Gradient Descent at the Edge of Stability: A Warmup

Bingrui Li

2022.12.16

Outline

- **Preliminaries**
- What is EoS phenomena?
- When does EoS occur?
- Why does EoS occur?(Examples and Intuition)
- How about SGD?

Not Included

- The rigorous explanation of main theorems

GD on Quadratic objective

Consider $L(\theta) = \frac{1}{2}\sum_{i=1}^{n} \lambda_i \theta_i^2$, where $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d > 0$.
Gradient descent with fixed step size η gives

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$
$$\theta_{t+1,i} = (1 - \eta\lambda_i)\theta_{t,i}$$

- To guarantee convergence to the global minima, we must have $\eta < 2/\lambda_1$, i.e. $\lambda_1 < 2/\eta$.
- Each dimension decreases independently
- If $\lambda_2 < 2/\eta < \lambda_1$, all dimension except the first dimension converges

Implicit bias in GD/SGD (in deep learning)

When multiple minima exists, the algorithm plays an active role for selecting the solution

- NTK
- Flatness/Sharpness
  - EoS
  - Stability

…

Outline

- Preliminaries
- **What is EoS phenomena?**
- When does EoS occur?
- Why does EoS occur?
- How about SGD?

EoS phenomena

- Progressive Sharpening

1. For any reasonable step size η the sharpness of iterates increases throughout training until it reaches $2/\eta$

- Edge of Stability

2. For the rest of training, the sharpness hovers right at, or just above, the value $2/\eta$

3. The train loss behaves non-monotonically, yet consistently decreases over long timescales.

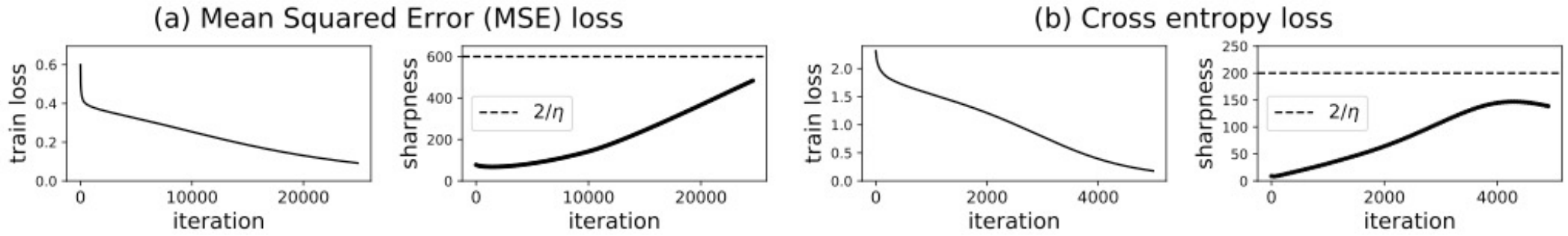


Figure 3: **So long as the sharpness is less than $2/\eta$, it tends to continually increase during gradient descent.** We train a network to completion (99% accuracy) using gradient descent with a very small step size. We consider both MSE loss (**left**) and cross-entropy loss (**right**).

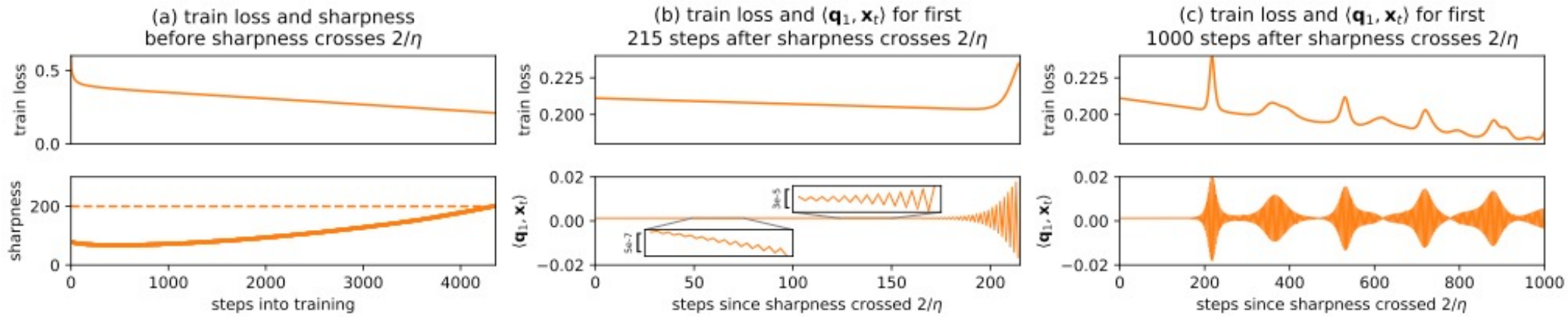


Figure 4: **Once the sharpness crosses $2/\eta$, gradient descent becomes destabilized**. We run gradient descent at $\eta = 0.01$. **(a)** The sharpness eventually reaches $2/\eta$. **(b)** Once the sharpness crosses $2/\eta$, the iterates start to oscillate along $\mathbf{q}_1$ with ever-increasing magnitude. **(c)** Somehow, GD does not diverge entirely; instead, the train loss continues to decrease, albeit non-monotonically.

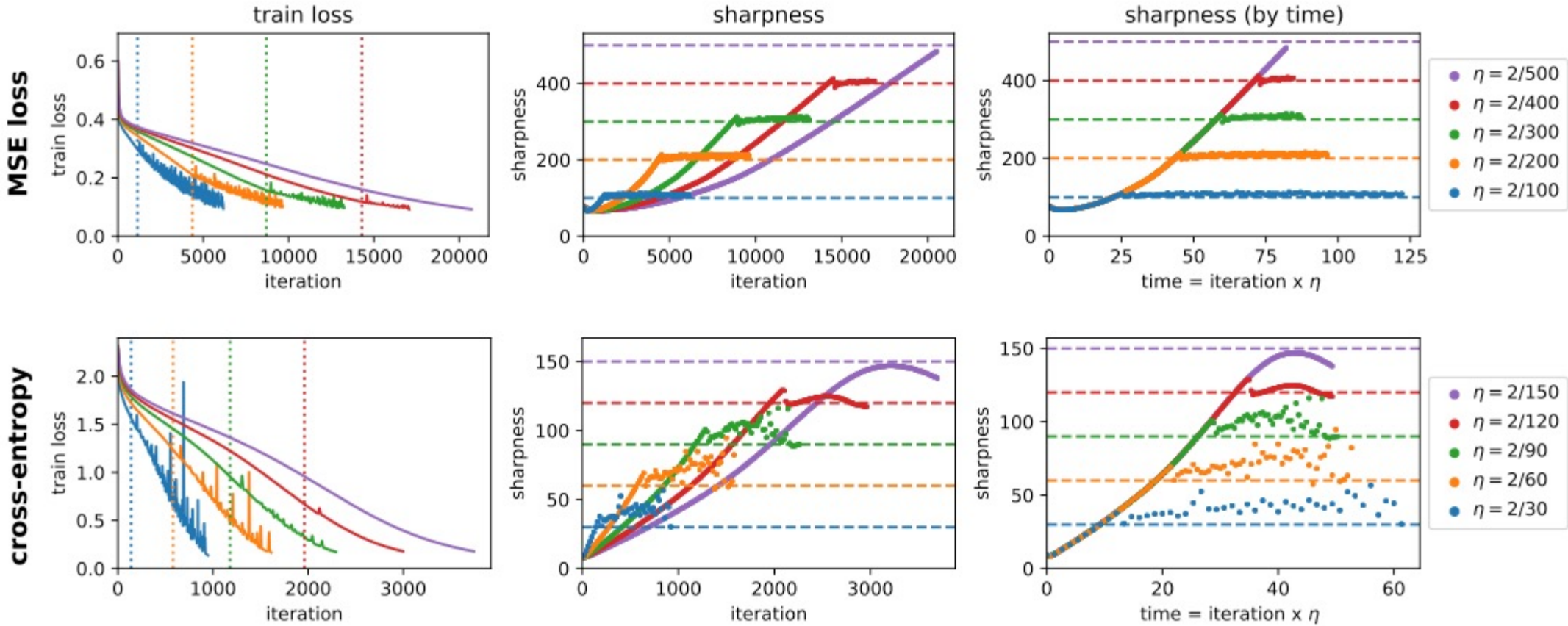
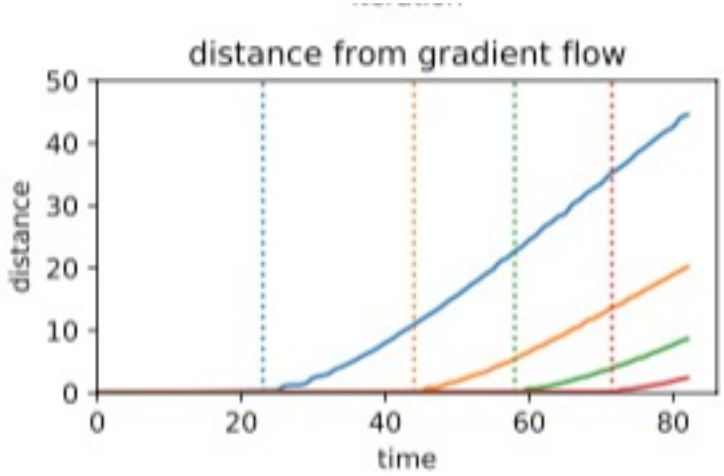
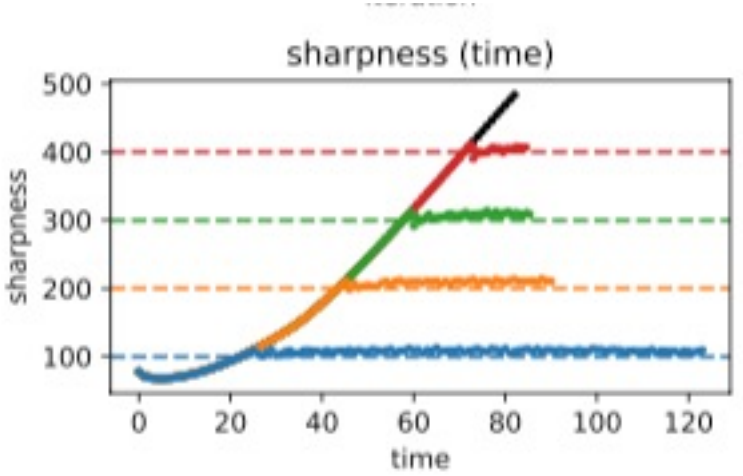


