

# ON THE CONVERGENCE DIRECTION OF GRADIENT DESCENT

**Shuo Chen**

Institute of Information Science  
Beijing Jiaotong University  
Beijing, China  
schen1307@foxmail.com

**Xiaolong Li**

Institute of Information Science  
Beijing Jiaotong University  
Beijing, China  
lixl@bjtu.edu.cn

**Jiaying Peng**

School of Mathematical Science  
Capital Normal University  
Beijing, China  
jiayingpeng@cnu.edu.cn

**Yao Zhao**

Institute of Information Science  
Beijing Jiaotong University  
Beijing, China  
yzhao@bjtu.edu.cn

## ABSTRACT

Gradient descent (GD) is a fundamental optimization method in deep learning, yet its asymptotic directional properties remain less understood. In this paper, we prove that if GD converges, its trajectory either aligns toward a fixed direction or oscillates along a specific line. The fixed-direction convergence occurs under small learning rates, while the oscillatory convergence behavior emerges for large learning rates. This result offers a new lens for understanding long-term GD dynamics. Experimentally, we find that this directional convergence behavior also appears in stochastic gradient descent and Adam. Furthermore, we discuss how these theoretical findings regarding oscillatory convergence might offer a perspective on the sharpness dynamics observed in the Edge of Stability (EoS) regime. Our work provides both theoretical clarity and practical insight into the behavior of dynamics for multiple optimization methods as well as EoS.

## 1 INTRODUCTION

Gradient descent (GD) is one of the most extensively studied optimization algorithms. While classical analysis typically focuses on small learning rates, the precise structure of the GD trajectory remains less understood. Recent empirical lines of research, such as the observations regarding the *Edge of Stability* (EoS) (Cohen et al., 2021), highlight that GD can exhibit complex, non-monotonic dynamics where the loss fluctuates over short timescales. Motivated by these rich dynamical behaviors, this work focuses on the asymptotic directional properties of GD, investigating theoretically how the trajectory aligns or oscillates relative to the learning rate.

A common framework for analyzing GD assumes that the function to be minimized is convex and  $L$ -smooth (Nesterov, 2003). Formally, let  $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$  be a differentiable function defined on a domain  $D$ , satisfying the  $L$ -smooth condition, i.e., for all  $\mathbf{x}, \mathbf{y} \in D$ ,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|. \quad (1)$$

Conventionally, GD dynamics  $\{\mathbf{x}_k\}_{k \geq 0}$  is defined as,

$$\mathbf{x}_{k+1} = F(\mathbf{x}_k), \quad (2)$$

where

$$F(\mathbf{x}) = \mathbf{x} - \eta \nabla f(\mathbf{x}). \quad (3)$$

By the classical descent lemma (Ahn et al., 2022), this framework ensures stable convergence and allows for a well-established convergence rate, if the learning rate satisfies  $0 < \eta < 2/L$ . However, the condition on  $L$ -smooth equation 1 requires that the constant  $L$  is sufficiently large. As a result, this condition usually fails on neural networks.

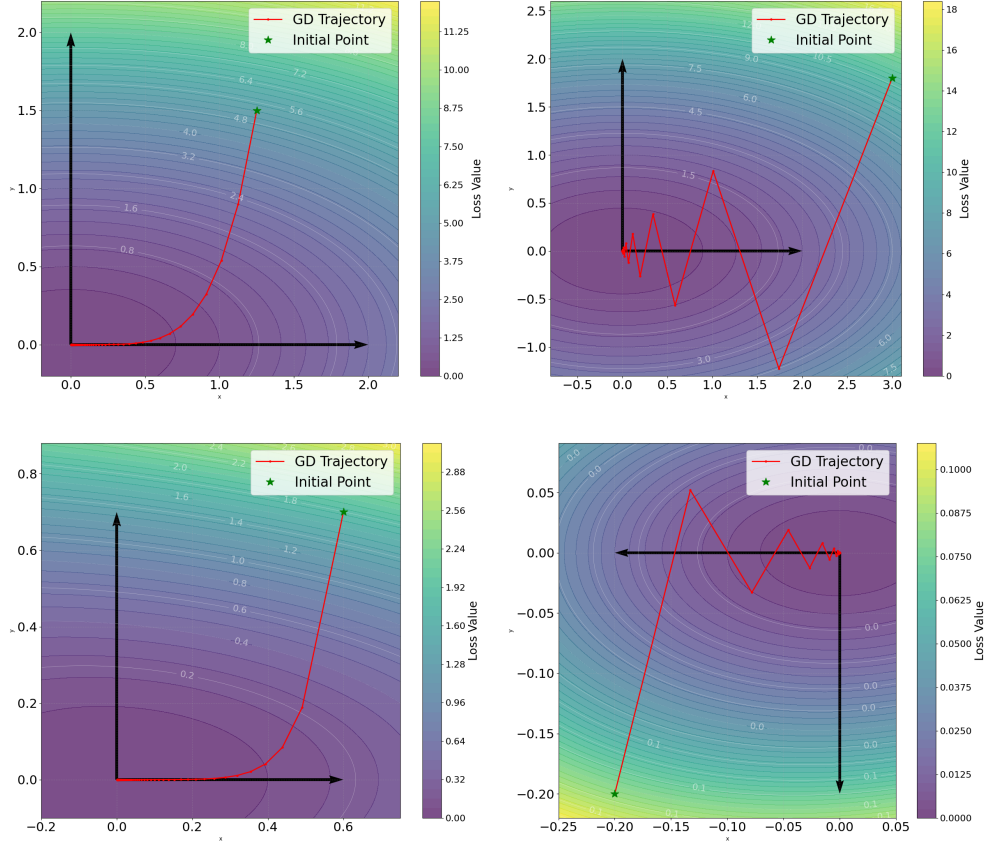


Figure 1: GD trajectories for  $f(x, y) = x^2/2 + 2y^2$  with  $\eta = 0.1$  (upper left, the convergence direction exists as  $x$ -axis) and  $\eta = 0.42$  (upper right, the convergence direction alternates along  $y$ -axis), and for  $f(x, y) = x^2/2 + 2y^2 + xy^2 + y^3$  with  $\eta = 0.1$  (lower left, the convergence direction exists as  $x$ -axis) and  $\eta = 0.42$  (lower right, the convergence direction alternates along  $y$ -axis).

Moreover, GD usually requires a learning rate satisfying  $0 < \eta < 2/\lambda_n$  for convergence (Ahn et al., 2022), where  $\lambda_n$  is the largest eigenvalue of the Hessian matrix at the local minimum. On the other hand, empirical results show that GD often continues to make progress even when the learning rate exceeds this bound at the first several GD iterations. More specifically,  $2/\eta$  is smaller than the sharpness of  $x_k$  for the first several  $x_k$ 's in the GD trajectory. Here, the sharpness is defined as the largest eigenvalue of the Hessian matrix at  $x_k$ . This observation raises fundamental questions about the underlying mechanisms that govern GD's long-term behavior and convergence. Many works have been devoted to analyzing the GD dynamics from the perspective of stability, loss landscape, and optimization trajectories (Ahn et al., 2022; Zhu et al., 2023; Damian et al., 2023; Lee & Jang, 2023; Chen et al., 2024; Cai et al., 2024).

