Trajectory Alignment: Understanding the Edge of Stability Phenomenon via Bifurcation Theory

Minhak Song Chulhee Yun

KAIST

NeurIPS 2023

Poster: Session 3, Wed 13 Dec

OptiML
Optimization and Machine Learning Laboratory

Edge of Stability (EoS)

Full-batch GD: $\Theta_{t+1} = \Theta_t - \eta \nabla L(\Theta_t)$

Edge of Stability (EoS)

Full-batch GD: $\Theta_{t+1} = \Theta_t - \eta \nabla L(\Theta_t)$

Descent Lemma: If $\eta < \frac{2}{\lambda_{\max}(\nabla^2 L)}$, then loss drops each iteration

Edge of Stability (EoS)

Full-batch GD: $\Theta_{t+1} = \Theta_t - \eta \nabla L(\Theta_t)$

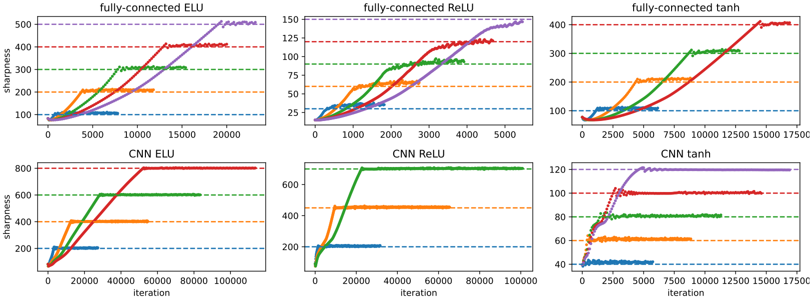
Descent Lemma: If $\eta < \frac{2}{\lambda_{\max}(\nabla^2 L)}$, then loss drops each iteration

Edge of Stability (EoS)

Full-batch GD: $\Theta_{t+1} = \Theta_t - \eta \nabla L(\Theta_t)$

Descent Lemma: If $\eta < \frac{2}{\lambda_{\max}(\nabla^2 L)}$, then loss drops each iteration

EoS: sharpness ($= \lambda_{\max}(\nabla^2 L)$) increases along GD trajectory then saturates at $2/\eta$ [Cohen et al., 2021]

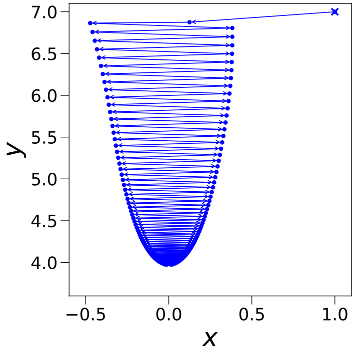


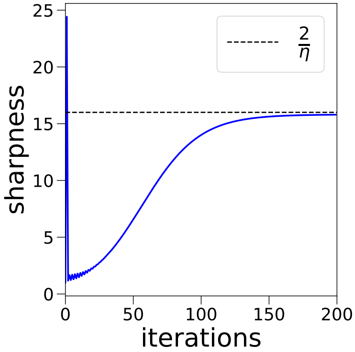
Toy model

Objective function: $L(x, y) = \log(\cosh(xy))$, step size: $\eta = 2/16$

Toy model

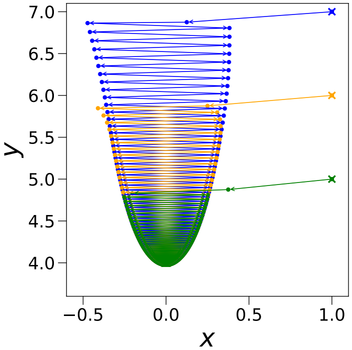
Objective function: $L(x, y) = \log(\cosh(xy))$, step size: $\eta = 2/16$

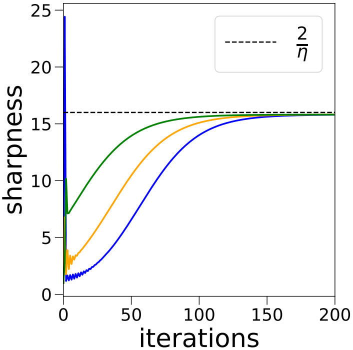




Toy model

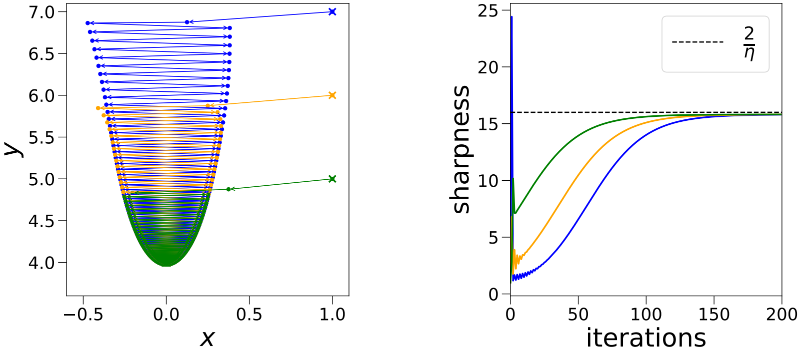
Objective function: $L(x, y) = \log(\cosh(xy))$, step size: $\eta = 2/16$





