arxiv: 2206.04172

NYU

Beyond the Edge of Stability via Two-step Gradient Updates

Lei Chen, Joan Bruna
New York University

I. Introduction

[Cohen2021]: GD is observed to still **converge** regardless of local instability, i.e., $\lambda_{\max} \approx 2/\eta$.

Question: why does GD not explode?

When optimizing the quadratic $f(x) = 0.5\lambda x^2$, GD explodes once $\eta > 2/\lambda$.

Ours: many problems allow **stable oscillations** around minima when $\eta > 2/\lambda$, including NNs.

Stable Oscillation (SO)

Definition Let $F_\eta : \Omega \to \Omega$ be GD with learning rate η for a function f. A period-2 stable oscillation is

1. $\exists x \in \Omega$, such that $F_\eta^2(x) \triangleq F_\eta(F_\eta(x)) = x$, and
2. x is not a minima of f.

Summary

In the setting of $\eta > 2/\lambda_{\max}(H(\bar{x}))$, we show

(i) Existence of SO and convergence on 1D functions,
(ii) Provable convergence on single-neuron ReLU net,
(iii) Observations of convergence on matrix factorization.

II. Stable Oscillation on Low-dim Functions

Existence: general 1D functions

Consider any 1D differentiable function $f(x)$ around a local minima $\bar{x}$, satisfying
(i) $f^{(3)}(\bar{x}) \neq 0$, and
(ii) $3[f^{(3)}]^2 - f'' f^{(4)} > 0$ at $\bar{x}$.
Then, there exists ϵ with sufficiently small $|\epsilon|$ and $\epsilon \cdot f^{(3)} > 0$ such that: for any point x_0 between $\bar{x}$ and $\bar{x} - \epsilon$, there exists a learning rate η such that $F_\eta^2(x_0) = x_0$, and

$$\frac{2}{f''(\bar{x})} < \eta < \frac{2}{f''(\bar{x}) - \epsilon \cdot f^{(3)}(\bar{x})}.$$

For a function f that the lowest order non-zero derivative (except the f'') at $\bar{x}$ is $f^{(k)}(\bar{x})$ with $k \geq 4$, the above conditions are changed to
(i) if k is odd and $\epsilon \cdot f^{(k)}(\bar{x}) > 0, f^{(k+1)}(\bar{x}) < 0$, or
(ii) if k is even and $f^{(k)}(\bar{x}) < 0$.

Existence: L_2 loss on general 1D functions

Base model: $g(x)$ Target value: y
Loss: $f(x) = (g(x) - y)^2$

From conditions on general 1-D functions, stable oscillation exists around $\bar{x} = g^{-1}(y)$ if
(i) $g'(\bar{x}) \neq 0$,
(ii) $g'(\bar{x}) g^{(3)}(\bar{x}) < 6[g''(\bar{x})]^2$.

Composition rule: if both $p(x), q(y)$ satisfy the above conditions at $x = \bar{x}, y = p(\bar{x})$, then $q(p(x))$ also satisfies the conditions to allow stable oscillation around $x = \bar{x}$.

Examples: *the base model g can be sin(x), tanh(x), high-order monomial, exp(x), log(x), sigmoid, gaussian...*

Convergence: a special 1D function

Loss: $f(x) = \frac{1}{4}(x^2 - 1)^2$ "*Symmetric scalar factorization*"
Learning rate: $1 < \eta < 1.121$
Initialization: any point $x_0 \in (0,1)$
Convergence: it converges to a period-2 orbit $\{x = \delta_i \,|\, i = 1,2\}$ where δ_1, δ_2 are the positive solutions of

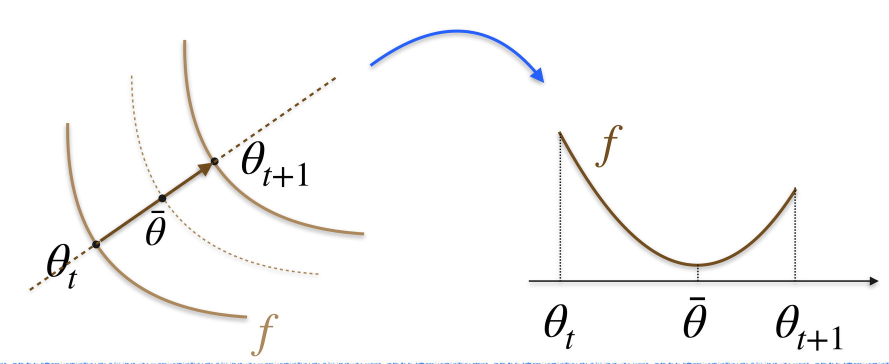
$$\eta = \frac{1}{\delta^2\left(\sqrt{\frac{1}{\delta^2} - \frac{3}{4}} + \frac{1}{2}\right)}.$$

Convergence: a special 2D function

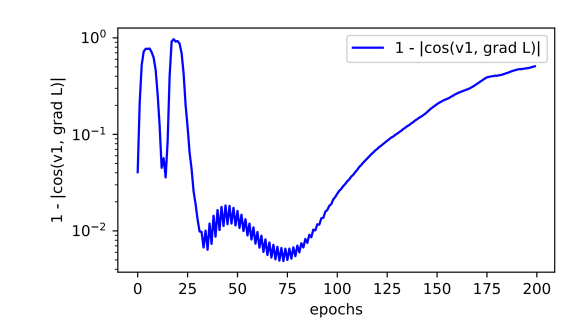
Loss: $f(x, y) = \frac{1}{2}(xy - 1)^2$ "*Asymmetric scalar factorization*"
Learning rate: $1 < \eta < 1.121$
Initialization: some conditions that guarantee $x, y > 0$ always
Convergence: it converges to a period-2 orbit as $\{(x = y = \delta_i) \,|\, i = 1,2\}$ where δ_1, δ_2 are the same as above.