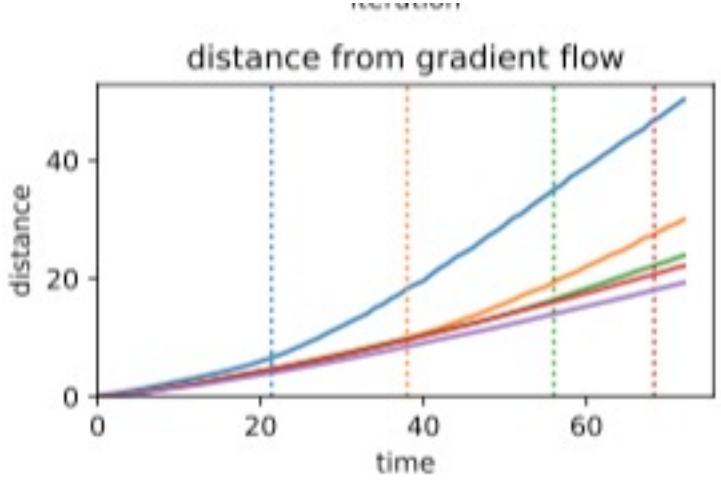
Figure 5: **After the sharpness reaches $2/\eta$, gradient descent enters the Edge of Stability**. A network is trained with gradient descent at a range of step sizes (see legend), using both MSE loss (**top row**) and cross-entropy (**bottom row**). **Left**: the train loss curves, with a vertical dotted line at the iteration where the sharpness first crosses $2/\eta$. **Center**: the sharpness, with a horizontal dashed line at the value $2/\eta$. **Right**: sharpness plotted by time (= iteration $\times$ η) rather than iteration.

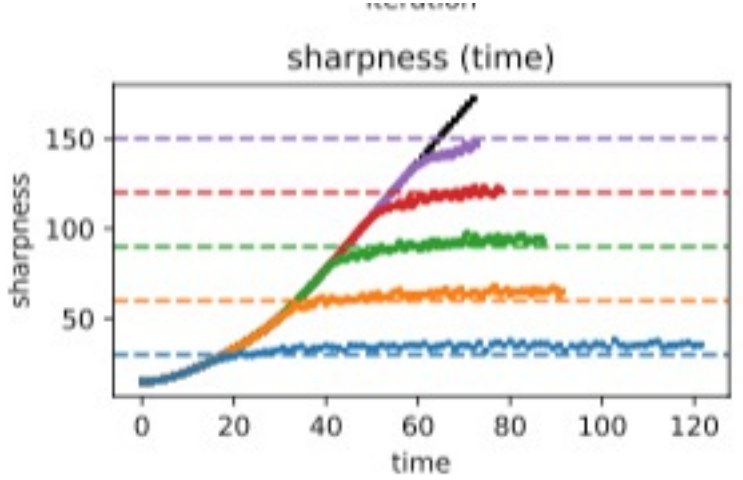
Distance from gradient flow

Setting:
2-layer FC network
MSE loss
Up: tanh activation
Bottom: ReLU activation









Stepsize drop make sharpening continue

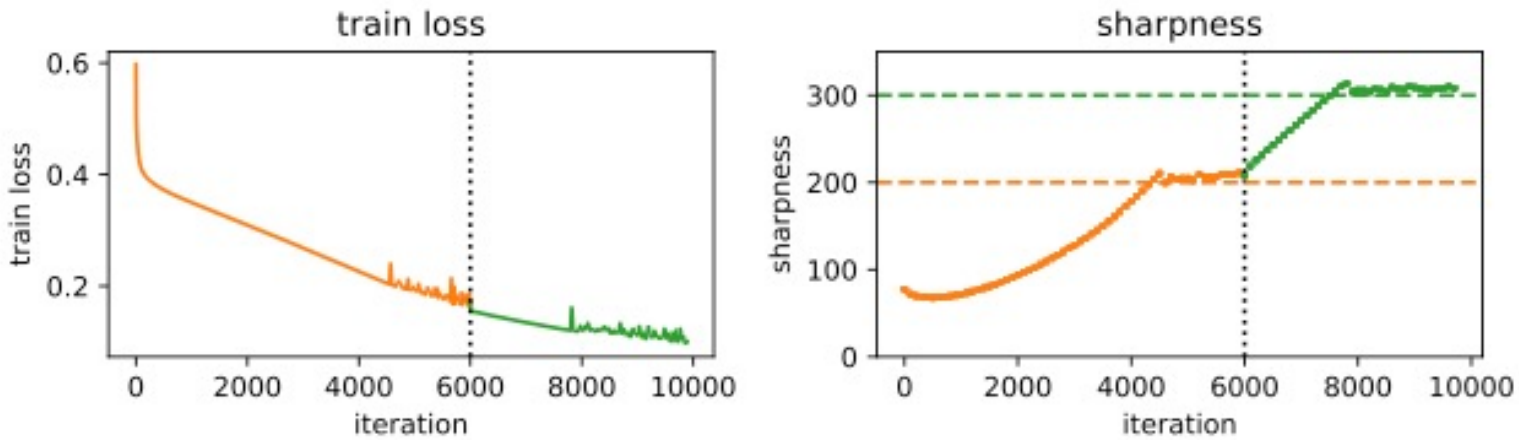
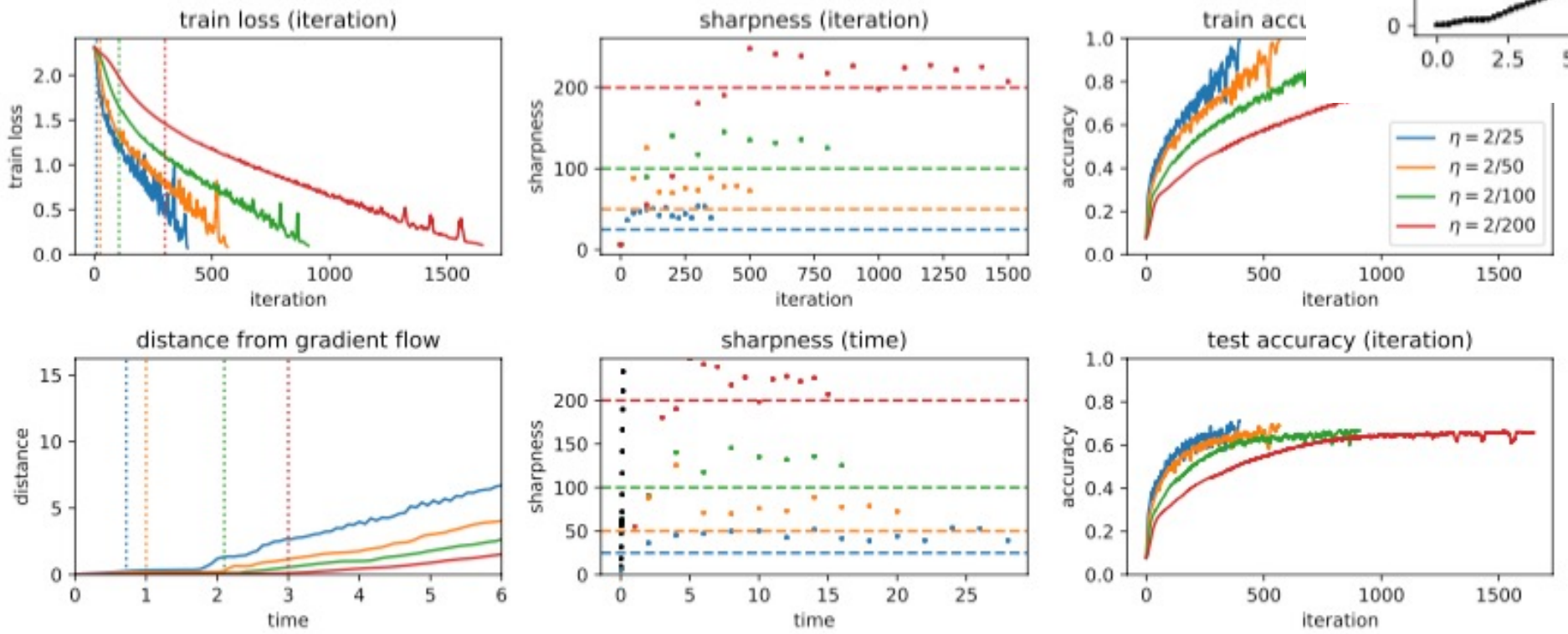


