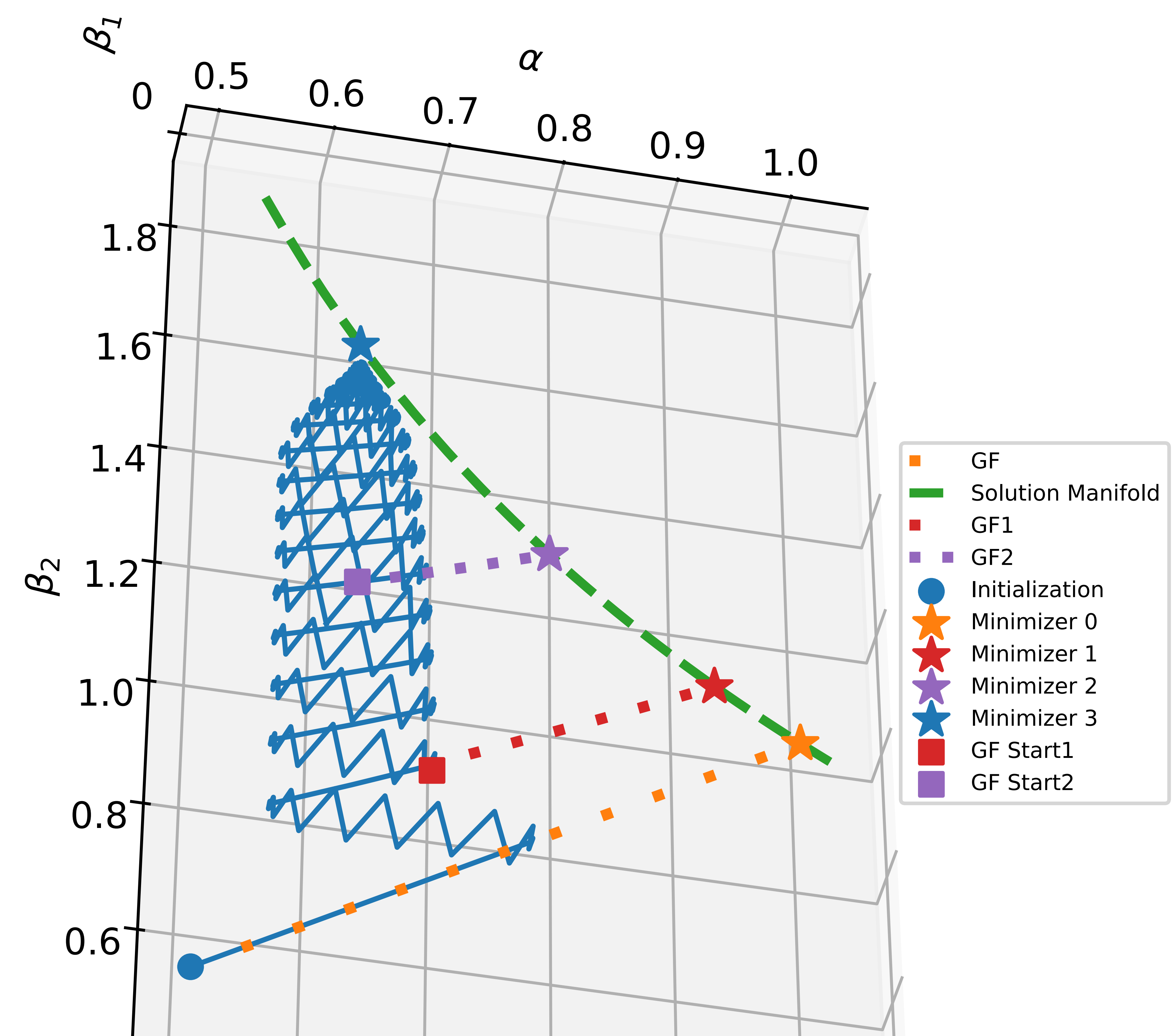
- A larger λ_2/λ_1 suggests greater signal-noise ratio, leading to faster loss descent.
- The decay rate is independent of the stepsize η .

Connection to Existing Work

Connection 1: Gradient Flow Solution (GFS, [Kreiser et al. 2023](#)):

