Understanding Edge-of-Stability Training Dynamics with a Minimalist Example

Xingyu Zhu*, Zixuan Wang*, Xiang Wang, Mo Zhou, Rong Ge
Duke University; Tsinghua University

May. 2023

Gradient Descent (GD):
Consider a neural network with parameter θ and objective $\mathcal{L}(\theta) \in \mathbb{R}$. Gradient descent with step size η updates θ by

$$\theta_{t+1} = \theta_t - \eta \nabla \mathcal{L}(\theta_t). \tag{1}$$

Sharpness of Neural Network:
Assuming the objective is twice differentiable, the sharpness of the network at θ is the largest eigenvalue of its parameter Hessian

$$S(\theta) \triangleq \lambda_1(\nabla^2 \mathcal{L}(\theta)). \tag{2}$$

Duke UNIVERSITY

Lemma 1 (The Descent Lemma)

Consider parameter θ optimized by gradient descent on $\mathcal{L}(\theta)$ with step size η, if the sharpness $S(\theta) \leq L$ for all θ, then

$$\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t) - \frac{\eta}{2}(2 - \eta L)\|\nabla \mathcal{L}(\theta)\|^2. \tag{3}$$

- When $L < 2/\eta$, $\mathcal{L}(\theta_t)$ decrease **monotonically** and θ converges to fixed point.
- When $S(\theta) > 2/\eta$, GD diverges in local quadratic approximation of $\mathcal{L}$ at θ.

We call $2/\eta$ the **stability threshold** of sharpness.

What is Edge-of-Stability (EoS)?

Edge-of-Stability (Cohen, Kaur, et al., 2021):
When training a deep neural networks using *full-batch* gradient descent with step size η, the sharpness of the model $S(\theta)$ will oscillate around $2/\eta$ while the objective $\mathcal{L}(\theta)$ keeps decreasing **non-monotonically**.

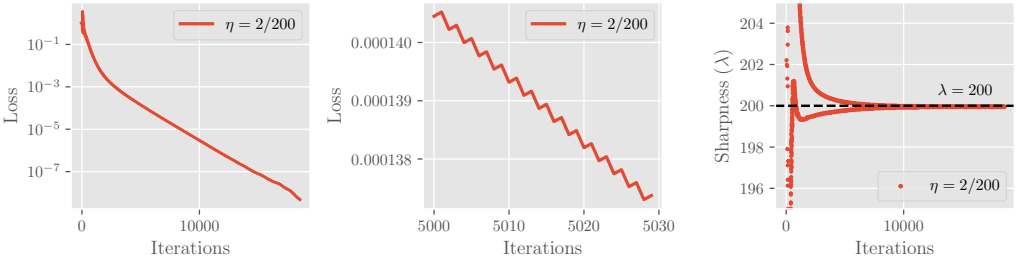


Figure 1: Loss and Sharpness of Training a 5-layer FC network on CIFAR-10 ($\eta = 0.01$)

What is Interesting about Edge-of-Stability?

Duke UNIVERSITY

1. **Convergence cannot be explained by traditional L-smooth analysis**
   - Diverging in quadratic approximation vs converging in real landscape
   - Monotonic vs non-monotonic decreasing of loss
2. **Step size determines converging sharpness**
   - Different from gradient flow trajectory (out of the NTK regime)
   - Implicit regularization effect of learning rate - better generalization? (Keskar et al., 2016; Jiang et al., 2019)
3. **Common in deep neural networks, but not for 2-layer networks**
   - How does depths relate to training dynamics?

Contents

1. Related Works (Previous and Concurrent)
2. Characterizing two EoS phenomena of interest
3. Proposing a minimalist example exhibiting such phenomena
4. Provide rigorous dynamics analysis near $2/\eta$ for the minimalist model
5. Connections to more general models
6. Chaotic behavior and EoS

Previous Works

1. Non-monotonic decreasing of loss has been empirically observed and discussed by Jastrzebski et al. (2020) and Lewkowycz et al. (2020). The edge of stability phenomenon was formalized by Cohen, Kaur, et al. (2021).
2. Under some special assumptions such as scale invariant objective (Lyu, Li, and Arora, 2022) / modified objective $\sqrt{\mathcal{L}}$ (Arora, Li, and Panigrahi, 2022), GD trajectories follows a sharpness minimizing flow near the manifold of minima. These trajectories lead to **flattest minima**.
3. Degree-2 factorization problem such as $\mathcal{L}(x, y) = (1 - xy)^2$ provably stabilizes and converges **below** $2/\eta$ (Wang et al., 2022; Chen and Bruna, 2022).