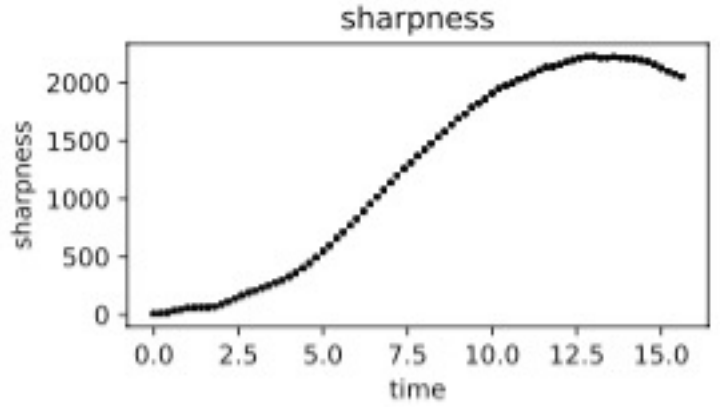
Figure 7: **After a learning rate drop, progressive sharpening resumes**. We start training at $\eta = 2/200$ (orange) and then after 6000 iterations (dotted vertical black line), we cut the step size to $\eta = 2/300$ (green). Observe that as soon as the step size is cut, the sharpness starts to rise.

Outline

- Preliminaries
- What is EoS phenomena?
- **When does EoS occur?**
  - **model and data, width and depth, step size, initialization**
- Why does EoS occur?
- How about SGD?

Model and Data complexity





VGG-11
CIFAR10

"model and dataset complexity substantially increase the sharpness"

Figure 90: We train a **VGG with BN** to completion using **gradient descent** at different step sizes (see legend in the top right pane). **Top left**: we plot the train loss, with a vertical dotted line marking the iteration where the sharpness first crosses $2/\eta$. **Top center**: we plot the evolution of

Model and Data complexity

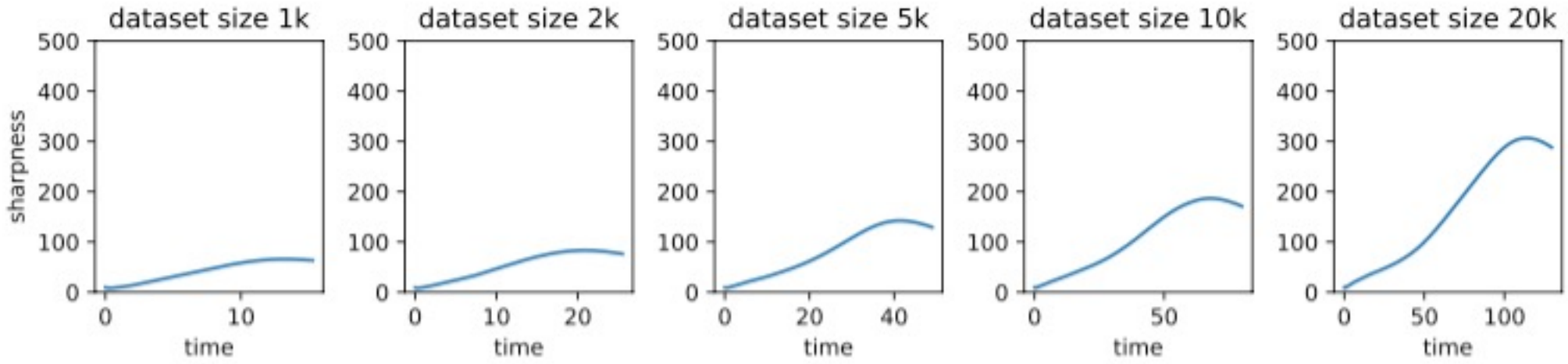


Figure 18: **The effect of dataset size**. We use gradient flow to train a network on varying-sized subsets of CIFAR-10. Observe that progressive sharpening occurs to a greater degree as the dataset size increases.

Width and Depth: for PS

- Sharpness increases as the network become narrower and deeper

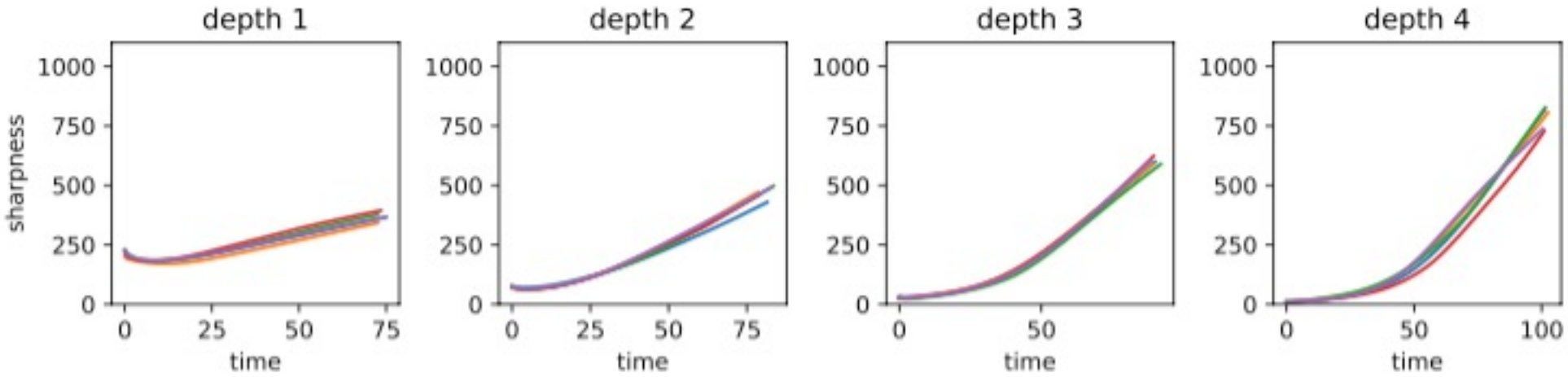


Figure 17: **The effect of depth: mean squared error**. We use gradient flow to train networks of various depths, ranging from 1 hidden layer to 4 hidden layers, using MSE loss. We train each network from five different random initializations (different colors). Observe that progressive sharpening occurs to a greater degree for deeper networks.