In this work, the convergence direction of GD is theoretically investigated. Specifically, depending on the learning rate, we prove that, if convergent, the GD trajectory either aligns with a fixed direction or oscillates along a specific line. To illustrate our idea, consider a simple convex quadratic function  $f(x, y) = x^2/2 + 2y^2$ , in which  $(0, 0)$  is the unique global minimum. In this case, the GD dynamics are given by  $x_k = (1 - \eta)^k x_0$  and  $y_k = (1 - 4\eta)^k y_0$ . Consider then the convergence direction, i.e., the limit of normalized vector

$$\mathbf{v}_k \triangleq \frac{(x_k, y_k)}{\|(x_k, y_k)\|}. \quad (4)$$

We have, for  $x_0 \neq 0$  and  $y_0 \neq 0$ ,  $\lim_{k \rightarrow \infty} \mathbf{v}_k = (\text{sign}(x_0), 0)$  if  $0 < \eta < 2/5$ , and  $\lim_{k \rightarrow \infty} (-1)^k \mathbf{v}_k = (0, \text{sign}(y_0))$  if  $2/5 < \eta < 1/2$  (see Figure 1 for illustration). Moreover, such phenomenon also holds if small disturbance exists. For example, consider a quadratic function

with high-order disturbance,  $f(x, y) = x^2/2 + 2y^2 + xy^2 + y^3$ , The similar results arise up under the similar settings (see Figure 1 as well).

This result suggests that GD iterates align with a particular direction when approaching the minimum, rather than converging in arbitrary directions. Our analysis generalizes this observation by characterizing the exact convergence direction of the GD trajectory, providing a refined understanding of GD’s asymptotic behavior.

Our key contributions are summarized as follows:

- **Convergence direction of GD:** We show that when GD converges, its trajectory admits a convergence direction that depends on the learning rate  $\eta$ . Specifically, if  $0 < \eta < 2/(\lambda_1 + \lambda_n)$ , the GD trajectory aligns toward a fixed direction. In contrast, if  $2/(\lambda_1 + \lambda_n) < \eta < 2/\lambda_n$ , the GD trajectory exhibits directional oscillations along a specific line. Here,  $\lambda_1$  and  $\lambda_n$  are the smallest and largest eigenvalues of the Hessian matrix at the local minimum respectively.
- **Insights for modern optimization methods:** We observe that popular optimization algorithms such as stochastic gradient descent (SGD) and Adam also exhibit similar alignment behaviors in practice. Our findings may inform the design of new optimization algorithms that explicitly exploit this directional alignment.

## 2 RELATED WORKS

A central theoretical foundation of our work builds on the celebrated proof of the *Gradient Conjecture*, originally posed by René Thom and proven in Parusinski et al. (2000). The conjecture concerns the behavior of gradient flow trajectories near critical points of real analytic functions. Specifically, consider a trajectory  $\mathbf{x}(t)$  of the gradient vector field of a real analytic function  $f$ , with  $\mathbf{x}(t) \rightarrow \mathbf{x}_0$  as  $t \rightarrow \infty$ . The Gradient Conjecture asserts that the normalized secants,  $\lim_{t \rightarrow \infty} \frac{\mathbf{x}(t) - \mathbf{x}_0}{\|\mathbf{x}(t) - \mathbf{x}_0\|}$ , converge. Actually, Parusinski et al. (2000) not only prove this conjecture but establishes a stronger statement: the radial projection of the trajectory onto the unit sphere has finite length. Their argument leverages Łojasiewicz inequalities, asymptotic critical values, and the construction of control functions. Notably, functions of the form  $g(\mathbf{x}) = f(\mathbf{x})/r^l - a - r^\alpha$  grow fast enough along trajectories to guarantee directional convergence, where  $r = \|\mathbf{x}\|$ ,  $\alpha$  is a small positive constant,  $l$  is a rational number, and  $a$  is a negative constant. In Xing et al. (2018), the authors analyze the trajectory of SGD in neural networks and observe that consecutive gradient steps tend to follow a valley-like structure in the loss landscape, often moving in directions that alternate or oscillate. Their results highlight that the directionality of gradient steps carries meaningful geometric information about the optimization path, even across non-convex and noisy settings. In Morchdi et al. (2023), the authors investigate the phenomenon of gradient oscillation in neural network training by analyzing the correlation between consecutive gradient directions. They observe that a significant portion of optimization progress occurs during periods when consecutive gradient steps exhibit strong negative correlation, highlighting the complex and often non-monotonic behavior of gradient directions in practical training dynamics. Inspired by these works, we try to extend the continuous-time Gradient Conjecture to GD. We show that if convergent, the GD iterates exhibit analogous directional behavior, which either converge to a fixed direction or oscillate along a line. This connection provides a bridge between the geometry of continuous gradient flow and discrete optimization dynamics.

The phenomenon of EoS is an interesting topic in understanding GD dynamics. In Cohen et al. (2021), the authors first observe that GD frequently operates in a regime where the largest eigenvalue of the Hessian hovers around the critical threshold  $2/\eta$ , challenging traditional stability analyses. This discovery motivates further research into how optimization methods interact with sharpness and stability. A key extension of EoS appears in Sharpness-Aware Minimization (SAM) (Long & Bartlett, 2024). In that work, the authors show that SAM dynamically adjusts the Hessian norm via the gradient and perturbation radius, stabilizing training and promoting flatter minima. Moreover, Dai et al. (2024) discusses the impact of the batch normalization layer toward EoS. Recent work by Grimmer (2024) highlights that GD with periodically long steps can achieve provably faster convergence, despite violating descent at individual iterations.

Other works explore GD or SGD dynamics in multiple regimes under large learning rates (Hoffer et al., 2017; Li et al., 2019; Wu et al., 2023). Related studies examine curvature-aware learning rates

(Thomas et al., 2020; Wang & Ma, 2024; Cohen et al., 2024), and nonmonotone line search method (Fox et al., 2024), which adaptively adjust learning rates near the EoS without manual tuning. The works (Ma et al., 2022; Lee & Jang, 2023) further show that loss landscapes are subquadratic near minima and propose interaction-aware sharpness measures for mini-batch settings. At the same time, using momentum may modify GD’s EoS behavior. In Phunyaphibarn et al. (2024), the authors show that Polyak’s momentum leads to sharp curvature drops (“catapults”) and raises the maximum stable sharpness, stabilizing training while promoting flatter minima. Another important direction is parameter averaging, and Nitanda et al. (2024) proves that averaged SGD smooths noisy updates and implicitly biases toward wider minima. More research on implicit bias can be found in Arora et al. (2022); Nacson et al. (2023). This complements EoS findings, as averaging helps mitigate the instability caused by large learning rates and sharp directions. Recent work also links EoS with multi-phase dynamics. For instance, Wang et al. (2022); Cai et al. (2024) observe that large learning rates in neural networks may induce the EoS phase and gradually more stable phases. In Yang et al. (2024), the authors show that untuned SGD can achieve optimal rates but is sensitive to unknown smoothness constants, leading to gradient explosions. In Wang & Ma (2024), the authors identify dynamic transitions in gradient flow training, suggesting EoS interacts with deeper phase structures in training. More studies on dynamic of different phases can be found in Damian et al. (2023).

Finally, the role of noise in SGD ties to EoS. For instance, Mulayoff et al. (2021) links large learning rates with smoother learned functions and shows that stable solutions depend on depth, supporting the regularization view of EoS. More research on the connection of geometry and dynamics can be found in Lee et al. (2016); Zhu et al. (2019); Martens (2020); Wang & Wu (2023). Overall, these studies show that normalization, momentum, averaging, and noise all shape the dynamics of training near EoS, highlighting the role of instability-driven behavior in modern optimization.

### 3 MAIN RESULTS

Consider a loss function  $f \in C^3(\mathbb{R}^n)$  and its GD dynamics defined in equation 2 and equation 3 with learning rate  $\eta$ . For an isolated local minimum  $\mathbf{x}^*$  of  $f$ , consider the GD trajectory  $\{\mathbf{x}_k\}_{k \geq 0}$  converging to  $\mathbf{x}^*$ . Specifically, we define  $V_\eta = \{\mathbf{x}_0 \in \mathbb{R}^n : \lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{x}^*\}$ . Moreover, assume that for any subset  $W \subset \mathbb{R}^n$  with zero measure,  $F^{-1}(W)$  also has zero measure, where  $F$  is defined in equation 2. The zero measure condition of GD dynamics is a conventional assumption for technical reasons (Ahn et al., 2022; Chen et al., 2024). Then, our main result can be summarized as the following theorem.