Toy model

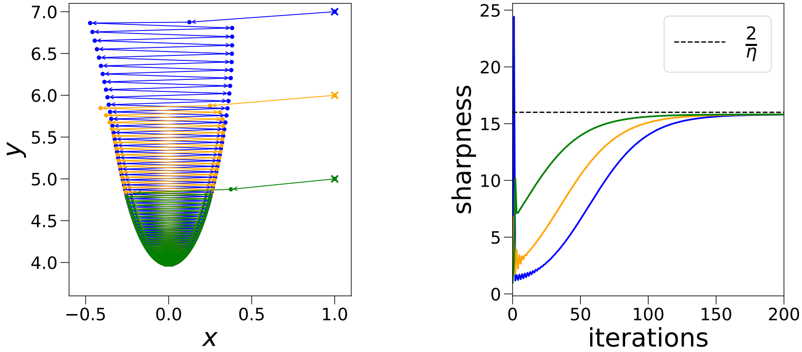
Objective function: $L(x, y) = \log(\cosh(xy))$, step size: $\eta = 2/16$



- Different GD trajectories align on the same curve: **Trajectory Alignment** occurs!

Toy model

Objective function: $L(x, y) = \log(\cosh(xy))$, step size: $\eta = 2/16$



- Different GD trajectories align on the same curve: **Trajectory Alignment** occurs!

Q. Trajectory alignment of GD in general setting?

Canonical reparameterization

Q. Trajectory alignment of GD in general setting?

Canonical reparameterization

Q. Trajectory alignment of GD in general setting?

$$\boxed{(p, q) \triangleq \left(\text{Residual}, \ \frac{\text{EoS threshold}}{\text{NTK sharpness}}\right)}$$

Canonical reparameterization

Q. Trajectory alignment of GD in general setting?

$$(p, q) \triangleq \left(\text{Residual}, \; \frac{\text{EoS threshold}}{\text{NTK sharpness}}\right)$$

Objective function: $L(\Theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i; \Theta) - y_i)$

Assumption: ℓ is convex, Lipschitz loss with $\ell'(0) = 0$, $\ell''(0) = 1$.

Canonical reparameterization

Q. Trajectory alignment of GD in general setting?

$$(p, q) \triangleq \left(\text{Residual}, \ \frac{\text{EoS threshold}}{\text{NTK sharpness}}\right)$$

Objective function: $L(\Theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i; \Theta) - y_i)$

Assumption: ℓ is convex, Lipschitz loss with $\ell'(0) = 0$, $\ell''(0) = 1$.

$$\begin{aligned}(p, q) &= \left(\frac{1}{n}\sum_{i=1}^{n}(f(x_i; \Theta) - y_i), \ \frac{2/\eta}{\lambda_{\max}(\mathrm{NTK})}\right) \\ &= \left(\frac{1}{n}\sum_{i=1}^{n}(f(x_i; \Theta) - y_i), \ \frac{2n}{\eta \| \sum_{i=1}^{n}(\nabla_\Theta f(x_i; \Theta))^{\otimes 2} \|_2^2}\right)\end{aligned}$$

Canonical reparameterization

Q. Trajectory alignment of GD in general setting?

$$(p, q) \triangleq \left(\text{Residual}, \ \frac{\text{EoS threshold}}{\text{NTK sharpness}}\right)$$

Objective function: $L(\Theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i; \Theta) - y_i)$

Assumption: ℓ is convex, Lipschitz loss with $\ell'(0) = 0$, $\ell''(0) = 1$.

$$\begin{aligned}(p, q) &= \left(\frac{1}{n} \sum_{i=1}^{n} (f(x_i; \Theta) - y_i), \ \frac{2/\eta}{\lambda_{\max}(\text{NTK})}\right) \\ &= \left(\frac{1}{n} \sum_{i=1}^{n} (f(x_i; \Theta) - y_i), \ \frac{2n}{\eta \| \sum_{i=1}^{n} (\nabla_\Theta f(x_i; \Theta))^{\otimes 2} \|_2^2}\right)\end{aligned}$$