Width and Depth: for PS

- Sharpness increases as the network become narrower and deeper

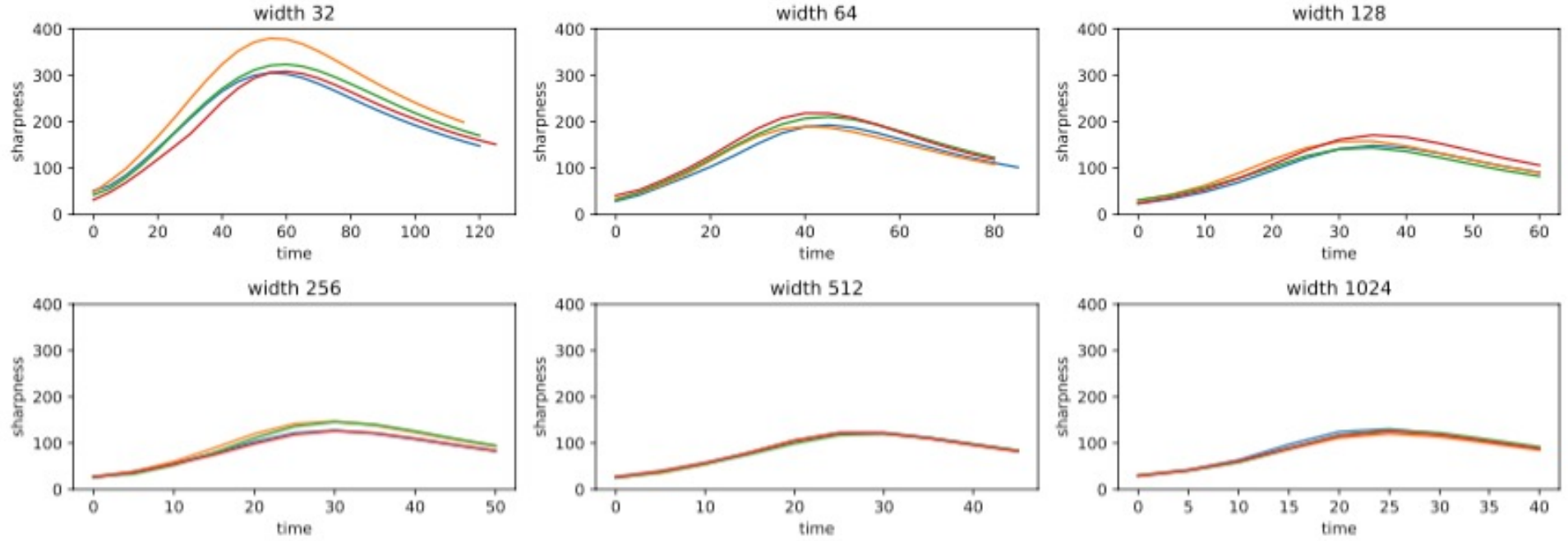
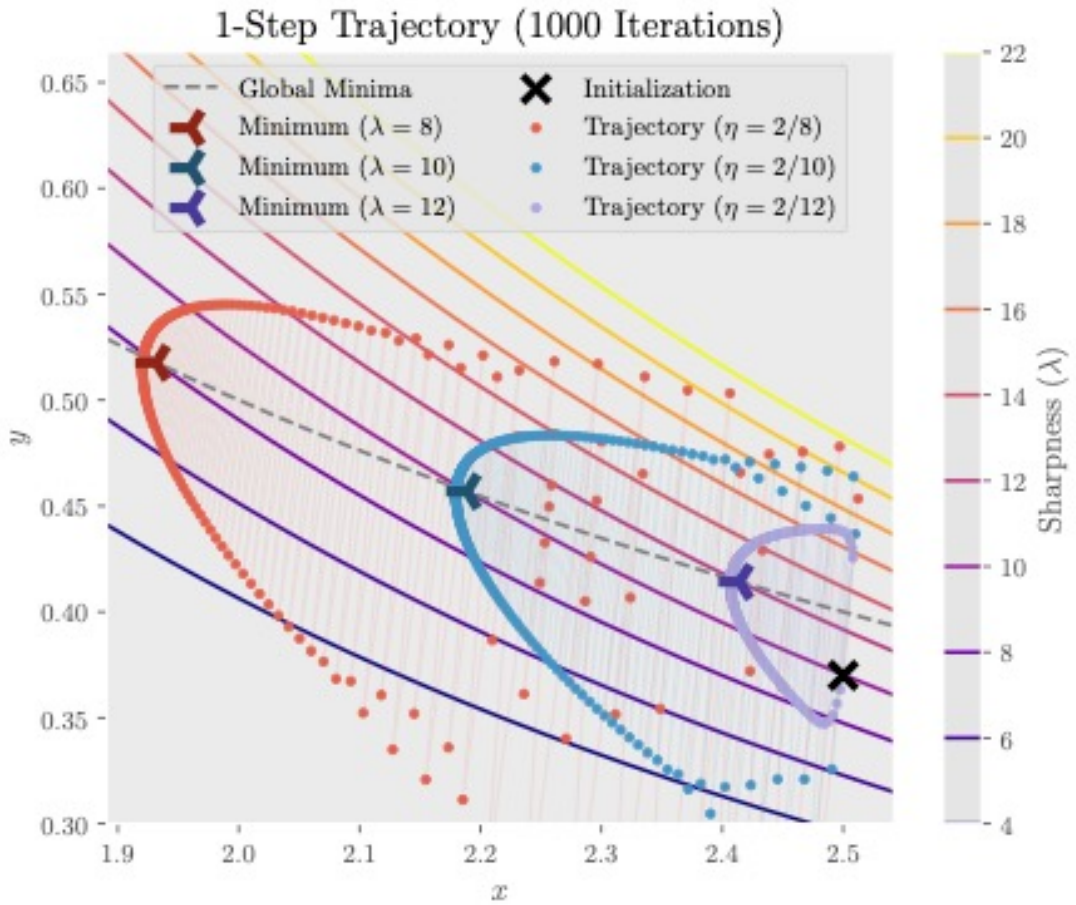
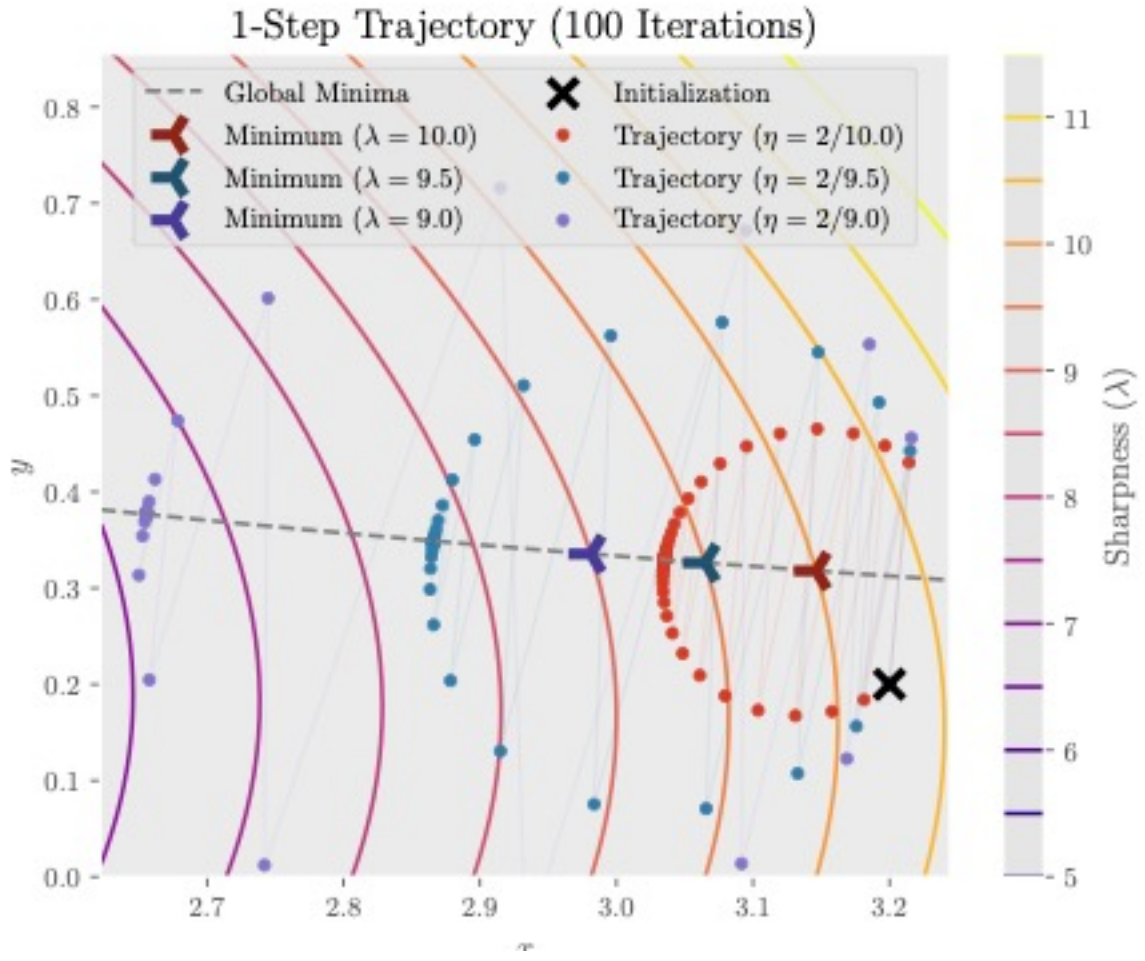


Figure 14: **Standard parameterization: evolution of the sharpness**. We use gradient flow to train standard-parameterized networks, and we track the evolution of the sharpness during training. For each width, we train from five different random initializations (different colors). Observe that the sharpness rises more when training narrow networks than when training wide networks.

Width and Depth: for EoS

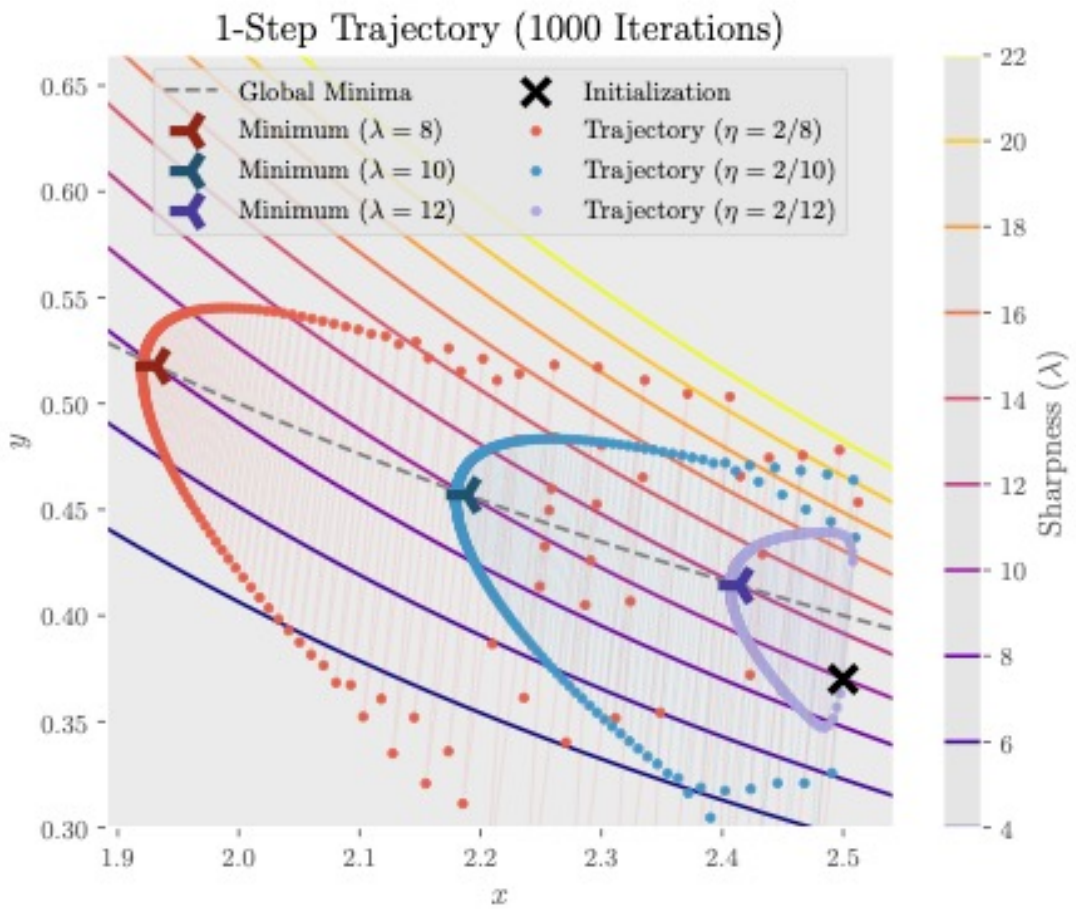
- Other work give a “simple” example that objective converges to some mimina flatter than EoS minima when more shallow. (inherently non-quadratic)