**Theorem 1:** For the above defined function  $f$ , suppose the eigenvalues  $\{\lambda_1, \dots, \lambda_n\}$  of  $\nabla^2 f(\mathbf{x}^*)$  satisfy  $0 < \lambda_1 < \lambda_2 \leq \dots \leq \lambda_{n-1} < \lambda_n$ , for almost all initial points  $\mathbf{x}_0 \in V_\eta$ ,

- If  $0 < \eta < 2/(\lambda_1 + \lambda_n)$ , the convergence direction of GD exists, i.e.,  $\lim_{k \rightarrow \infty} \frac{\mathbf{x}_k - \mathbf{x}^*}{\|\mathbf{x}_k - \mathbf{x}^*\|}$  exists.
- If  $2/(\lambda_1 + \lambda_n) < \eta < 2/\lambda_n$ , the alternative convergence direction of GD exists, i.e.,  $\lim_{k \rightarrow \infty} (-1)^k \frac{\mathbf{x}_k - \mathbf{x}^*}{\|\mathbf{x}_k - \mathbf{x}^*\|}$  exists.

In this theorem, the condition for eigenvalues means that  $f$  is locally a strong convex function at  $\mathbf{x}^*$ . Moreover, for given  $\eta > 0$ , the initial point set  $V_\eta$  is an open subset of  $\mathbb{R}^n$  (Chen et al., 2024). Next, we proceed to sketch the proof and the complete proof is provided in the appendices.

#### 3.1 PROOF SKETCH WITH $0 < \eta < 2/(\lambda_1 + \lambda_n)$

Without loss of generality, we assume  $\mathbf{x}^* = \mathbf{0}$ . Furthermore, we may assume that the Hessian matrix  $\nabla^2 f(\mathbf{0})$  is diagonal (see Appendix A for details), i.e.,  $\nabla^2 f(\mathbf{0}) = \text{diag}(\lambda_1, \dots, \lambda_n)$ . Define

$$a = 1 - \eta\lambda_1, \quad b = \max_{2 \leq i \leq n} \{1 - \eta\lambda_i\}. \quad (5)$$

Since  $\eta < 2/(\lambda_1 + \lambda_n) < 2/2\lambda_1 = 1/\lambda_1$ , it follows that  $a > b \geq 0$ . Let  $x_{k,i}$  be the  $i$ -th component of  $\mathbf{x}_k$ , then the GD iteration can be expressed componentwisely as, for each  $1 \leq i \leq n$ ,

$$x_{k+1,i} = x_{k,i} - \eta \partial_i f(\mathbf{x}_k) = (1 - \eta\lambda_i)x_{k,i} + g_i(\mathbf{x}_k), \quad (6)$$

where  $g_i$  is defined by, for  $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ ,

$$g_i(\mathbf{x}) = \eta(\lambda_i x_i - \partial_i f(\mathbf{x})). \quad (7)$$



For simplicity, denote  $g_{k,i} = g_i(\mathbf{x}_k)$  throughout the following proof. Clearly,  $g_i \in \mathcal{C}^2(\mathbb{R}^n)$  and thus  $g_{k,i}$  is a twice differentiable function with respect to  $\mathbf{x}_k$ . Moreover, by GD iteration equation 2,  $g_{k,i}$  can be regarded as a twice differentiable function of  $\mathbf{x}_0$ . By the above preparation, we first establish a forward-invariant set for GD and provide key estimates for  $g_{k,i}$ . These results are summarized in the lemma below.

**Lemma 1:** For the function  $f$  defined in Theorem 1, there exists an open set  $\Omega \subset V_\eta$  containing  $\mathbf{x}^* = \mathbf{0}$  and a constant  $C > 0$  such that for any initial point  $\mathbf{x}_0 \in \Omega$ , the following properties hold,

- For all  $k \geq 0$ ,  $\mathbf{x}_k \in \Omega$ .
- For all  $k \geq 0$  and  $1 \leq i \leq n$ ,  $|g_{k,i}| \leq C\|\mathbf{x}_k\|^2$ .
- For all  $k \geq 0$  and  $1 \leq i, j \leq n$ ,  $\left| \frac{\partial g_{k,i}}{\partial x_{k,j}} \right| \leq C\|\mathbf{x}_k\|$ .

The proof of Lemma 1 draws upon the main theorem derived in Chen et al. (2024), which is stated below for clarity.

**Theorem 2:** [Theorem 2, Chen et al. (2024)] Let  $f \in \mathcal{C}^2(\mathbb{R}^n)$  and  $\mathbf{x}^*$  is a local minimum. Suppose that there exists  $r > 0$  such that  $f(\mathbf{x}) > f(\mathbf{x}^*)$  holds for any  $\mathbf{x}$  satisfying  $\|\mathbf{x} - \mathbf{x}^*\| = r$ . Moreover, assume that  $\nabla f(\mathbf{x}) \neq 0$  for all  $\mathbf{x} \in \overline{B(\mathbf{x}^*, r)} \setminus \{\mathbf{x}^*\}$ . If  $0 < \eta < 2/\lambda_n$ , where  $\lambda_n$  is the largest eigenvalue of  $\nabla^2 f(\mathbf{x}^*)$ , then there exists an open set  $U \subset B(\mathbf{x}^*, r)$  containing  $\mathbf{x}^*$  such that the following forward-invariance property holds,

$$\mathbf{x} \in U \implies F(\mathbf{x}) = \mathbf{x} - \eta \nabla f(\mathbf{x}) \in U. \quad (8)$$

Here,  $B(\mathbf{x}^*, r)$  is the open ball centered at  $\mathbf{x}^*$  with radius  $r$ , and  $\overline{B(\mathbf{x}^*, r)}$  is its closure. The forward-invariance is one of the most important properties for GD, which states that once the GD trajectory falls in to the forward-invariance set  $U$ , it will never escape from it.

Noticed that Lemma 1 remains valid for all  $0 < \eta < 2/\lambda_n$ , and it will be utilized in the proof for both cases of small or large learning rates. Its proof is detailed in Appendix B. Next, leveraging  $\Omega$  and  $C$  from Lemma 1, we select a constant  $\varepsilon > 0$  satisfying

$$B(\mathbf{0}, \varepsilon) \subset \Omega, \quad C\varepsilon + (a + nC\varepsilon)^2 < a, \quad \varepsilon \leq \frac{a-b}{3n^2C}. \quad (9)$$

With this choice, we first prove that the Theorem 1 holds for almost all initial points  $\mathbf{x}_0 \in B(\mathbf{0}, \varepsilon)$ . We now provide the following lemma.

**Lemma 2:** Consider the loss function  $f$  defined in Theorem 1 with learning rate  $0 < \eta < 2/(\lambda_1 + \lambda_n)$ . For the constant  $\varepsilon$  defined in equation 9, the following forward-invariance properties holds for  $k \geq 0$ ,

$$\mathbf{x}_0 \in B(\mathbf{0}, \varepsilon) \implies \mathbf{x}_k \in B(\mathbf{0}, \varepsilon). \quad (10)$$

By this lemma, we see that  $B(\mathbf{0}, \varepsilon)$  is also a forward invariant set, i.e., the GD trajectories dived into  $B(\mathbf{0}, \varepsilon)$  remain confined within  $B(\mathbf{0}, \varepsilon)$ . The detailed proof of Lemma 2 is provided in Appendix C.