Concurrent Works

1. When the trajectory enters the EoS regime and oscillates along direction v, the third order information $\nabla^3 \mathcal{L}(\theta)(v, v)$ is approximately $\nabla S(\theta)$. Under the assumption of progressive sharpening, the degree-3 term constitutes a 4-stage heuristical dynamics of EoS (Damian, Nichani, and Lee, 2022).
2. Ahn et al. (2022) studied a simple single-neuron model and proved that the bias term is learnt when training enters the EoS regime.
3. Kreisler et al. (2023) studied the EoS dynamics in scalar networks, and proved that the GFS (gradient flow solution) sharpness is monotonically decreasing.

Concrete Characterization of Edge-of-Stability

Duke UNIVERSITY

1. Sharpness Concentration

Fix a learning rate η, there exists a large range of initializations $\mathbb{S}$ with different initial sharpness such that the GD trajectory with step size η from any $\theta \in \mathbb{S}$ converges to a minima with sharpness very close to $2/\eta$.

2. Sharpness Adaptivity

Fix an initialization θ, there exists a wide range of step sizes $\mathbb{K}$ such that for any $\eta \in \mathbb{K}$, GD trajectory starting at θ with step size η converges to a minima sharpness very close to $2/\eta$.

Minimalist Example for Edge of Stability

Consider the simple objective

$$\mathcal{L}(x, y) = \tfrac{1}{4}(1 - x^2 y^2)^2. \qquad (4)$$

Equivalent to a 4-layer scalar network with coupling entries (x, x, y, y) trained on $(1, 1)$ with square loss.

In the first quadrant:

1. Global minima $xy = 1$.
2. 2 **EoS Minima** (with sharpness $2/\eta$)

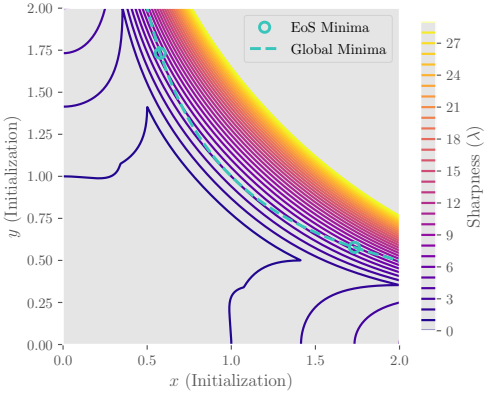


Figure 2: Sharpness and Loss of the Minimalist Objective

The simple objective

$$\mathcal{L}(x, y) = \tfrac{1}{4}(1 - x^2 y^2)^2 \tag{5}$$

exhibits both sharpness concentration and sharpness adaptivity.

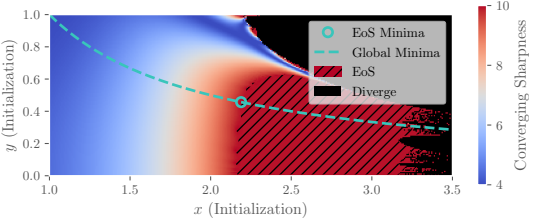


Figure 3: Sharpness Concentration ($\eta = 0.2$)

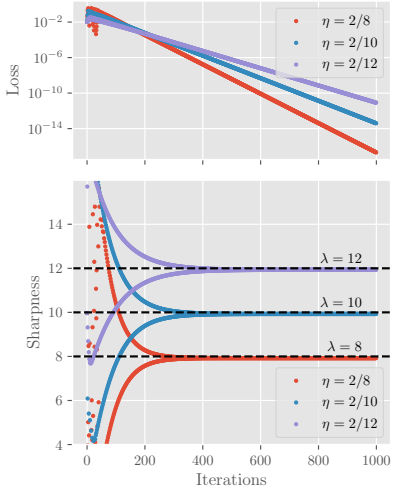


Figure 4: Sharpness Adaptivity

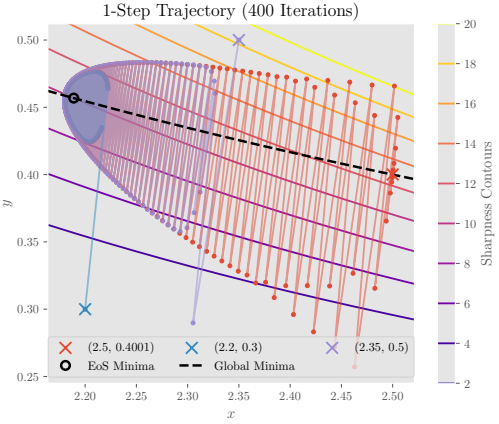


Figure 5: Sharpness Concentration
(Different initialization, same η)