The limit of $\dot{\theta}(t) = \nabla L(\theta)$

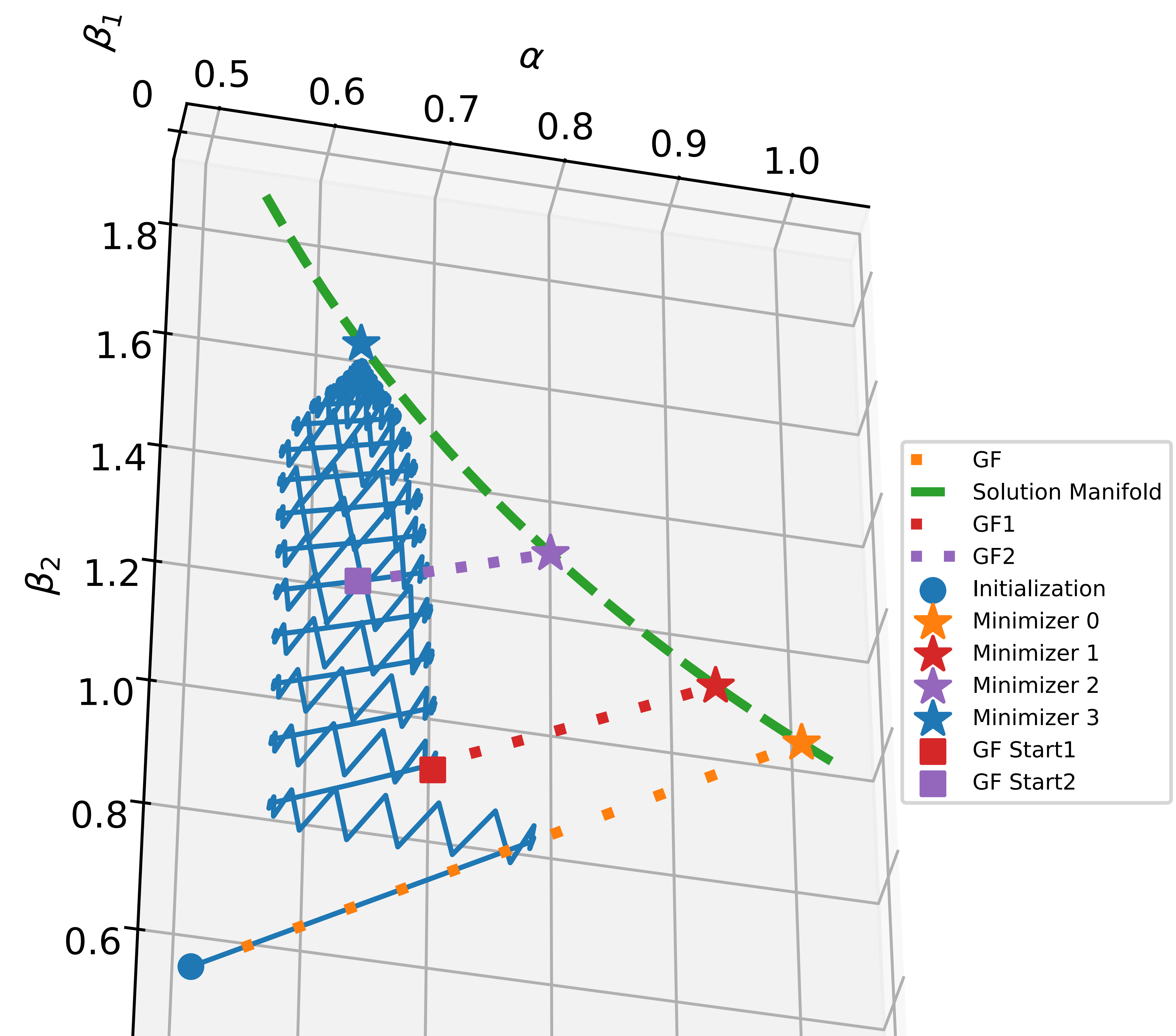
- Cover GFS analysis to two-dimensional inputs.
- Prove **monotonic decrease** of GFS sharpness.
- Clarify **distinction** between GFS and GD sharpness.



Connection to Existing Work

Connection 1: Gradient Flow Solution (GFS, [Kreiser et al. 2023](#)):