Having established these lemmas, we now prove Theorem 1 for the case of small learning rate  $0 < \eta < 2/(\lambda_1 + \lambda_n)$ . Firstly, define

$$S = \left\{ \mathbf{x}_0 \in B(\mathbf{0}, \varepsilon) : \forall k \geq 0, |x_{k,1}| < \sum_{i=2}^n |x_{k,i}| \right\}. \quad (11)$$

One can prove that  $S$  has zero measure (the detailed proof of this technical issue is provided in Appendix D). Then we only need to consider  $\mathbf{x}_0 \in B(\mathbf{0}, \varepsilon) \setminus S$ . Specifically, by definition of  $S$  in equation 11, we only need to consider the initial point  $\mathbf{x}_0 \in B(\mathbf{0}, \varepsilon)$  such that there exists  $k^* \geq 0$ ,

$$|x_{k^*,1}| \geq \sum_{i=2}^n |x_{k^*,i}|. \quad (12)$$

In this case, it's easy to see that  $|x_{k^*,1}| \geq |x_{k^*,i}|$  for  $2 \leq i \leq n$ . Thus, by Cauchy's inequality,  $\|\mathbf{x}_{k^*}\|^2 \leq nx_{k^*,1}^2$ . Then, according to equation 6 and Lemma 1,

$$|x_{k^*+1,1}| \geq a|x_{k^*,1}| - |g_{k^*,1}| \geq a|x_{k^*,1}| - C\|\mathbf{x}_{k^*}\|^2 \geq a|x_{k^*,1}| - nCx_{k^*,1}^2. \quad (13)$$

Moreover, for  $2 \leq i \leq n$ , also by equation 6 and Lemma 1,

$$|x_{k^*+1,i}| \leq |1 - \eta\lambda_i| |x_{k^*,i}| + |g_{k^*,i}| \leq b |x_{k^*,i}| + C \|\mathbf{x}_{k^*}\|^2 \leq b |x_{k^*,i}| + nCx_{k^*,1}^2. \quad (14)$$

It follows that

$$\sum_{i=2}^n |x_{k^*+1,i}| \leq b \sum_{i=2}^n |x_{k^*,i}| + n(n-1)Cx_{k^*,1}^2. \quad (15)$$

Based on equation 12 and equation 15, we have,

$$\sum_{i=2}^n |x_{k^*+1,i}| \leq b |x_{k^*,1}| + n(n-1)Cx_{k^*,1}^2. \quad (16)$$

Then, by equation 9, equation 13 and equation 16, with  $|x_{k^*,1}| \leq \|\mathbf{x}_{k^*}\| < \varepsilon$ ,

$$|x_{k^*+1,1}| \geq \sum_{i=2}^n |x_{k^*+1,i}|. \quad (17)$$

In this way, by Lemma 2, as  $\|\mathbf{x}_0\| < \varepsilon$ , we know that  $\|\mathbf{x}_k\| < \varepsilon$  holds for any  $k \geq 0$ . Then, by induction, we conclude that for all  $k \geq k^*$ ,

$$|x_{k,1}| \geq \sum_{i=2}^n |x_{k,i}|. \quad (18)$$

Moreover, to ensure our conclusion, we have to show if  $x_{k^*,1} > 0$ , then for all  $k \geq k^*$ ,  $x_{k,1} > 0$ . Actually, if  $x_{k^*,1} > 0$ , then by equation 6, equation 9 and the same induction in equation 13,

$$x_{k+1,1} \geq ax_{k,1} - nCx_{k,1}^2 \geq (a - nC\varepsilon) x_{k,1} > 0. \quad (19)$$

The above claim can be then proved inductively. In the same way, if  $x_{k^*,1} < 0$ , then for all  $k \geq k^*$ ,  $x_{k,1} < 0$ . In this situation, we know that  $x_{k,1} \neq 0$  for all  $k \geq k^*$ , if  $x_{k^*,1} \neq 0$ . Furthermore, with equation 18, based on the same induction of equation 13 and  $|x_{k,1}| \leq \varepsilon$ , we have, for all  $k \geq k^*$ ,

$$|x_{k+1,1}| \geq a |x_{k,1}| - nCx_{k,1}^2 \geq (a - nC\varepsilon) |x_{k,1}|. \quad (20)$$

Next, similarly, with equation 18, based on the same induction of equation 14, for  $k \geq k^*$  and  $2 \leq i \leq n$ , we have,

$$|x_{k+1,i}| \leq b |x_{k,i}| + C \|\mathbf{x}_k\|^2 \leq b |x_{k,i}| + nCx_{k,1}^2. \quad (21)$$

Since  $\varepsilon$  satisfies equation 9, then  $a - nC\varepsilon > b$ , one can pick  $\alpha$  satisfying  $1 < \alpha < 2$  such that  $q = b/(a - nC\varepsilon)^\alpha < 1$ . Then, according to equation 20 and equation 21, for  $k \geq k^*$  and  $2 \leq i \leq n$ , we have,

$$\frac{|x_{k+1,i}|}{|x_{k+1,1}|^\alpha} \leq \frac{b |x_{k,i}|}{(a - nC\varepsilon)^\alpha |x_{k,1}|^\alpha} + \frac{nC |x_{k,1}|^2}{(a - nC\varepsilon)^\alpha |x_{k,1}|^\alpha} = q \frac{|x_{k,i}|}{|x_{k,1}|^\alpha} + \frac{nC |x_{k,1}|^{2-\alpha}}{(a - nC\varepsilon)^\alpha}. \quad (22)$$

Since  $|x_{k,1}|^{2-\alpha} < 1$ , equation 22 can be reduced to

$$\frac{|x_{k+1,i}|}{|x_{k+1,1}|^\alpha} \leq q \frac{|x_{k,i}|}{|x_{k,1}|^\alpha} + \frac{nC}{(a - nC\varepsilon)^\alpha}. \quad (23)$$

Still, by induction, equation 23 implies that, for all  $k \geq k^*$  and  $2 \leq i \leq n$ ,

$$\frac{|x_{k,i}|}{|x_{k,1}|^\alpha} \leq q^{k-k^*} \frac{|x_{k^*,i}|}{|x_{k^*,1}|^\alpha} + \frac{nC}{(a - nC\varepsilon)^\alpha} \sum_{j=0}^{k-k^*} q^j. \quad (24)$$

As  $\sum_{j=0}^{k-k^*} q^j < 1/(1-q)$ , we have, for all  $k \geq k^*$  and  $2 \leq i \leq n$ ,

$$\frac{|x_{k,i}|}{|x_{k,1}|^\alpha} \leq q^{k-k^*} \frac{|x_{k^*,i}|}{|x_{k^*,1}|^\alpha} + \frac{nC}{(1-q)(a - nC\varepsilon)^\alpha}. \quad (25)$$

Hence, there exists a constant  $C_1 > 0$  such that the following estimation holds for  $k \geq k^*$  and  $2 \leq i \leq n$ ,  $\frac{|x_{k,i}|}{|x_{k,1}|^\alpha} \leq C_1$ . As a result, for all  $2 \leq i \leq n$ ,  $\lim_{k \rightarrow \infty} \left| \frac{x_{k,i}}{x_{k,1}} \right| = 0$ . Therefore, based on the above induction, on one hand, if  $x_{k^*,1} \neq 0$ ,

$$\lim_{k \rightarrow \infty} \frac{\mathbf{x}_k}{\|\mathbf{x}_k\|} = (\text{sign}(x_{k^*,1}), 0, \dots, 0). \quad (26)$$

This situation corresponds to the case where the GD trajectory eventually becomes dominated by the direction of the eigenvector corresponding to the smallest eigenvalue of  $\nabla^2 f(\mathbf{x}^*)$ . This behavior is intuitive and expected for almost all initial points in  $B(\mathbf{0}, \varepsilon)$ .