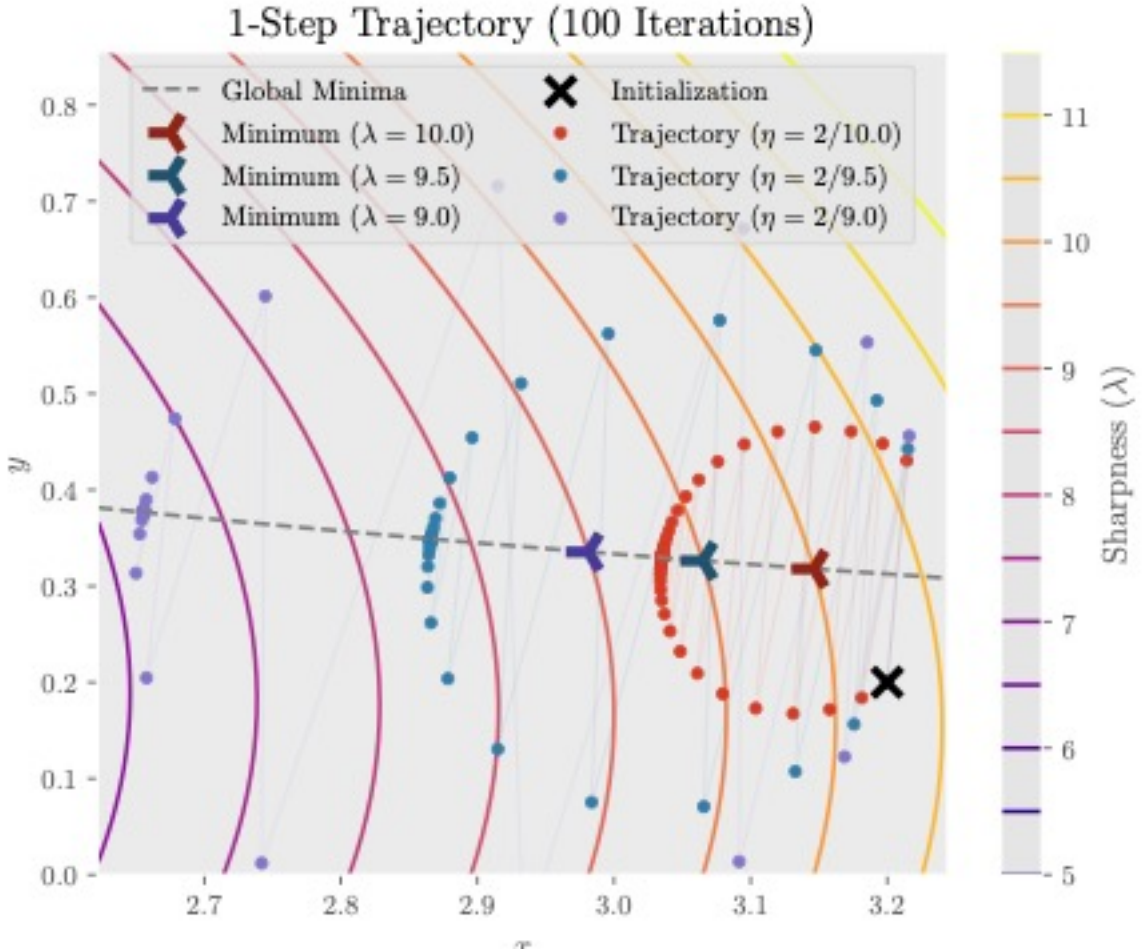




Width and Depth: for EoS

- The difference between (1-x^2y^2)^2 (left) and (1-xy)^2 (right)





EoS occurs at reasonable step size-1

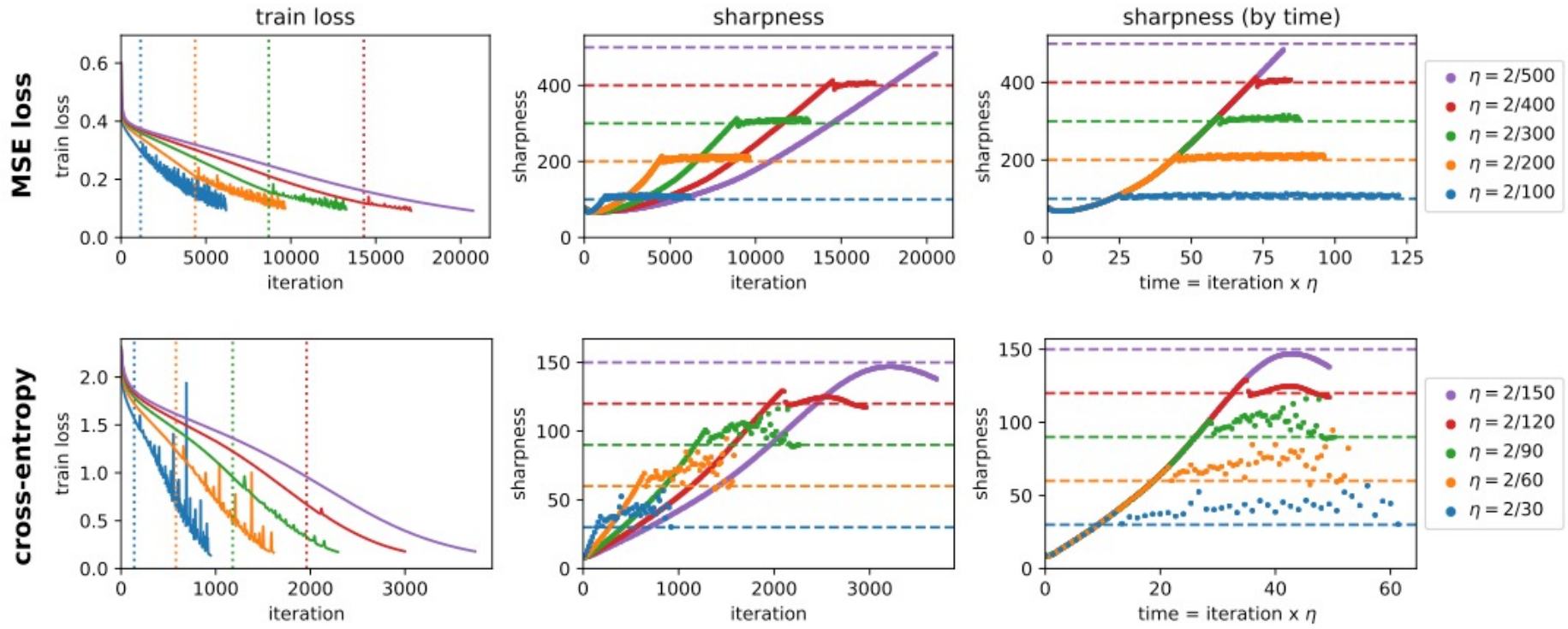


Figure 5: **After the sharpness reaches $2/\eta$, gradient descent enters the Edge of Stability**. A network is trained with gradient descent at a range of step sizes (see legend), using both MSE loss (**top row**) and cross-entropy (**bottom row**). **Left**: the train loss curves, with a vertical dotted line at the iteration where the sharpness first crosses $2/\eta$. **Center**: the sharpness, with a horizontal dashed line at the value $2/\eta$. **Right**: sharpness plotted by time (= iteration $\times$ η) rather than iteration.