Balancing effect: $|x - y| \to 0$ *despite of different init.*
***Previous** balancing effects:*
(i) [Du2018] GF: $x^2 - y^2$ remains unchanged.
(ii) [Wang2022] GD below EoS: $x^2 - y^2$ gets smaller, but not 0.

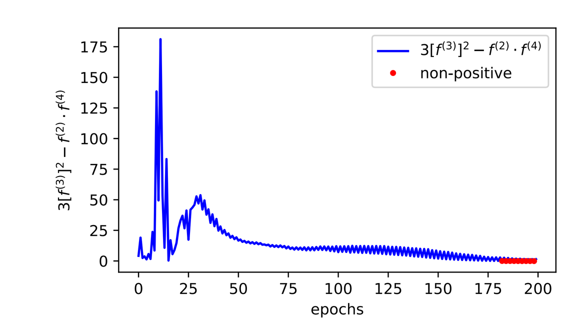
Experiments: *MLPs on MNIST*



Obs 1: v_1(Hessian) aligns with ∇Loss in direction



Obs 2: $3[f^{(3)}]^2 - f'' f^{(4)} > 0$ holds at $\bar{\theta}$



III. Case Study: Two-layer Single-neuron ReLU Network

Setting *Nonlinear*

(a) Student net: $f(x; \theta) = v \cdot \sigma(w^\top x), v \in \mathbb{R}$, $w, x \in \mathbb{R}^d$,
(b) Teacher model: $y \,|\, x = \sigma(\tilde{w}^\top x)$,
(c) Population loss: $L(\theta) = \mathbb{E}_{x \in \mathcal{S}^{d-1}}[f(x;\theta) - y\,|\,x]^2$.

Flattest minima

For any minimizer with $vw = \tilde{w}$, the largest eigenvalue of Hessian is

$$\lambda_1 = \frac{(\|w\| - v)^2 + 2\|\tilde{w}\|}{d} \geq 2\frac{\|\tilde{w}\|}{d}.$$

Sharpness at the flattest minima is $2\frac{\|\tilde{w}\|}{d}$
EoS learning rate is $\frac{d}{\|\tilde{w}\|}$

Convergence

For $\eta = K \cdot \frac{d}{\|\tilde{w}\|}$ with $K \in (1, 1.121)$, it converges to

1. **Directional alignment:**
$\text{proj}_{\tilde{w}_\perp} w \to 0$ as $\mathcal{O}\left((1 - 0.030K)^t\right)$
2. **Balancing effect:**
$\big| v - \|w\| \big| \to 0$
3. **Stable oscillation:** *Same as the 2-D case*
$v = \|w\|$ is in a period-2 orbit

References

[Cohen2021] Cohen et al., "*Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability*". ICLR, 2021
[Wang2022] Wang et al., "*Large Learning Rate Tames Homogeneity: Convergence and Balancing Effect*". ICLR, 2022
[Du2018] Du et al., "*Algorithmic Regularization in Learning Deep Homogeneous Models: Layers are Automatically Balanced*". NeurIPS 2018

IV. Case Study: Matrix Factorization

Setting *High-dim*

(a) Learnable weights: $\mathbf{Y}, \mathbf{Z} \in \mathbb{R}^{d \times d}$,
(b)Target: PSD $\mathbf{C} \in \mathbb{R}^{d \times d}$ with $\lambda_1 = 1$,
(c) Loss: $L(\mathbf{Y}, \mathbf{Z}) = \frac{1}{2}\|\mathbf{Y}\mathbf{Z}^\top - \mathbf{C}\|_F^2$.

1D condition at any minimizer

For any minimizer with $\mathbf{Y}\mathbf{Z}^\top = \mathbf{C}$, consider the 1D function L_Δ at the cross section of the loss landscape L and the leading eigen-direction Δ of Hessian.

L_Δ satisfies the 1D condition at the minimizer as

$$3[L_\Delta^{(3)}]^2 - L_\Delta^{(2)} L_\Delta^{(4)} > 0$$

MF allows stable oscillation in 1D subspace!

Convergence *(observations)*

For $\eta \in (1, 1.121)$ and $\eta(1 + \lambda_2) < 2$, it converges to

1. **Balancing effect:**

$$\mathbf{Y} = \delta_i u v^\top + \sum_{j=2}^{d} \sigma_{y,j} u_{y,j} v_{y,j}^\top$$

$$\mathbf{Z} = \delta_i u v^\top + \sum_{j=2}^{d} \sigma_{z,j} u_{z,j} v_{z,j}^\top$$

2. **Oscillation in 1D subspace:**

$$\mathbf{Y}\mathbf{Z}^\top - \mathbf{C} = (\delta_i^2 - 1) u u^\top$$