On the other hand, if  $x_{k^*,1} = 0$ , we see that  $\mathbf{x}_{k^*} = \mathbf{0}$  since  $\mathbf{x}_0 \notin S$ , meaning that the GD trajectory reached  $\mathbf{x}^* = \mathbf{0}$  after only finite iterations. Define

$$W_k = \{\mathbf{x}_0 \in B(\mathbf{0}, \varepsilon) : \mathbf{x}_k = \mathbf{0}\}, \quad (27)$$

then we have  $\mathbf{x}_0 \in W_{k^*}$ . According to the assumption on the GD dynamics in Theorem 1,  $W_k$  is a zero measure set since  $W_k = F^{-k}(\{\mathbf{0}\})$ , and thus the set  $S \cup (\cup_{k=0}^\infty W_k)$  has zero measure. In this way, for each initial point  $\mathbf{x}_0 \in B(\mathbf{0}, \varepsilon) \setminus (S \cup (\cup_{k=0}^\infty W_k))$ , according to equation 26, the convergence direction of GD exists. Finally, return to the set  $V_\eta$ , define

$$\widetilde{W}_k = \{\mathbf{x}_0 \in V_\eta : \mathbf{x}_k = \mathbf{0} \text{ or } \mathbf{x}_k \in S\}. \quad (28)$$

It is clear that the measure of  $\widetilde{W}_k$  is zero, thus  $\cup_{k \geq 0} \widetilde{W}_k$  also has zero measure. Then for each initial point  $\mathbf{x}_0 \in V_\eta \setminus (\cup_{k \geq 0} \widetilde{W}_k)$ , as its GD trajectories will finally enter the set  $B(\mathbf{0}, \varepsilon) \setminus (S \cup (\cup_{k=0}^\infty W_k))$ , the theorem is finally proved.

The detailed proof of Theorem 1 with large learning rate is provided in Appendix E.

### 3.2 SHARPNESS OSCILLATION AND DISCUSSION OF EOS

Consider a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfying the assumptions of Theorem 1 with a local minimum  $\mathbf{x}^* = \mathbf{0}$ . Suppose that the Hessian  $\nabla^2 f(\mathbf{0})$  have distinct eigenvalues satisfying  $0 < \lambda_1 < \lambda_2 < \dots < \lambda_n$ . By classical matrix perturbation theory (Kato, 1995; Horn & Johnson, 2013; Bhatia, 2013), we know that there exists a neighborhood  $U$  of  $\mathbf{x}^*$  such that the eigenvalues  $\lambda_1(\mathbf{x}), \dots, \lambda_n(\mathbf{x})$  of  $\nabla^2 f(\mathbf{x})$  are differentiable functions of  $\mathbf{x}$ , and  $\lambda_1(\mathbf{x}) < \dots < \lambda_n(\mathbf{x})$ . Then, by Taylor expansion, we have, for the maximal eigenvalue  $\lambda_n(\mathbf{x})$  and  $\mathbf{x} \in U$ ,

$$\lambda_n(\mathbf{x}) = \lambda_n(\mathbf{0}) + \boldsymbol{\omega}^\top \mathbf{x} + o(\|\mathbf{x}\|), \quad (29)$$

in which, by definition,  $\lambda_n(\mathbf{0}) = \lambda_n$  is the maximal eigenvalue at  $\mathbf{0}$  and  $\boldsymbol{\omega} = \lambda'_n(\mathbf{0})$ . Accordingly, by Theorem 1, the GD trajectory  $\{\mathbf{x}_k\}_{k \geq 0}$  converges or alternatively converges along a certain direction  $\mathbf{v} \in \mathbb{R}^n$ . As a result, we may conclude that: For small learning rate  $\eta \in (0, 2/(\lambda_1 + \lambda_n))$ , there exists a constant  $C_\eta$  such that

$$\lambda_n(\mathbf{x}_k) = \lambda_n + C_\eta \|\mathbf{x}_k\| + o(\|\mathbf{x}_k\|). \quad (30)$$

Therefore, ignoring the higher order term of  $\|\mathbf{x}_k\|$ , we have,  $\lambda_n(\mathbf{x}_k) \downarrow \lambda_n$  or  $\lambda_n(\mathbf{x}_k) \uparrow \lambda_n$  depending on the sign of  $C_\eta$ . For large learning rate  $\eta \in (2/(\lambda_1 + \lambda_n), 2/\lambda_n)$ , there exists a constant  $C_\eta$  such that

$$\lambda_n(\mathbf{x}_k) = \lambda_n + (-1)^k C_\eta \|\mathbf{x}_k\| + o(\|\mathbf{x}_k\|). \quad (31)$$

Similarly, ignoring the higher order term of  $\|\mathbf{x}_k\|$ , we have, for instance, when  $C_\eta > 0$ ,  $\lambda_n(\mathbf{x}_k)$  will oscillate converge to  $\lambda_n$ , i.e.,  $\lambda_n(\mathbf{x}_{2k}) \downarrow \lambda_n$  and  $\lambda_n(\mathbf{x}_{2k+1}) \uparrow \lambda_n$ . Moreover, as  $\lambda_n < 2/\eta < \lambda_1 + \lambda_n$ , compared with the above case, the sharpness  $\lambda_n(\mathbf{x}_{2k})$  is more likely larger than  $2/\eta$  for the case of small  $k$ , i.e., for the first several  $\mathbf{x}_k$  in GD trajectory. The above discussion may provide new insights to the sharpness fluctuation phenomenon of EoS.

We now present an example to illustrate our idea. Consider here a loss function of two variables,  $f(x, y) = x^2 + y^2/2 + x^2y + x^3$ , which has a local minimum  $(0, 0)$  with  $\lambda_1 = 1$  and  $\lambda_2 = 2$ . The GD dynamics and sharpness evolution under different learning rates are shown in Figure 2. In Figure 2a and 2b, the learning rate is taken as  $\eta = 0.1 < 2/(\lambda_1 + \lambda_2) = 2/3$ . In these two figures, the GD trajectory converges to the minimum along  $\mathbf{v} = (0, \mp 1)$ , and the maximum eigenvalue  $\lambda_2(\mathbf{x}_k)$  converges monotonically to  $\lambda_2 = 2$ , consistent with our theoretical prediction for small

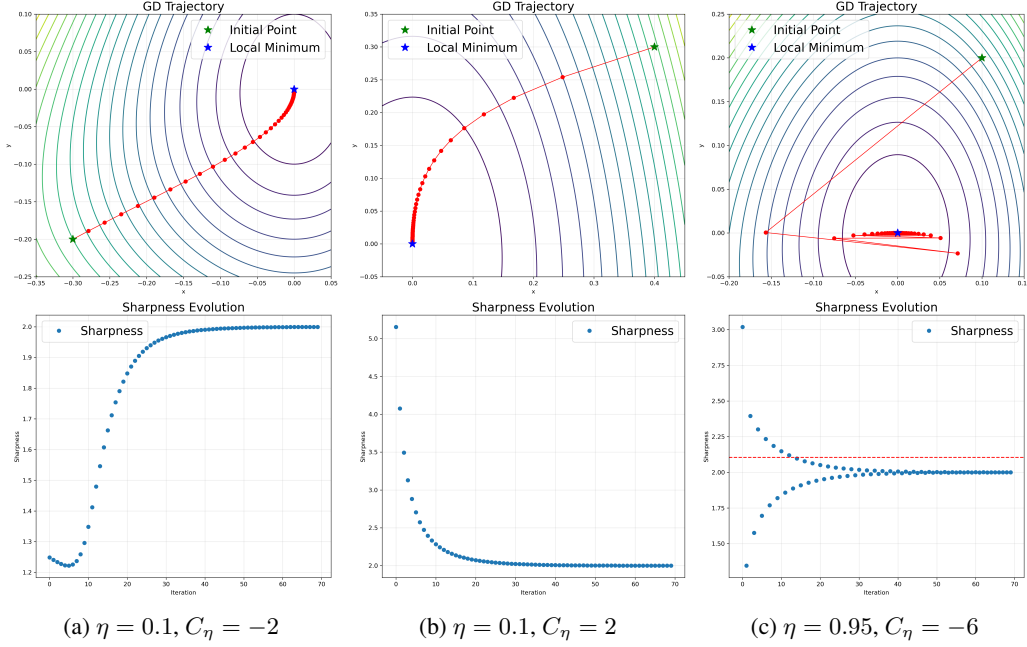


Figure 2: GD behavior on  $f(x, y) = x^2 + y^2/2 + x^2y + x^3$  under different learning rates. The top row shows the GD trajectories, and the bottom row shows the evolution of the sharpness along GD trajectory.