EoS occurs at reasonable step size-2

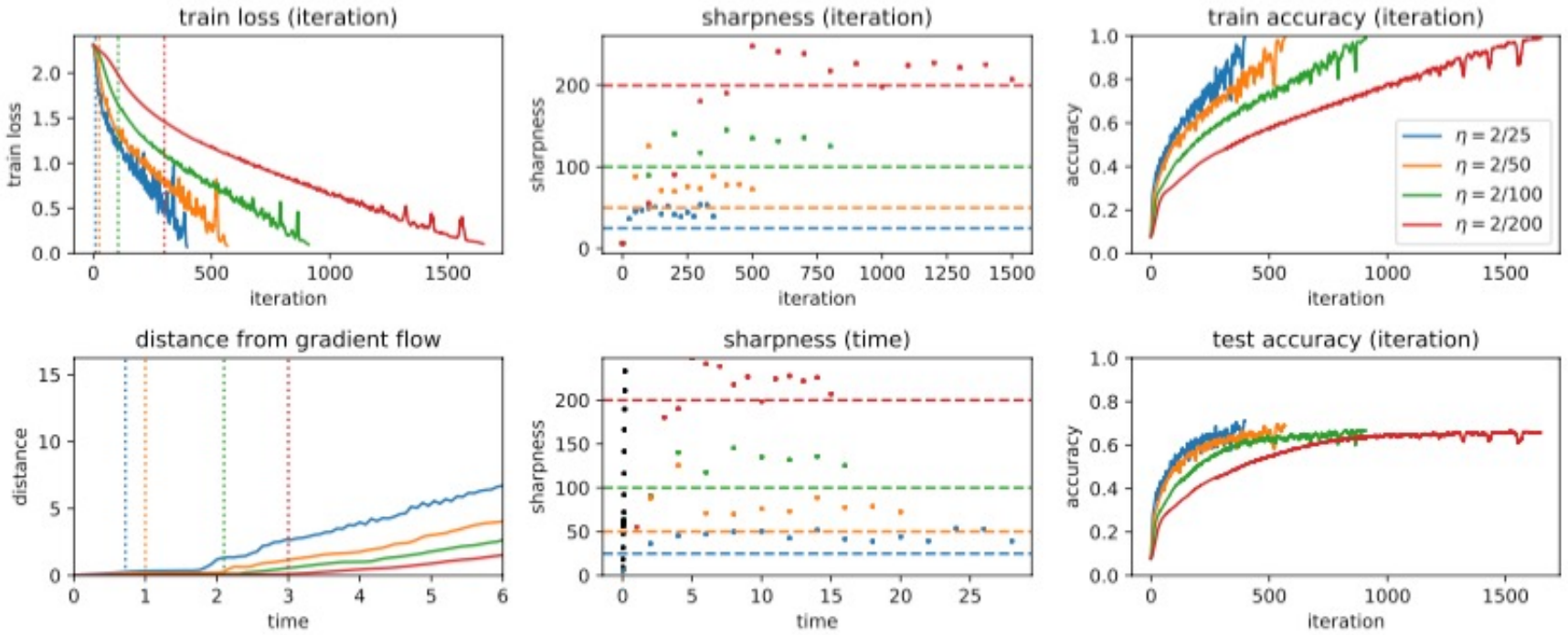


Figure 90: We train a **VGG with BN** to completion using **gradient descent** at different step sizes (see legend in the top right pane). **Top left**: we plot the train loss, with a vertical dotted line marking the iteration where the sharpness first crosses $2/\eta$. **Top center**: we plot the evolution of

EoS is more practical!

Implications

- Rethinking L-smoothness assumption
  - Following the optimization trajectory, the L-smoothness condition gradually breaks.
  - The non-monotonical loss decrease indicates the ``descent lemma'' does not hold, contradictory with L-smoothness
  - It (always) can be applied locally
- Only continuous time analysis (e.g. gradient flow) is not enough
  - Gradient flow will not perceive the $2/\eta$ threshold
  - The progressive sharpening continues
- The EoS is inherently non-quadratic

=> Technical difficulties for these inter-related phenomena !!

Outline

- Preliminaries
- What is EoS phenomena?
- When does EoS occur?
- **Why does EoS occur? (Examples and Intuition)**
- How about SGD?

Theory of EoS*

- Normalized GD
- A Minimalist Example
- Self-stabilization

Normalized GD

- Consider

Normalized GD with LR η by $x_\eta(t)$, with $x_\eta(0) \equiv x_{\text{init}}$ for all η:

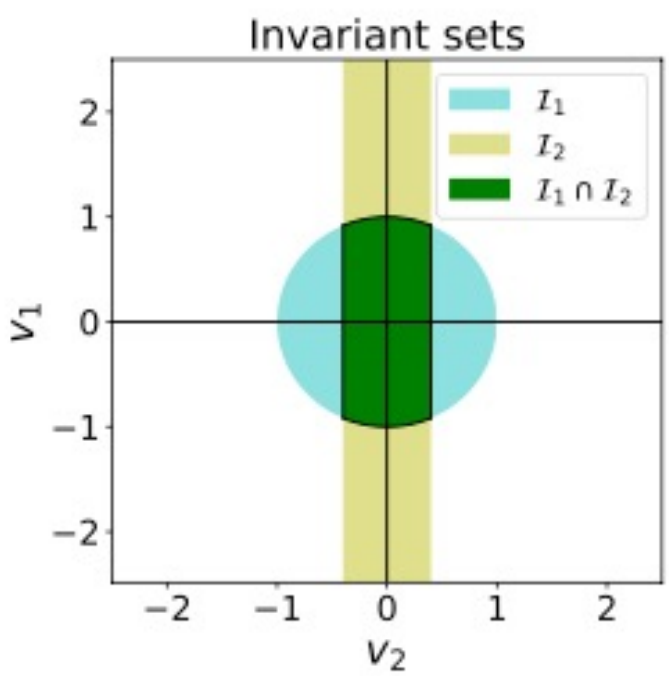
$$\textbf{Normalized GD:} \quad x_\eta(t+1) = x_\eta(t) - \eta \frac{\nabla L(x_\eta(t))}{\|\nabla L(x_\eta(t))\|}$$

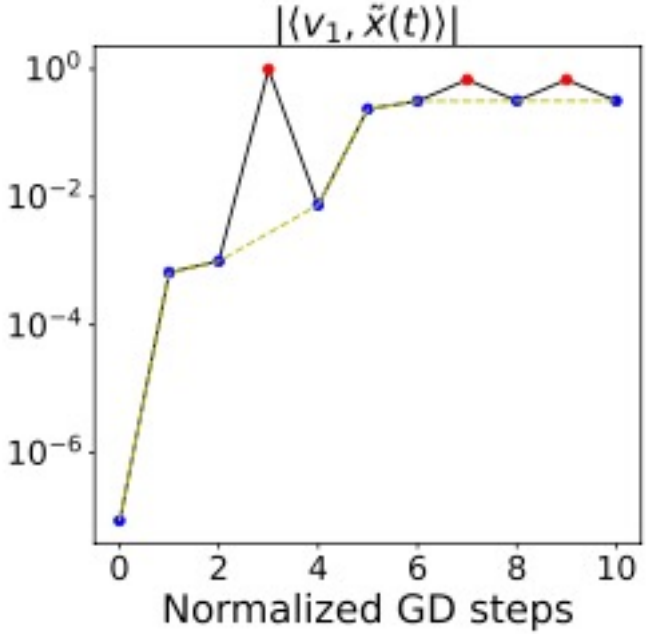
- The algorithm even cannot converge for the quadratics

Normalized GD: Two phase

- Phase I (Convergence/Preparation)
  - monotonical loss decrease till unstable
- Phase II (Alignment/Limit flow)
  - with top eigenvectors
  - sharpness reduction

(Fig for quadratics)





Normalized GD: Two phase

Cons

- close to the manifold of global minimizers
- unfixed step size
- noise injected
- tracking gradient flow?

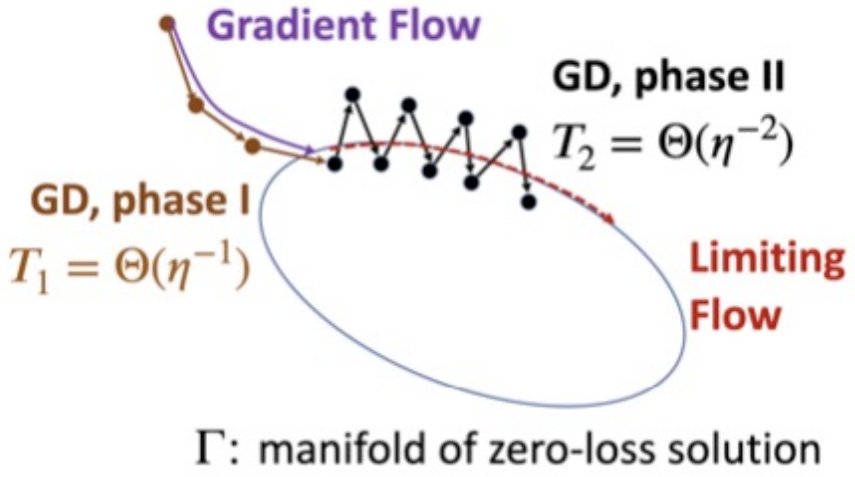


Figure 3: Illustration for two-phase dynamics of Normalized GD and GD on $\sqrt{L}$ on a 1D zero loss manifold Γ. For sufficiently small LR η, Phase I is close to Gradient Flow and lasts for $\Theta(\eta^{-1})$ steps, while Phase II is close to the limiting flow which decreases the sharpness of the loss and lasts for $\Theta(\eta^{-2})$ steps. GD iterate oscillates along the top eigenvector of the Hessian with the period equal to two steps. (cf. Figure 2 in [Li et al., 2022b])