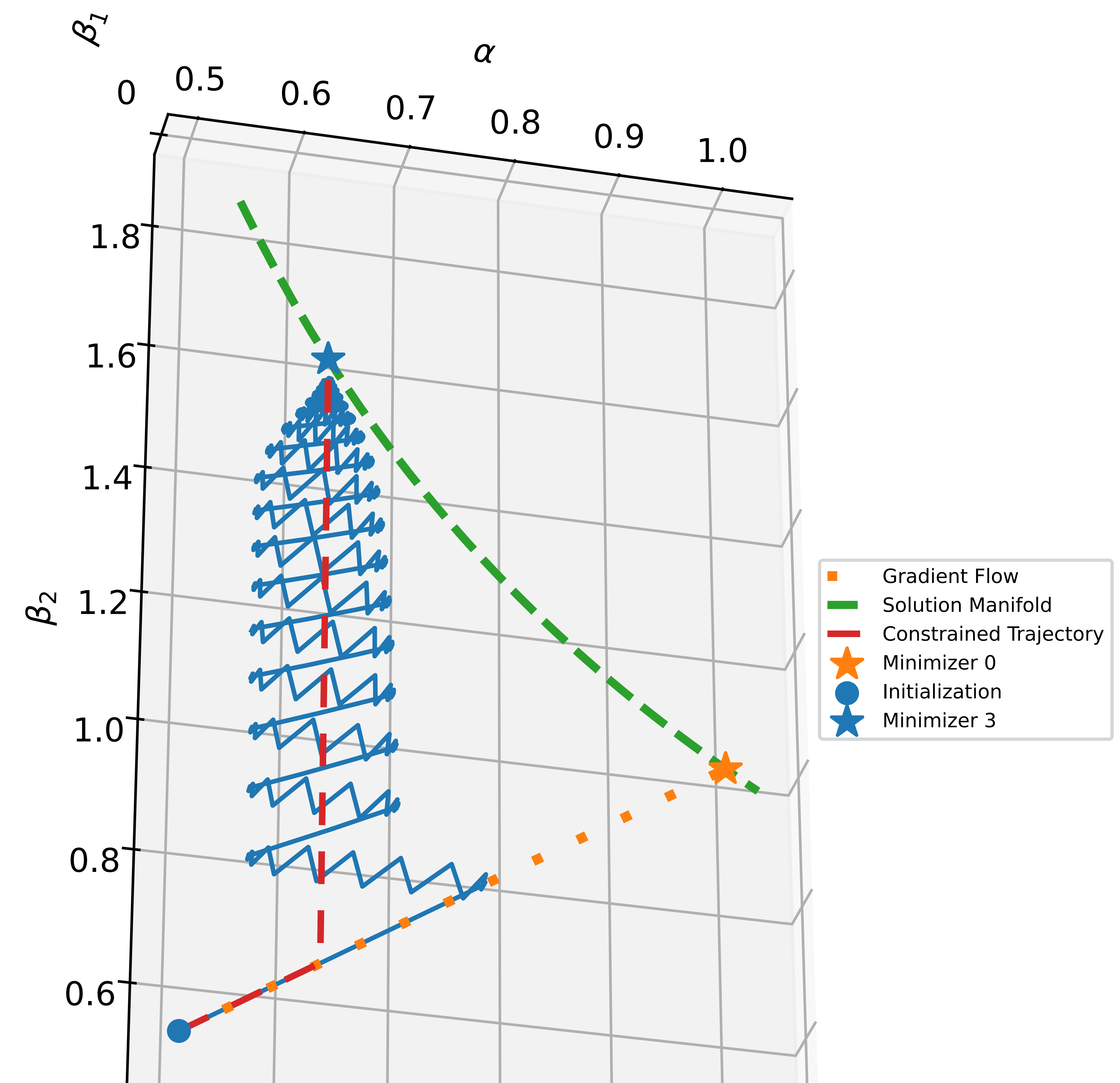
- Cover GFS analysis to two-dimensional inputs.
- Prove **monotonic decrease** of GFS sharpness.
- Clarify **distinction** between GFS and GD sharpness.



Connection to Existing Work

Connection 2: Constrained Trajectory (Damian et al. 2023):