learning rates. On the other hand, in Figure 2c, the learning rate  $\eta = 0.95$  satisfies  $2/(\lambda_1 + \lambda_2) = 2/3 < \eta < 2/\lambda_2 = 1$ . As a result, the GD trajectory exhibits oscillations along the sharpest direction  $\mathbf{v} = (1, 0)$ . Correspondingly, the maximum eigenvalue of the Hessian oscillates around the asymptotic value  $\lambda_2$ , and may occasionally exceed the theoretical upper bound  $2/\eta \approx 2.105$ , which is shown as the dashed red line in this figure. This fluctuation is a manifestation of the EoS phenomenon, but the long-term convergence of the oscillation envelope toward  $\lambda_2$  supports the conclusion of our theorem. More results about this experiment are provided in Appendix F.

### 3.3 CONVERGENCE DIRECTIONS FOR MODERN OPTIMIZERS

Having rigorously established in Theorem 1 that GD trajectories admit a convergence direction, we are curious whether this phenomenon might extend to more widely used optimization algorithms in deep learning. Surprisingly, our experiments reveal that the existence of convergence directions is not confined to the deterministic and idealized setting of vanilla GD. Similar behaviors also emerge when using SGD with momentum and Adam in multi-class classification tasks.

To explore this, we conduct experiments on the CIFAR-10 dataset, using a convolutional neural network architecture (the details are provided in Appendix G). We train this network using two widely adopted optimization methods: SGD with momentum and Adam. For each optimizer, we monitor both the training loss and the directional alignment of successive update steps throughout training. In particular, we track the cosine similarity between consecutive GD parameter updates, defined as,

$$\cos \langle \Delta \mathbf{x}_{k+1}, \Delta \mathbf{x}_k \rangle = \frac{\langle \Delta \mathbf{x}_k, \Delta \mathbf{x}_{k+1} \rangle}{\|\Delta \mathbf{x}_k\| \|\Delta \mathbf{x}_{k+1}\|}, \quad (32)$$

where  $\Delta \mathbf{x}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ , which captures whether the optimizer’s trajectory aligns with a stable direction over time. The results are shown in Figure 3. For all optimization methods,  $\cos \langle \Delta \mathbf{x}_{k+1}, \Delta \mathbf{x}_k \rangle$  tends to 1 when the loss function converges stably in descending manner, which indicates the convergence direction emerges and supports our claim.

This finding has important implications for our understanding of optimization dynamics. It suggests that the asymptotic alignment behavior may reveal several promising research directions. The consistency of convergence directions across algorithms could inform better learning rate schedules,

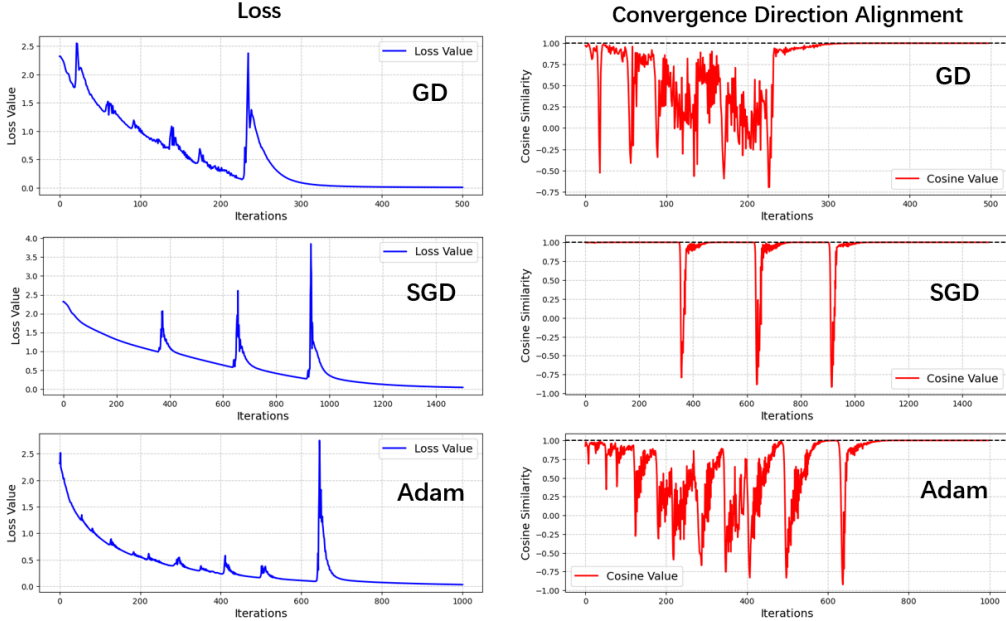


Figure 3: Convergence directions for GD, SGD, and Adam. Each row presents the training loss (left) and the convergence direction alignment (right) for a different optimizer.

initializations, or optimizers. Understanding their emergence, especially in adaptive methods, may offer insights into convergence and generalization.

## 4 CONCLUSION

In this paper, we investigate the asymptotic directional behavior of GD and establish new theoretical results. Motivated by the Gradient Conjecture from continuous-time gradient flow, we prove that under mild assumptions, the discrete-time GD also admits a similar directional property. Moreover, we discuss the oscillating behavior of sharpness by deriving the variation of eigenvalues near the minimum. Furthermore, we also find that this result holds for SGD and Adam.

This work opens several promising directions for future research.

- We only consider the locally strong convex function in the proposed theorem. It is important to investigate whether directional convergence can be extended to more general functions and other optimization algorithms, such as momentum-based or adaptive methods.
- A quantitative study of convergence speed in the direction could offer sharper theoretical guarantees and practical insights. Such results may inform the design of new optimization algorithms that leverage directional behavior for improved efficiency or generalization. Understanding these aspects could help unify optimization methods under a broader theoretical framework.
- An intriguing future direction is to explore whether the phenomenon of convergence directions observed in GD is further amplified or structurally modified under the use of periodically long steps, as suggested by the non-monotonic but globally contracting patterns in Grimmer (2024).
- Building upon the Kurdyka-Łojasiewicz (KL) inequality based works proposed in Qiu et al. (2024); Attouch et al. (2010), a promising direction is to extend the analysis to functions defined with the KL condition. Another interesting line of work is to investigate the finite-time behavior and generalization performance of stochastic optimization under the approximate descent paradigm. Finally, integrating the KL inequality with data-dependent sampling strategies or overparameterized models could provide new insights into convergence dynamics in modern machine learning.

## ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grants 62372037 and U24B20179.

## REFERENCES

- Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 247–257, 2022.
- Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 948–1024, 2022.
- Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Math. Oper. Res.*, 35(2):438–457, 2010.
- Rajendra Bhatia. *Matrix Analysis*. Springer New York, 2013.
- Yuhang Cai, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter L. Bartlett. Large stepsize gradient descent for non-homogeneous two-layer networks: Margin improvement and fast optimization. In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- Shuo Chen, Jiaying Peng, Xiaolong Li, and Yao Zhao. On the unstable convergence regime of gradient descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11373–11380, 2024.
- 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 *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.
- Jeremy Cohen et al. Adaptive learning rates in the edge of stability. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2024.
- Yan Dai, Kwangjun Ahn, and Suvrit Sra. The crucial role of normalization in sharpness-aware minimization. In *Advances in Neural Information Processing Systems*, volume 36, 2024.
- Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.
- Curtis Fox, Leonardo Galli, Mark Schmidt, and Holger Rauhut. Nonmonotone line searches operate at the edge of stability. In *Proceedings of the OPT 2024 Workshop on Optimization for Machine Learning*, 2024.
- Benjamin Grimmer. Provably faster gradient descent via long steps. *SIAM J. Optim.*, 34(3):2588–2608, 2024.
- Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2013.
- Tosio Kato. *Perturbation Theory for Linear Operators*. Springer Berlin Heidelberg, 1995.
- J. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer New York, 2012.

- Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Proceedings of the 29th Annual Conference on Learning Theory*, 2016.
- Seunghyeon Lee and Chulhee Jang. A new characterization of the edge of stability based on a sharpness measure aware of batch gradient distribution. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.
- Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- Philip M. Long and Peter L. Bartlett. Sharpness-aware minimization and the edge of stability. *Journal of Machine Learning Research*, 25(179):1–20, 2024.
- Chao Ma, Daniel Kunin, Lei Wu, and Lexing Ying. Beyond the quadratic approximation: The multiscale structure of neural network loss landscapes. *Journal of Machine Learning*, 2022.
- James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.
- Chedi Morchdi, Yi Zhou, Jie Ding, and Bei Wang. Exploring gradient oscillation in deep neural network training. In *Annual Allerton Conference on Communication, Control, and Computing*, pp. 1–7. IEEE, 2023.
- Rotem Mulayoff, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability: A view from function space. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- J.R. Munkres. *Analysis On Manifolds*. Advanced Books Classics. Avalon Publishing, 1997.
- Mor Shpigel Nacson, Rotem Mulayoff, Greg Ongie, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability in multivariate shallow ReLU networks. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.
- Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer US, 2003.
- Atsushi Nitanda, Ryuhei Kikuchi, Shugo Maeda, and Denny Wu. Why is parameter averaging beneficial in SGD? An objective smoothing perspective. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pp. 3565–3573, 2024.
- Adam Parusinski, Tadeusz Mostowski, and Krzysztof Kurdyka. Proof of the gradient conjecture of R. Thom. *Annals of Mathematics*, 152(3):763–792, 2000.
- Prin Phunyahphibarn, Junghyun Lee, Bohan Wang, Huishuai Zhang, and Chulhee Yun. Gradient descent with Polyak’s momentum finds flatter minima via large catapults. In *Proceedings of the 2nd Workshop on High-dimensional Learning Dynamics (HiLD) at ICML 2024*, 2024.
- Junwen Qiu, Bohao Ma, Xiao Li, and Andre Milzarek. A kl-based analysis framework with applications to non-descent optimization methods. *CoRR*, abs/2406.02273, 2024.
- Valentin Thomas, Fabian Pedregosa, Bart van Merriënboer, Pierre-Antoine Manzagol, Yoshua Bengio, and Nicolas Le Roux. On the interplay between noise and curvature and its effect on optimization and generalization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 3503–3513, 2020.
- Mingze Wang and Chao Ma. Understanding multi-phase optimization dynamics and rich nonlinear behaviors of ReLU networks. In *Advances in Neural Information Processing Systems*, volume 36, 2024.
- Mingze Wang and Lei Wu. A theoretical analysis of noise geometry in stochastic gradient descent. *arXiv preprint arXiv:2310.00692*, 2023.
- Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along GD trajectory: Progressive sharpening and edge of stability. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

- Jingfeng Wu, Vladimir Braverman, and Jason D. Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with SGD. *CoRR*, abs/1802.08770, 2018.
- Junchi Yang, Xiang Li, Ilyas Fatkhullin, and Niao He. Two sides of one coin: the limits of untuned SGD and the power of adaptive methods. In *Advances in Neural Information Processing Systems*, volume 36, 2024.
- Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.
- Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 7654–7663, 2019.



## APPENDIX A: ON THE DIAGONAL ASSUMPTION FOR THE HESSIAN MATRIX $\nabla^2 f(\mathbf{0})$

In this appendix, we justify the assumption made in the proof sketch that  $\nabla^2 f(\mathbf{0})$  is diagonal. Consider the loss function defined in Theorem 1, as  $\nabla^2 f(\mathbf{0})$  is symmetric, there exists an orthogonal matrix  $P \in \mathbb{R}^{n \times n}$  such that

$$\nabla^2 f(\mathbf{0}) = P^\top D P, \quad (33)$$

where  $D = \text{diag}(\lambda_1, \dots, \lambda_n)$ . Now define a change of coordinates  $\mathbf{y} = P\mathbf{x}$ , and consider the transformed function

$$g(\mathbf{y}) = f(P^\top \mathbf{y}). \quad (34)$$

In the new coordinates, since

$$\nabla g(\mathbf{y}) = P \nabla f(P^\top \mathbf{y}), \quad (35)$$

by GD dynamics equation 2, we have,

$$\mathbf{y}_{k+1} = P\mathbf{x}_{k+1} = P\mathbf{x}_k - \eta P \nabla f(\mathbf{x}_k) = \mathbf{y}_k - \eta \nabla g(\mathbf{y}_k). \quad (36)$$

Therefore, the GD dynamics are preserved under the transformation. Moreover,

$$\nabla^2 g(\mathbf{0}) = P \nabla^2 f(\mathbf{0}) P^\top = D, \quad (37)$$

which is diagonal. Hence, by working in the transformed coordinates, we may assume that the Hessian at the local minimum is diagonal without altering the GD behavior. This justifies the assumption  $\nabla^2 f(\mathbf{0}) = \text{diag}(\lambda_1, \dots, \lambda_n)$  used in the proof sketch of Theorem 1.

## APPENDIX B: PROOF OF LEMMA 1

To prove Lemma 1, we first check the function  $f$  considered in Theorem 1 satisfies all conditions in Theorem 2. Since  $f \in \mathcal{C}^3$  and  $\nabla^2 f(\mathbf{0})$  is positive definite, there exists  $r > 0$  such that  $f$  is strongly convex on  $\overline{B(\mathbf{0}, r)}$ , where  $\overline{B(\mathbf{0}, r)}$  is the closure of the open ball  $B(\mathbf{0}, r)$ . Therefore,  $f$  satisfies all the conditions required by Theorem 2. Next, we will apply Theorem 2 and proceed with the proof of Lemma 1. By Taylor expansion, we have,

$$\partial_i f(\mathbf{x}) = \partial_i f(\mathbf{0}) + \sum_{j=1}^n \partial_{i,j} f(\mathbf{0}) x_j + \sum_{j,l=1}^n \left( \int_0^1 (1-t) \partial_{i,j,l} f(t\mathbf{x}) dt \right) x_j x_l \quad (38)$$

and

$$\partial_{i,j} f(\mathbf{x}) = \partial_{i,j} f(\mathbf{0}) + \sum_{l=1}^n \left( \int_0^1 \partial_{i,j,l} f(t\mathbf{x}) dt \right) x_l. \quad (39)$$

On the other hand, according to the definition equation 7, we have,

$$\partial_j g_i(\mathbf{x}) = \begin{cases} -\eta \partial_{i,j} f(\mathbf{x}), & j \neq i, \\ \eta \lambda_i - \eta \partial_{i,i} f(\mathbf{x}), & j = i. \end{cases} \quad (40)$$