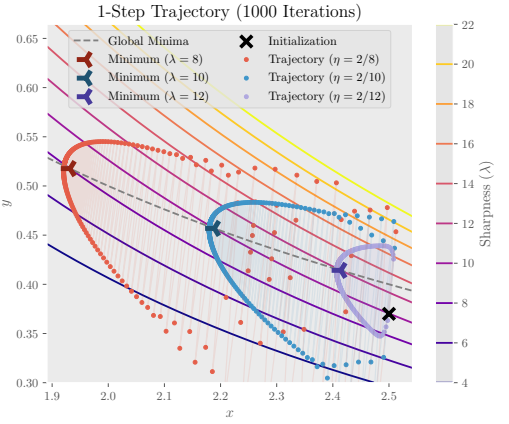


Figure 6: Sharpness Adaptivity
(Same initialization, different η)

EoS Training Trajectory for the Minimalist Example

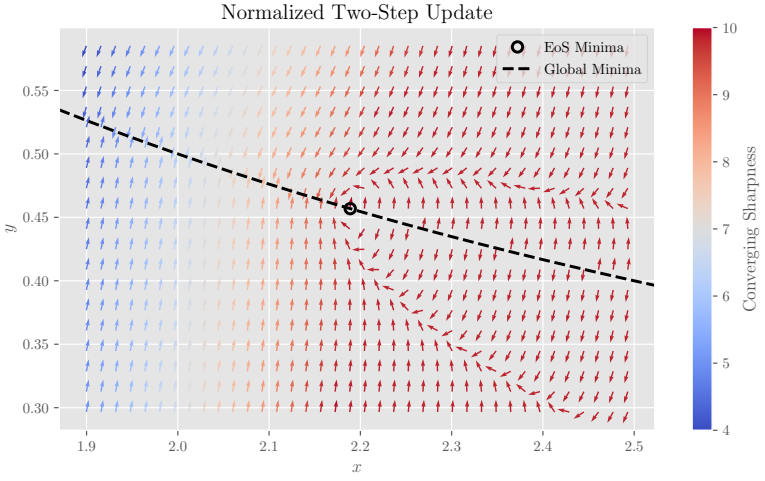


Figure 7: Directions of 2-Step Movement Near the EoS Minimum

The minima with sharpness $2/\eta$ is

$$(\breve{x}, \breve{y}) \triangleq \left(\frac{1}{\sqrt{2}} \left((-4 + \eta^{-2})^{\frac{1}{2}} + \eta^{-1} \right)^{\frac{1}{2}}, \sqrt{2} \left((-4 + \eta^{-2})^{\frac{1}{2}} + \eta^{-1} \right)^{-\frac{1}{2}} \right).$$

Consider the reparameterization

$$c \triangleq \sqrt{x^2 - y^2}, d \triangleq xy.$$

Then we have $(\breve{c}, \breve{d}) \triangleq \left((\eta^{-2} - 4)^{\frac{1}{4}}, 1 \right)$.
Moreover, let

$$a = c - \breve{c}, b = d - \breve{d}.$$

Our analysis will focus on (a, b).

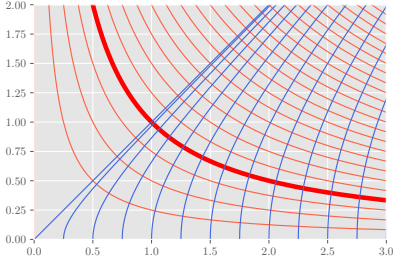


Figure 8: (c, d)-Reparameterization

Let $\kappa \triangleq \sqrt{\eta}$, the 1-step update under the (a, b)-reparameterization becomes

$$
\begin{aligned}
a' &= (\kappa^{-4} - 4)^{\frac{1}{4}} + \left(a + (\kappa^{-4} - 4)^{\frac{1}{4}}\right)\left(1 - \left((1+b)^3 - (1+b)\right)^2 \kappa^4\right)^{\frac{1}{2}}, \\
b' &= b + \left((1+b)^3 - 2(1+b)^5 + (1+b)^7\right)\kappa^4 \\
&\quad + \left((1+b) - (1+b)^3\right)\left(4(1+b)^2\kappa^4 + (a\kappa + (1 - 4\kappa^4)^{\frac{1}{4}})^4\right)^{\frac{1}{2}}.
\end{aligned} \tag{6}
$$

Denoting $\kappa \triangleq \sqrt{\eta}$, when $\kappa, |a|, |b|$ are all not too large, we have

$$\begin{aligned} a' &= a - 2b^2\kappa^3 + \mathcal{O}(\dots), \\ b' &= -b - 4ab\kappa - 3b^2 - b^3 + \mathcal{O}(\dots). \end{aligned} \tag{7}$$