- Minimum with sharpness $2/\eta$ corresponds to $(p, q) = (0, 1)$

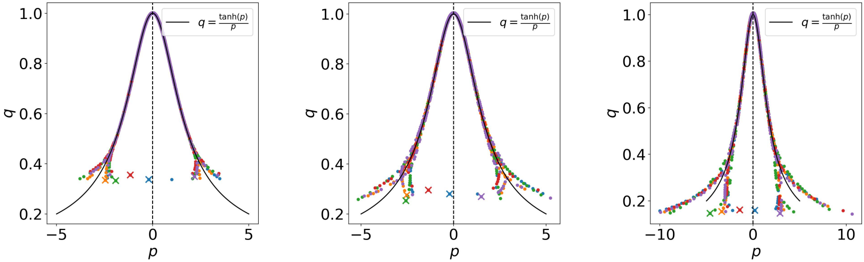
Experiment: single training point

Setting: log-cosh loss $\ell(p) = \log(\cosh(p))$, 3-layer FC networks

Experiment: single training point

Setting: log-cosh loss $\ell(p) = \log(\cosh(p))$, 3-layer FC networks

Observation: GD trajectories align on the curve $\boxed{q = \dfrac{\ell'(p)}{p}}$



(a) tanh FC network (b) ELU FC network (c) linear FC network

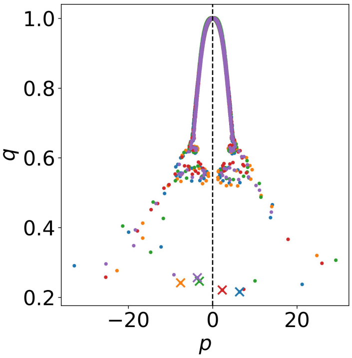
Experiment: multiple training points (CIFAR-10)

Setting: log-cosh loss $\ell(p) = \log(\cosh(p))$, CIFAR-10 2-class subset

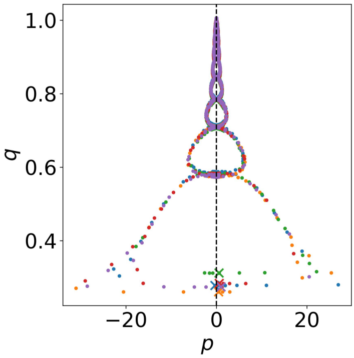
Experiment: multiple training points (CIFAR-10)

Setting: log-cosh loss $\ell(p) = \log(\cosh(p))$, CIFAR-10 2-class subset

Observation: GD trajectories align on the curve **independent of initialization**



(a) 3-layer MLP



(b) 3-layer CNN

Theory: Trajectory Alignment phenomenon provably occurs

Setting: training a two-layer linear network on a single data point

Theory: Trajectory Alignment phenomenon provably occurs

Setting: training a two-layer linear network on a single data point

Theorems 4.2 and 4.3 (informal, EoS regime)

If $q_0 < 1$, then there exists $t_a = O(\log(\eta^{-1}))$ such that for any $t \geq t_a$,

$$\boxed{\frac{q_t}{r(p_t)} = 1 + h(p_t)\eta^2 + O(\eta^4)},$$

where $h(p) \triangleq -\frac{1}{2}\left(\frac{pr(p)^3}{r'(p)} + p^2 r(p)^2\right)$ for $p \neq 0$ and $h(0) \triangleq -\frac{1}{2r''(0)}$.

Theory: Trajectory Alignment phenomenon provably occurs

Setting: training a two-layer linear network on a single data point

Theorems 4.2 and 4.3 (informal, EoS regime)

If $q_0 < 1$, then there exists $t_a = O(\log(\eta^{-1}))$ such that for any $t \geq t_a$,

$$\boxed{\frac{q_t}{r(p_t)} = 1 + h(p_t)\eta^2 + O(\eta^4)},$$

where $h(p) \triangleq -\frac{1}{2}\left(\frac{pr(p)^3}{r'(p)} + p^2 r(p)^2\right)$ for $p \neq 0$ and $h(0) \triangleq -\frac{1}{2r''(0)}$.
Moreover, (p_t, q_t) converge to the point $(0, q^*)$ such that

$$q^* = 1 - \frac{\eta^2}{2r''(0)} + O(\eta^4),$$

and the limiting sharpness is $\boxed{\frac{2}{\eta} - \frac{\eta}{|r''(0)|} + O(\eta^3)}$.

Summary

Summary

- We empirically demonstrate and provably establish the **trajectory alignment** phenomenon of GD in EoS regime.

Summary

- We empirically demonstrate and provably establish the **trajectory alignment** phenomenon of GD in EoS regime.
- Sheds light on the training dynamics of high-dimensional non-convex NN optimization using GD with large step size.

Summary

- We empirically demonstrate and provably establish the **trajectory alignment** phenomenon of GD in EoS regime.
- Sheds light on the training dynamics of high-dimensional non-convex NN optimization using GD with large step size.
- For more details, **join our poster session (Session 3, Wed 13 Dec)** or check our paper!

(openreview link)

References I

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.