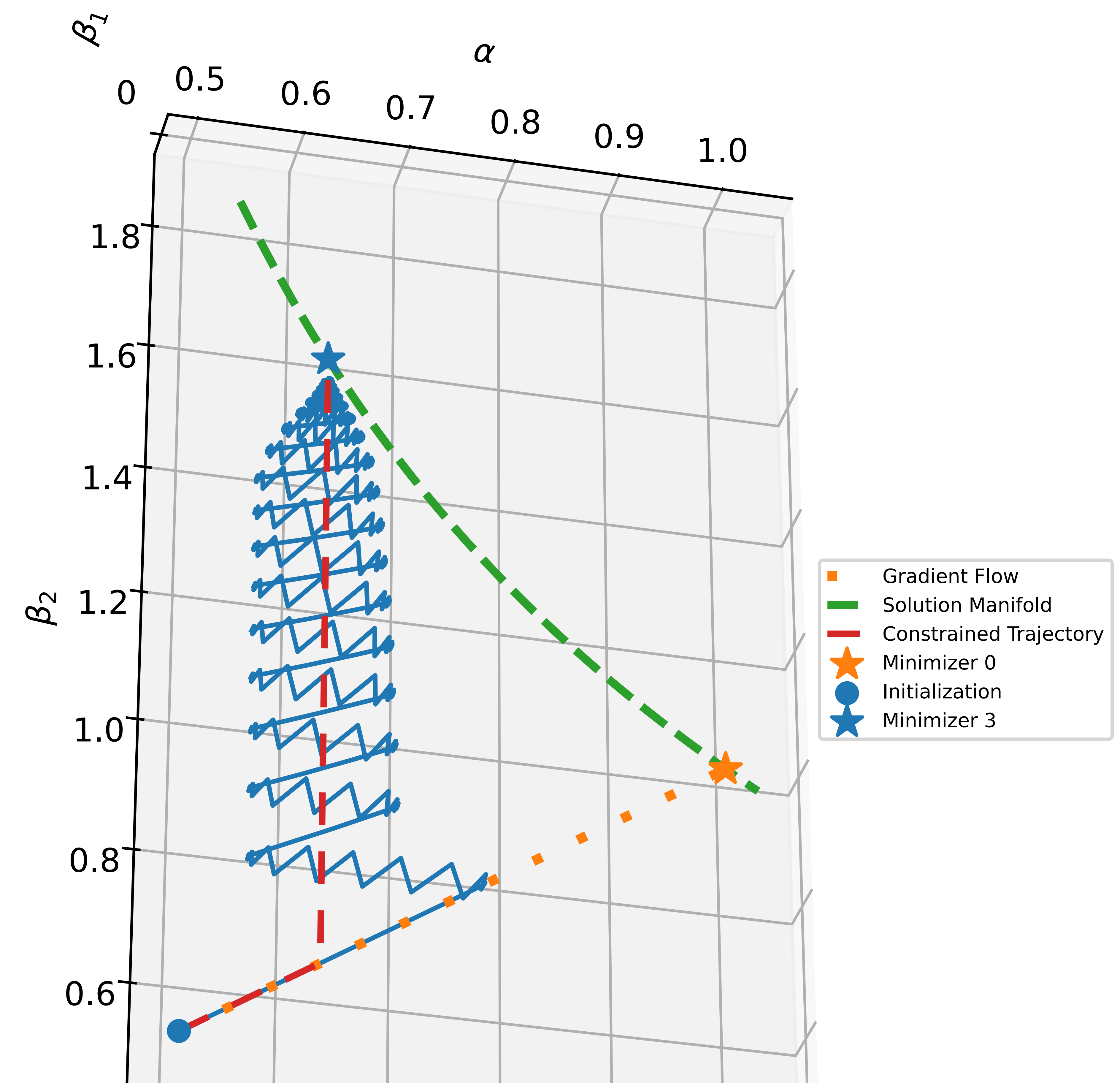
- Construct a well-behaved stable set $\mathcal{M}$, within which $S(\theta) \leq 2/\eta$ and $\beta_1 = 0$.
- Consider projected gradient descent (PGD) on $\mathcal{M}$. The loss converges at a linear rate $(1 - 2\lambda_2/\lambda_1)^2$.
- Remark: compare to the decay rate of $\hat{L} \Rightarrow \beta_1$ barely effects the loss descent.



Connection to Existing Work

Connection 2: Constrained Trajectory (Damian et al. 2023):

- Construct a well-behaved stable set $\mathcal{M}$, within which $S(\theta) \leq 2/\eta$ and $\beta_1 = 0$.
- Consider projected gradient descent (PGD) on $\mathcal{M}$. The loss converges at a linear rate $(1 - 2\lambda_2/\lambda_1)^2$.
- Remark: compare to the decay rate of $\hat{L} \Rightarrow \beta_1$ barely effects the loss descent.



Practical Implications

What this means for practice?

- Feature heterogeneity $\Rightarrow$ loss decay rate.
- Irrelevant features $\Rightarrow$ oscillations, not necessary slowdown.
- Larger stepsize $\nRightarrow$ faster convergence.

Beyond Our Simple Example:

- Feature relevance for optimization dynamics.
- Implicit regularization benefits of EoS.
- Future Directions...

Practical Implications

What this means for practice?

- Feature heterogeneity $\Rightarrow$ loss decay rate.
- Irrelevant features $\Rightarrow$ oscillations, not necessary slowdown.
- Larger stepsize $\nRightarrow$ faster convergence.

Beyond Our Simple Example:

- Feature relevance for optimization dynamics.
- Implicit regularization benefits of EoS.
- Future Directions...

Reference

- [1] Liu, L., Zhang, Z., Du, S., & Zhao, T. (2025). A Minimalist Example of Edge-of-Stability and Progressive Sharpening. arXiv preprint arXiv:2503.02809.
- [2] Cohen, J., Kaur, S., Li, Y., Kolter, J. Z., & Talwalkar, A. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. In International Conference on Learning Representations.
- [3] Damian, A., Nichani, E. and Lee, J. D. (2022). Self-stabilization: The implicit bias of gradient descent at the edge of stability. arXiv preprint arXiv:2209.15594.
- [4] Kreisler, I., Nacson, M. S., Soudry, D., & Carmon, Y. (2023, July). Gradient descent monotonically decreases the sharpness of gradient flow solutions in scalar networks and beyond. In Proceedings of the 40th International Conference on Machine Learning.

Thank You!