After the parameterization

- a is monotonically decreasing (Movement Direction)
- b' is flipping signs and contains 2,3-order term of b (Oscillation Direction)

2-step movement is cleaner:

$$\begin{aligned} a'' &= a - 4b^2\kappa^3 + \mathcal{O}(\dots), \\ b'' &= b - 8ab\kappa + 16b^3 + \mathcal{O}(\dots). \end{aligned} \tag{8}$$

This gives ODE approximation

$$\frac{\mathrm{d}b}{\mathrm{d}a} = \frac{b'' - b}{a'' - a} = \frac{16b^3 - 8ab\kappa}{4b^2\kappa^3} \tag{9}$$

with solutions

$$b^2 = \frac{1}{2}a\kappa + \frac{1}{16}\kappa^4 + C\exp(8a\kappa^{-3}). \tag{10}$$

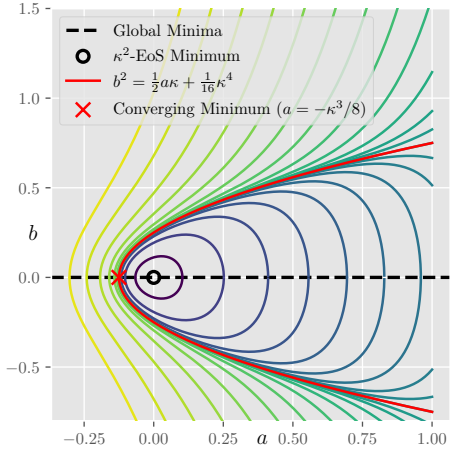


Figure 9: Solutions to (8)

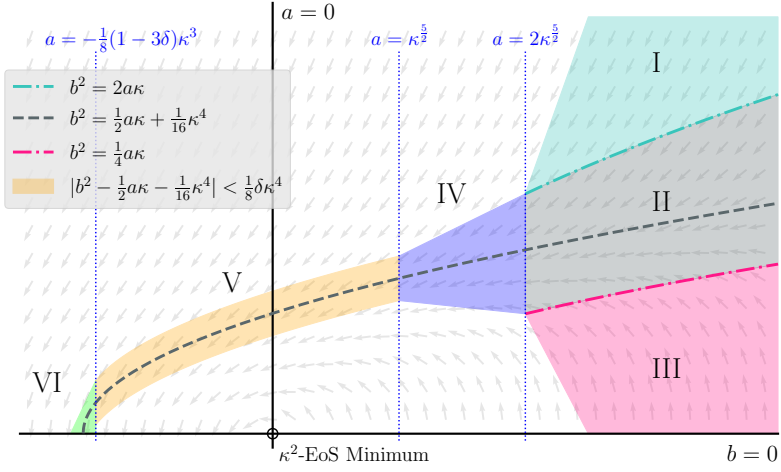


Figure 10: Convergence Diagram for the Minimalist Example

Duke UNIVERSITY

Theorem 2 (Sharpness Concentration in (a, b)-Parameterization, Informal)

For any κ less than some small constant, there exists a a box $\mathbb{S}_\kappa$ of width $\Theta(\kappa^{-1})$ and constant height that all GD trajectory with step size $\eta = \kappa^2$ initialized at (a_0, b_0) in $\mathbb{S}_\kappa$ converges ϵ-close to a minimum with a $\in (-\frac{5}{3}\kappa^3, -\frac{1}{10}\kappa^3)$ in poly $\left(\kappa^{-1}, \log(\epsilon^{-1}), \log(|b_0|^{-1})\right)$ steps.

Corollary 3 (Sharpness Concentration in (x, y)-Parameterization, Informal)

For all η less than some small constant, there exists a constant size region $\mathbb{S}_\eta$ that all GD trajectory with step size η initialized within $\mathbb{S}_\eta$ converges to a minimum with $\lambda \in (\frac{2}{\eta} - \frac{20}{3}\eta, \frac{2}{\eta})$

Duke UNIVERSITY

Corollary 4 (Sharpness Adaptivity in (x, y)-Parameterization, Informal)

For all α less than some small constant, there exists a constant size region $\mathbb{S}_\alpha$ and a set of learning rate $\mathbb{K}_\alpha$ of measure $\Theta(\alpha^2)$ that all GD trajectory with step size $\eta \in \mathbb{K}_\alpha$ initialized within $\mathbb{S}_\alpha$ converges to a minimum with $\lambda \in (\frac{2}{\eta} - \frac{20}{3}\eta, \frac{2}{\eta})$.