A Minimalist Example

We focus on the simple objective $\mathcal{L}(x, y, z, w) \triangleq \frac{1}{2}(1 - xyzw)^2$. Let the learnable parameters $x, y, z, w \in \mathbb{R}$ to be trained using gradient descent with a fixed step size $\eta \in \mathbb{R}^+$ that

$$(x_{t+1}, y_{t+1}, z_{t+1}, w_{t+1}) = (x_t, y_t, z_t, w_t) - \eta \nabla \mathcal{L}(x_t, y_t, z_t, w_t). \tag{1}$$

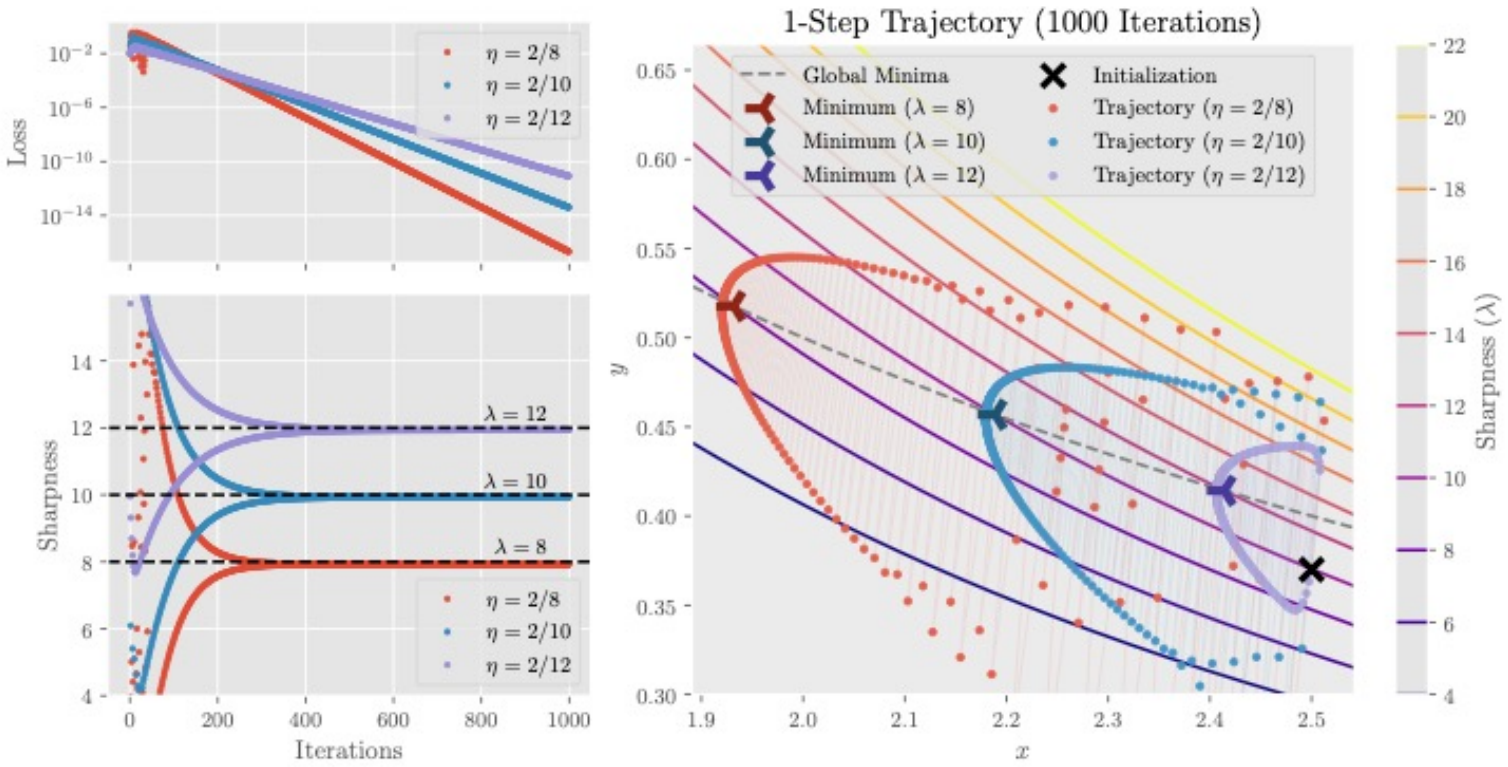
Here x_t denotes the value of parameter x after the t-th update. To further simplify the problem, we consider the symmetric initialization of $z_0 = x_0$, $w_0 = y_0$. Note that due to symmetry of objective, the identical entries will remain identical throughout the training process, so the training dynamics reduces to two dimensional and the 1-step update of x and y follows

$$x_{t+1} = x_t - x_t y_t^2 \eta (x_t^2 y_t^2 - 1), \quad y_{t+1} = y_t - x_t^2 y_t \eta (x_t^2 y_t^2 - 1). \tag{2}$$

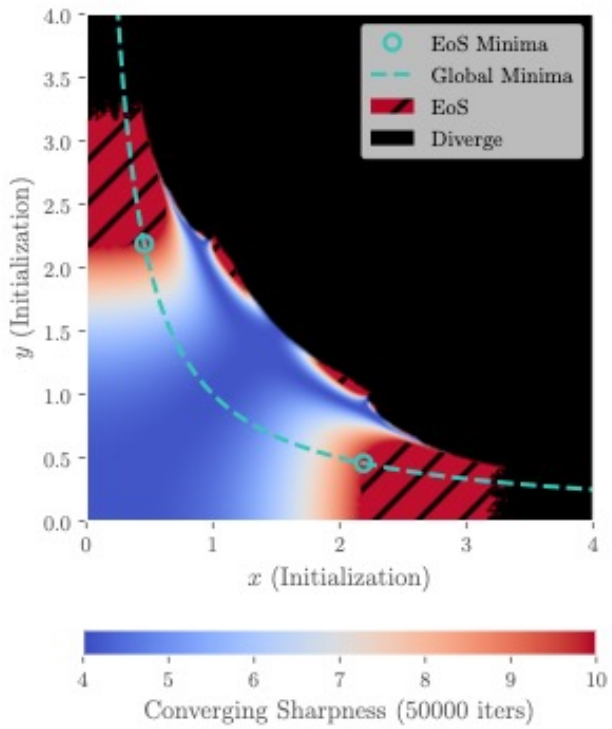
It's easy to show that the set of global minima for this function form the hyperbola $xy = 1$. Without

A Minimalist Example

$$(a, b) \triangleq \left(\left(x^2 - y^2\right)^{\frac{1}{2}} - \left(\eta^{-2} - 4\right)^{\frac{1}{4}}, xy - 1 \right).$$



(a) Evolution of training loss, sharpness, and trajectory of GD on the 4 scalar example from the same initialization with different learning rates.



(b) Initializations converging close to $\lambda = 2/\eta$ ($\eta = 0.2$)

A Minimalist Example

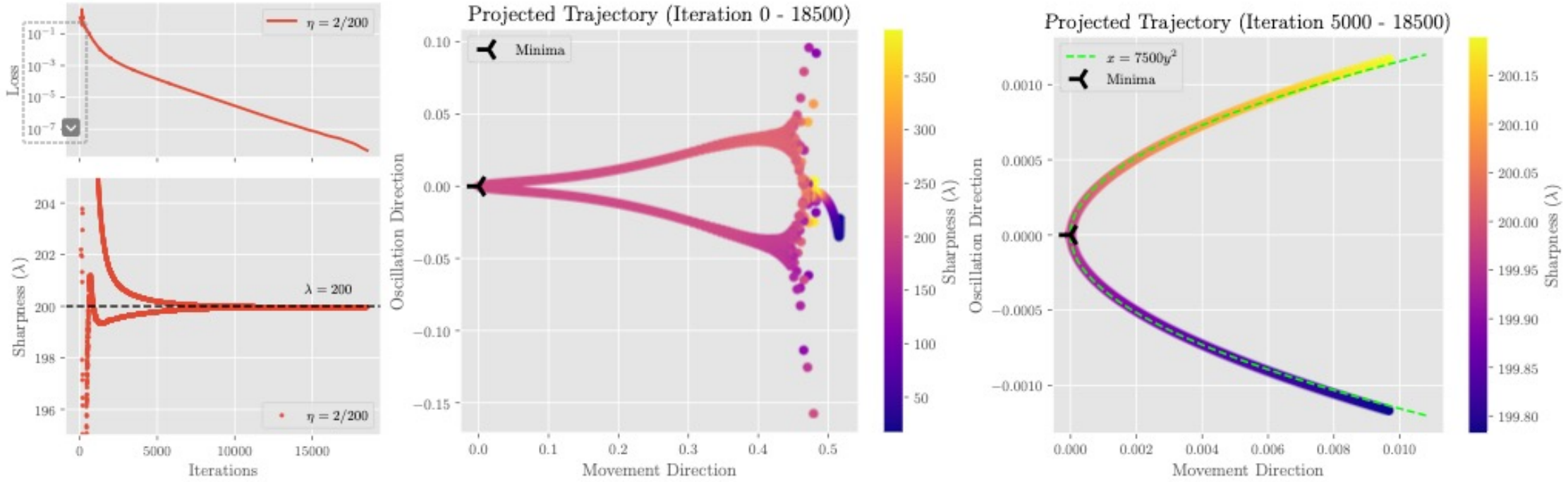


Figure 7: **Training trajectory of 5-layer ELU-activated FC Network**. We train the model using with $\eta = 0.01$ for 18500 iterations. The sharpness converges to 199.97 while $2/\eta = 200$. The local trajectory (right) can be very well approximated by the parabola $x = 7500y^2$.