Difference Between Degree-2 and Degree-4 Models

Previously, Wang et al. (2022) and Chen and Bruna (2022) has conducted fine-grained convergence analysis for degree-2 models such as

$$\mathcal{L}(x, y) = (1 - xy)^2. \tag{11}$$

Stabilization exists in these cases, but the converging minima are usually significantly flatter than the stability threshold of $2/\eta$.

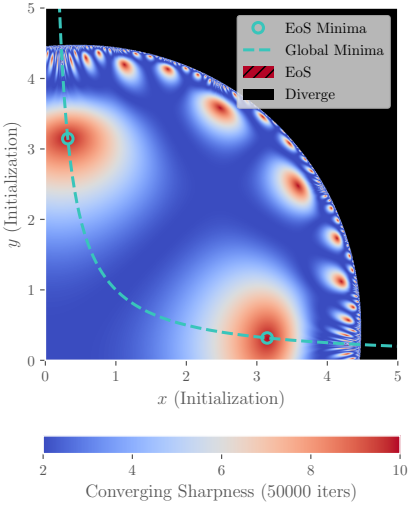


Figure 11: Converging λ of Degree-2 Model

Difference Between Degree-2 and Degree-4 Models

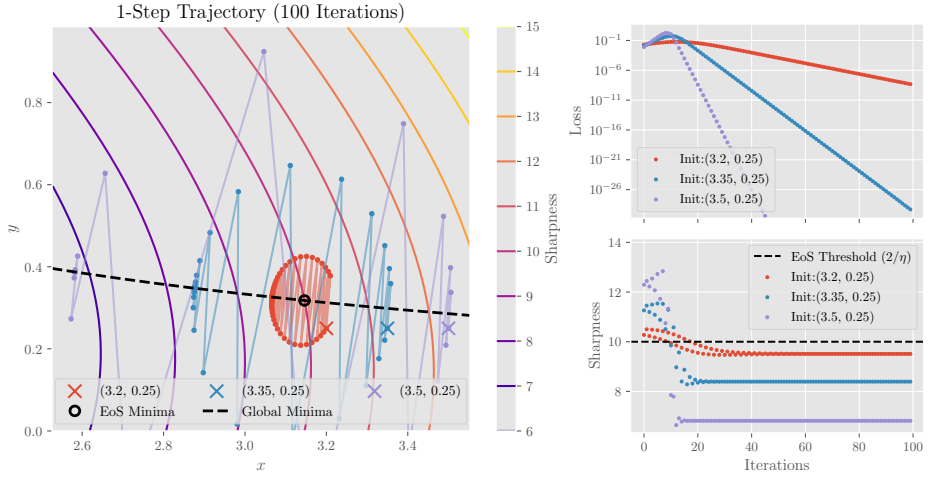


Figure 12: Lack of Sharpness Concentration for Degree-2 Models

Difference Between Degree-2 and Degree-4 Models

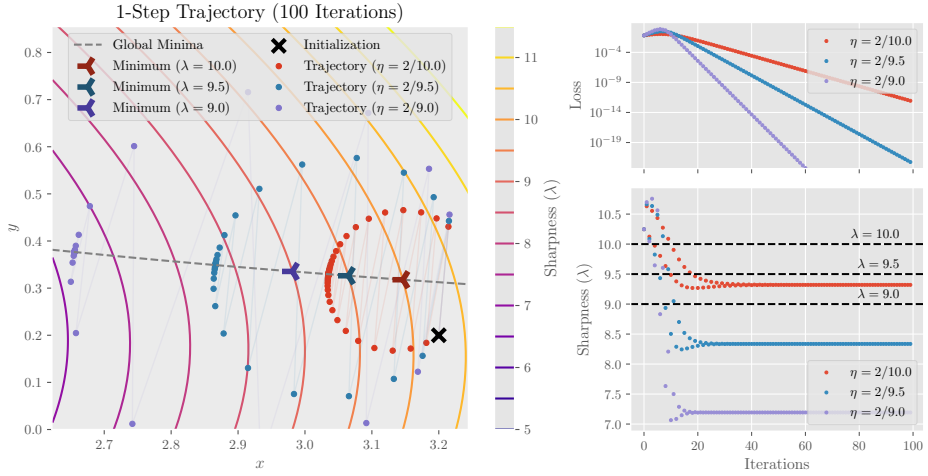


Figure 13: Lack of Sharpness Adaptivity for Degree-2 Models

Under the same reparameterization, the two step update of (a, b) under learning rate κ^2 can be approximated by

$$a'' \approx a - \sqrt{2}b^2\kappa^3, \quad b'' \approx b + 4\sqrt{2}ab\kappa. \tag{12}$$

This gives ODE approximation

$$\frac{\mathrm{d}b}{\mathrm{d}a} = \frac{b'' - b}{a'' - a} = \frac{-4ab\kappa}{b^2\kappa^3} = -\frac{4}{\kappa^2}\frac{a}{b} \tag{13}$$

with solutions as the family of ellipses

$$b^2 = 4(C - a^2)/\kappa^2. \tag{14}$$

Degree-3 Model with Mixed Behaviors

Consider $\mathcal{L}(x, y, z) = \frac{1}{2}(1 - xyz)^2$ with coupling initialization (x, y, y). This coupling will be preserve during training.

- Large entries - movement direction
- Small entries - oscillation direction
- Duplicated Small Entries
  Parabolic Trajectory - Strong EoS
- Single Small Entry
  Elliptical Trajectory - Weak EoS

Takeaway

Degree of Model Matters to EoS

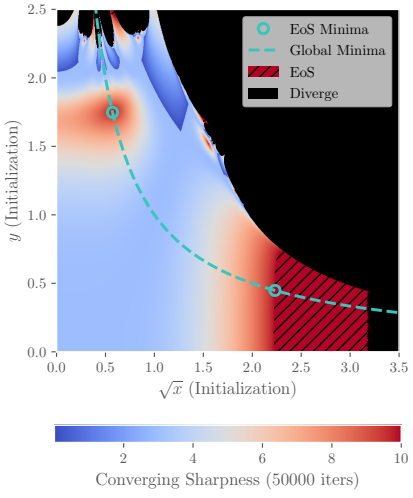


Figure 14: Converging λ of Degree-3 Model

GD on a 3-layer scalar network with initialization $(6, 0.4, 0.1)$. The small entries converge to the same value, and exhibits strong EoS.

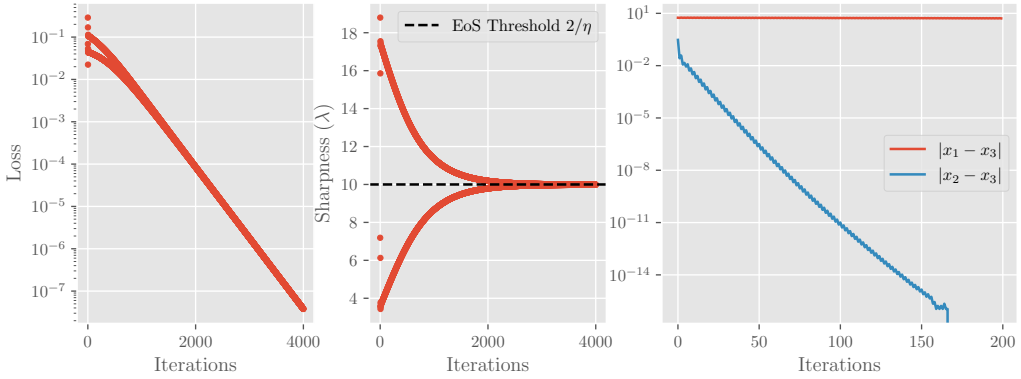


Figure 15: EoS (left, middle); Movement of Entries (right)

GD on a 7-layer scalar network with initialization $(2, 2.5, 3, 0.1, 0.2, 0.3, 0.4)$. The small entries converge to the same value, and exhibits strong EoS.

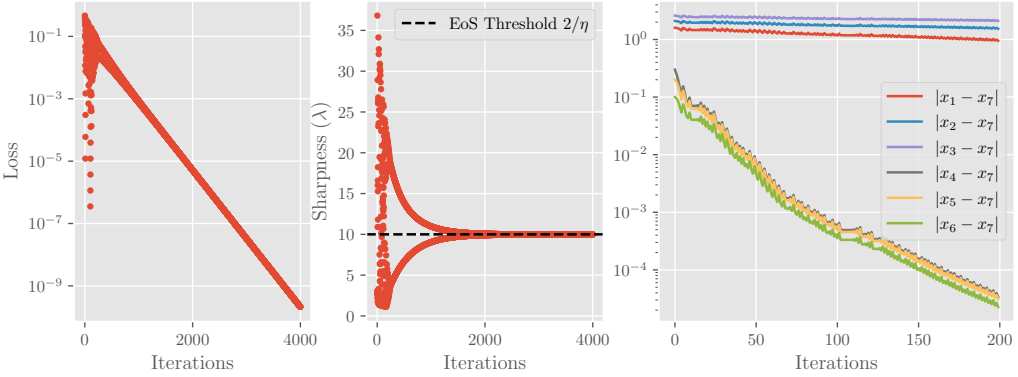


Figure 16: EoS (left, middle); Movement of Entries (right)

Summary of GD on scalar network with non-coupling initialization:

- **Large Entries:**
  Monotonically travel along a 1-d subspace (movement direction)
- **Small Entries:**
  Oscillate and converge to the same value (oscillation direction)

Reduce to the 2D case with parabolic trajectory if more than one "small entry".