A Minimalist Example

Cons

- No stable sharpening phase, (locally or intrinsically)
- Only for single simple example

Self-stabilization: sketch

On the unstable direction, we have

- $\nabla L(\theta + \alpha u) \approx \nabla L(\theta) + \alpha \nabla(\nabla L(\theta)^T u) + 0.5 \alpha^2 \nabla(u^T \nabla^2 L(\theta) u)$

where u is the top eigenvector

On the rest stable direction, the loss does decrease

Limitations

- Normalized GD
  - Not the true algorithm
  - Inject some noise the analysis
- A Minimalist Example
  - Analysis for a single degree-4 model
- Self-stabilization
  - The assumptions may not hold(from openreivew)

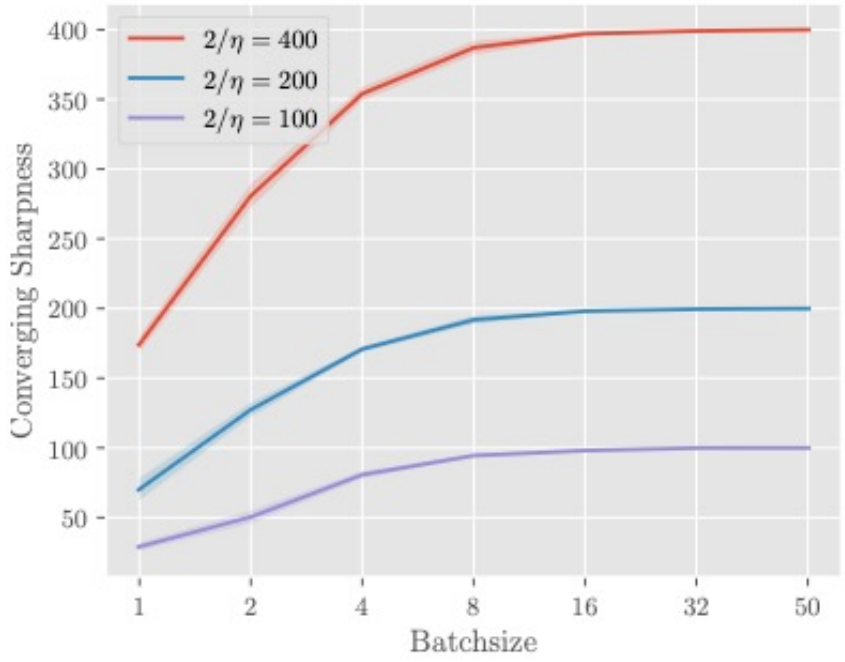
My understanding

- Locally
- Decomposition

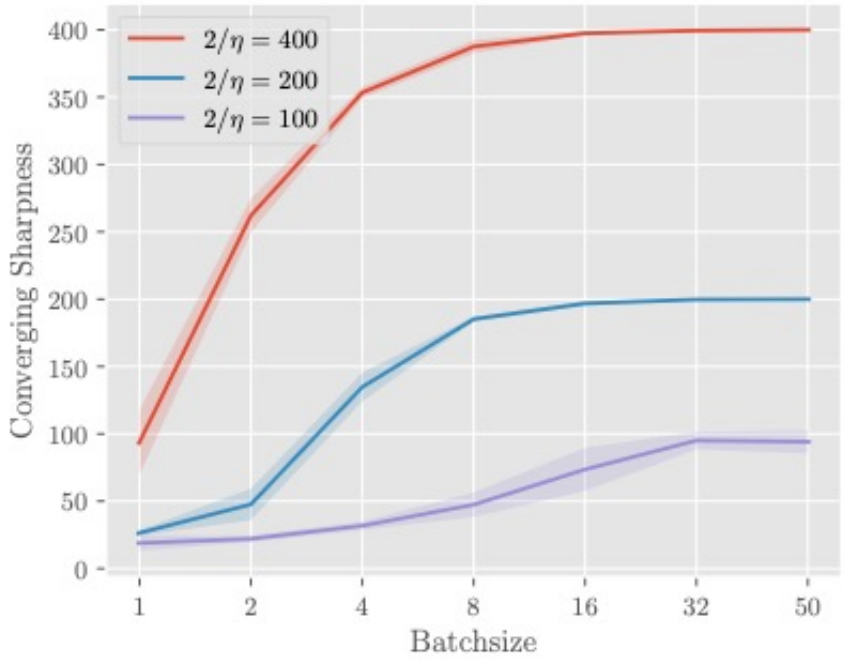
Outline

- Preliminaries
- What is EoS phenomena?
- When does EoS occur?
- Why does EoS occur?
- **How about SGD?**

EoS variant for SGD



(a) 5-layer FC network with ELU activation



(b) 5-layer FC network with tanh activation

gure 29: **FC networks trained with SGD of varying batchsizes.** ($\eta = 0.005, 0.01, 0.02$)

SGD is more picky

$$f_1(x) = \min\{x^2, 0.1(x-1)^2\},\ f_2(x) = \min\{x^2, 1.9(x-1)^2\}.$$

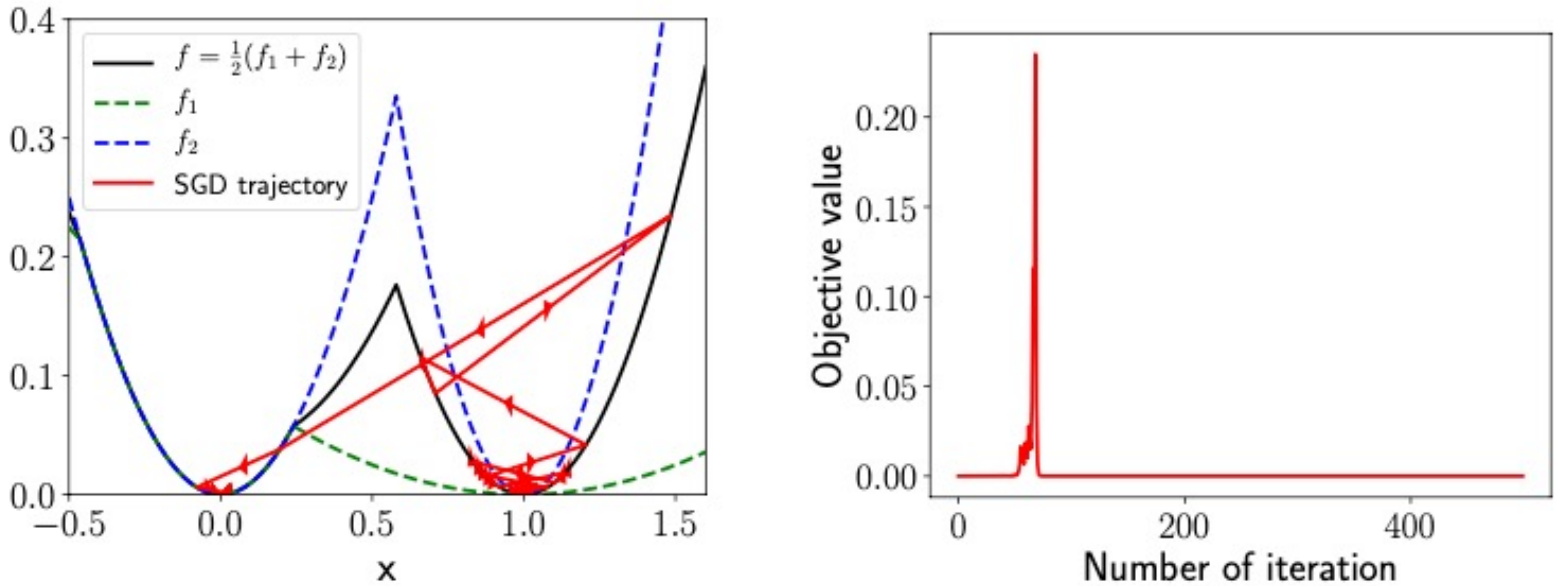


Figure 2: Motivating example. (Left) One trajectory of SGD with learning rate $\eta = 0.7, x_0 = 1 - 10^{-5}$, showing convergence to 0. GD with the same learning rate will converge to 1. (Right) The value of objective function, showing a burst during the escape.

Summary & Future

EoS

- What? When? Why?

Future

- How to give an accurate characterization for SGD
- How EoS helps us in practice
  - Adapt the step size
  - Find the minima with lower sharpness (for possible better generalization)

Thanks