Takeaway

Duplication of entries is not an artifact we have to impose for EoS to exist, but a natural consequence of gradient descent dynamics.

Dataset: 50-image binary subset of CIFAR-10: cars (-1) & planes (+1)
Model: 5-layer ELU/tanh-activated fully-connected network; MSE loss

Parabolic trajectory on **2D subspace** toward end of convergence

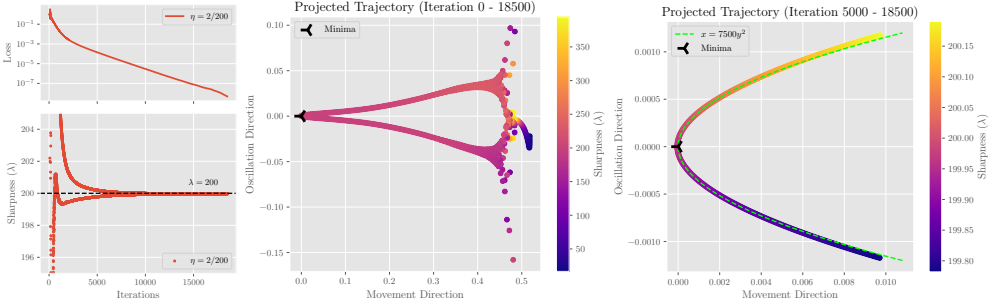
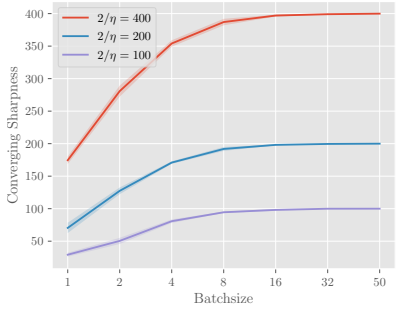
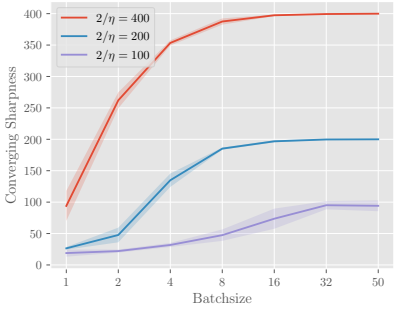


Figure 17: EoS (left); Global Projected Trajectory (mid); Local Projected Trajectory (right)



(a) ELU activation



(b) tanh activation

Figure 18: 5-layer FC networks trained on 50-image subset of CIFAR-10 with SGD of varying batchsizes ($\eta = 0.005, 0.01, 0.02$)

Even for the minimalist example, proving global convergence result is very difficult.

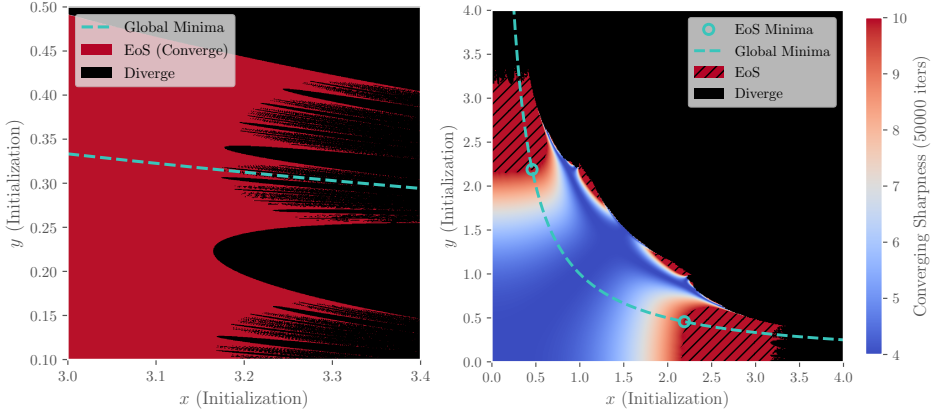


Figure 19: Fractal Structures in Boundary of Converging Initializations.

Bifurcation in GD dynamics

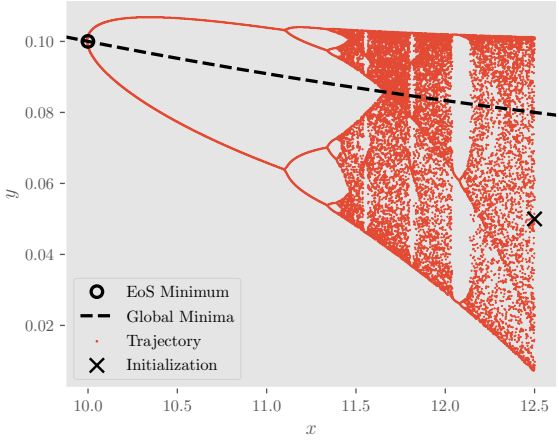


Figure 20: Bifurcation in GD Trajectory Initialized Near the Boundary of Convergence

Revisit the 2-step movement:

$$\begin{aligned} a'' &\approx a - 4b^2\kappa^3 \\ b'' &\approx b + 8ab\kappa - 16b^3 \\ &= b(1 + 8a\kappa - 16b^2). \end{aligned} \tag{15}$$

With small κ, b, a is relatively stationary. Dynamics of b can be approximated by the bifurcation diagram of the degree-3 polynomial mapping $b(1 + 8a\kappa - 16b^2)$ parameterized by a.

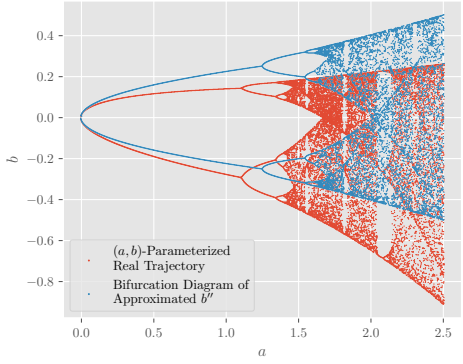


Figure 21: Predicted Bifurcation using $b(1 + 8a\kappa - 16b^2)$

Future Works

1. Understanding the coupling of entries by GD
2. Identify the source of degree-3 term causing EoS in general neural networks.
3. Leverage boundary of bifurcation map on global convergence analysis.
4. Study EoS in stochastic gradient descent / adaptive methods (Cohen, Ghorbani, et al., 2022).

References I

Ahn, Kwangjun et al. (2022). "Learning threshold neurons via the" edge of stability"". In: *arXiv preprint arXiv:2212.07469*.

Arora, Sanjeev, Zhiyuan Li, and Abhishek Panigrahi (2022). "Understanding Gradient Descent on Edge of Stability in Deep Learning". In: *arXiv preprint arXiv:2205.09745*.

Chen, Lei and Joan Bruna (2022). "On Gradient Descent Convergence beyond the Edge of Stability". In: *arXiv preprint arXiv:2206.04172*.

Cohen, Jeremy M, Behrooz Ghorbani, et al. (2022). "Adaptive gradient methods at the edge of stability". In: *arXiv preprint arXiv:2207.14484*.

Cohen, Jeremy M, Simran Kaur, et al. (2021). "Gradient descent on neural networks typically occurs at the edge of stability". In: *arXiv preprint arXiv:2103.00065*.

Damian, Alex, Eshaan Nichani, and Jason D Lee (2022). "Self-stabilization: The implicit bias of gradient descent at the edge of stability". In: *arXiv preprint arXiv:2209.15594*.

Jastrzebski, Stanislaw et al. (2020). "The break-even point on optimization trajectories of deep neural networks". In: *arXiv preprint arXiv:2002.09572*.

Jiang, Yiding et al. (2019). "Fantastic generalization measures and where to find them". In: *arXiv preprint arXiv:1912.02178*.

Keskar, Nitish Shirish et al. (2016). "On large-batch training for deep learning: Generalization gap and sharp minima". In: *arXiv preprint arXiv:1609.04836*.

Kreisler, Itai et al. (2023). "Gradient Descent Monotonically Decreases the Sharpness of Gradient Flow Solutions in Scalar Networks and Beyond". In.

Lewkowycz, Aitor et al. (2020). "The large learning rate phase of deep learning: the catapult mechanism". In: *arXiv preprint arXiv:2003.02218*.

References III

Lyu, Kaifeng, Zhiyuan Li, and Sanjeev Arora (2022). "Understanding the Generalization Benefit of Normalization Layers: Sharpness Reduction". In: *arXiv preprint arXiv:2206.07085*.

Wang, Yuqing et al. (2022). "Large Learning Rate Tames Homogeneity: Convergence and Balancing Effect". In: *International Conference on Learning Representations*.