It follows that

$$g_i(\mathbf{x}) = -\eta \sum_{j,l=1}^n \left( \int_0^1 (1-t) \partial_{i,j,l} f(t\mathbf{x}) dt \right) x_j x_l \quad (41)$$

and

$$\partial_j g_i(\mathbf{x}) = -\eta \sum_{l=1}^n \left( \int_0^1 \partial_{i,j,l} f(t\mathbf{x}) dt \right) x_l. \quad (42)$$

Since  $|x_j|, |x_l| \leq \|\mathbf{x}\|$  and  $\partial_{i,j,l} f$  is continuous on  $\overline{B(\mathbf{0}, r)}$ , there exists a constant  $C > 0$  such that for all  $\mathbf{x} \in \overline{B(\mathbf{0}, r)}$ , we have  $|g_i(\mathbf{x})| \leq C \|\mathbf{x}\|^2$  and  $|\partial_j g_i(\mathbf{x})| \leq C \|\mathbf{x}\|$  hold. In this way, as  $\eta < 2/\lambda_n$ , then according to Theorem 2, there exists an open set  $\Omega \subset B(\mathbf{0}, r)$  containing  $\mathbf{0}$  such that for all  $\mathbf{x}_0 \in \Omega$ , we have  $\mathbf{x}_k \in \Omega$ ,  $|g_{k,i}| \leq C \|\mathbf{x}_k\|^2$  and  $\left| \frac{\partial g_{k,i}}{\partial x_{k,j}} \right| \leq C \|\mathbf{x}_k\|$ . Lemma 1 is finally proved.

## APPENDIX C: PROOF OF LEMMA 2

For  $\mathbf{x}_0 \in \Omega$ , by Lemma 1 and equation 6, we know that  $\mathbf{x}_k \in \Omega$  holds for all  $k \geq 0$ , and

$$x_{k+1,i}^2 \leq a^2 x_{k,i}^2 + 2aC |x_{k,i}| \|\mathbf{x}_k\|^2 + C^2 \|\mathbf{x}_k\|^4. \quad (43)$$

Then,

$$\|\mathbf{x}_{k+1}\|^2 \leq a^2 \|\mathbf{x}_k\|^2 + 2aC \left( \sum_{i=1}^n |x_{k,i}| \right) \|\mathbf{x}_k\|^2 + nC^2 \|\mathbf{x}_k\|^4. \quad (44)$$

Since  $\sum_{i=1}^n |x_{k,i}| \leq \sqrt{n} \|\mathbf{x}_k\|$ , we have

$$\|\mathbf{x}_{k+1}\|^2 \leq a^2 \|\mathbf{x}_k\|^2 + 2aC\sqrt{n} \|\mathbf{x}_k\|^3 + nC^2 \|\mathbf{x}_k\|^4 = (a + \sqrt{n}C\|\mathbf{x}_k\|)^2 \|\mathbf{x}_k\|^2. \quad (45)$$

On the other hand, by equation 9, we have

$$a + \sqrt{n}C\varepsilon < \sqrt{a} < 1. \quad (46)$$

Then, according to equation 45 and equation 46, we can apply induction to prove that  $\|\mathbf{x}_k\| < \varepsilon$  if  $\|\mathbf{x}_0\| < \varepsilon$ . We then finish to prove this lemma.

APPENDIX D: PROOF OF THE ZERO MEASURE PROPERTY FOR THE SET  $S$  DEFINED IN EQUATION 11

Noticed that we consider here the case of small learning rate with  $0 < \eta < 2/(\lambda_1 + \lambda_n)$ . We first give the following lemmas.

**Lemma 3:** Consider the function  $g_{k,i}$  defined in equation 7, which is a twice differentiable function of  $\mathbf{x}_0$ . For all  $1 \leq i \leq n$ , the series

$$\sum_{k=0}^{\infty} \frac{g_{k,i}}{a^k} \quad (47)$$

is convergent for all  $\mathbf{x}_0 \in B(\mathbf{0}, \varepsilon)$ . Here, the constant  $\varepsilon$  is defined in equation 9,  $a$  is defined in equation 5.

*Proof.* Regarding the proof for Lemma 2, by equation 45, for  $\mathbf{x}_0 \in B(\mathbf{0}, \varepsilon)$ , we have

$$\|\mathbf{x}_{k+1}\| \leq (a + \sqrt{n}C\varepsilon) \|\mathbf{x}_k\|. \quad (48)$$

And thus,

$$\|\mathbf{x}_k\| \leq (a + \sqrt{n}C\varepsilon)^k \|\mathbf{x}_0\| \leq \varepsilon (a + \sqrt{n}C\varepsilon)^k. \quad (49)$$

As a result, by Lemma 1 and equation 49,

$$\frac{|g_{k,i}|}{a^k} \leq C\varepsilon^2 \left( \frac{a + \sqrt{n}C\varepsilon}{\sqrt{a}} \right)^{2k}. \quad (50)$$

Thus, as a result of equation 46,  $\sum_{k=0}^{\infty} g_{k,i}/a^k$  is convergent.  $\square$

**Lemma 4:** For the function  $g_{k,i}$  defined in equation 7, take

$$\gamma_k^{(i,j)} = \frac{\partial g_{k,i}}{\partial x_{0,j}}, \quad (51)$$

which is a differentiable function of  $\mathbf{x}_0$ . Then for the constant  $\varepsilon$  defined in equation 9 and  $a$  defined in equation 5,

$$\sum_{k=0}^{\infty} \frac{\gamma_k^{(i,j)}}{a^{k+1}} \quad (52)$$

is convergent for all  $\mathbf{x}_0 \in B(\mathbf{0}, \varepsilon)$ , and

$$\left| \sum_{k=0}^{\infty} \frac{\gamma_k^{(i,j)}}{a^{k+1}} \right| < 1. \quad (53)$$

*Proof.* Firstly, we define,

$$u_{k,j} = \sum_{i=1}^n \left| \frac{\partial x_{k,i}}{\partial x_{0,j}} \right|. \quad (54)$$

By equation 6, we have

$$u_{k+1,j} = \sum_{i=1}^n \left| \frac{\partial x_{k+1,i}}{\partial x_{0,j}} \right| \leq a \sum_{i=1}^n \left| \frac{\partial x_{k,i}}{\partial x_{0,j}} \right| + \sum_{i=1}^n \sum_{l=1}^n \left| \frac{\partial g_{k,i}}{\partial x_{k,l}} \frac{\partial x_{k,l}}{\partial x_{0,j}} \right|. \quad (55)$$

Then, by Lemma 1 and equation 55,

$$u_{k+1,j} \leq a u_{k,j} + nC \|\mathbf{x}_k\| u_{k,j} = (a + nC \|\mathbf{x}_k\|) u_{k,j}. \quad (56)$$

Since  $u_{0,j} = 1$  and  $\mathbf{x}_0 \in B(\mathbf{0}, \varepsilon)$ , by equation 56 and Lemma 2,

$$u_{k,j} \leq \left( \prod_{l=0}^{k-1} (a + nC \|\mathbf{x}_l\|) \right) u_{0,j} \leq (a + nC\varepsilon)^k. \quad (57)$$

Next, by the definition of  $\gamma_k^{(i,j)}$ , equation 54 and Lemma 1,

$$\left| \gamma_k^{(i,j)} \right| \leq \sum_{l=1}^n \left| \frac{\partial g_{k,i}}{\partial x_{k,l}} \frac{\partial x_{k,l}}{\partial x_{0,j}} \right| \leq C \|\mathbf{x}_k\| \sum_{l=1}^n \left| \frac{\partial x_{k,l}}{\partial x_{0,j}} \right| = C \|\mathbf{x}_k\| u_{k,j}. \quad (58)$$

Then, by equation 49 and equation 57,

$$\left| \gamma_k^{(i,j)} \right| \leq C\varepsilon (a + \sqrt{n}C\varepsilon)^k (a + nC\varepsilon)^k \leq C\varepsilon (a + nC\varepsilon)^{2k}. \quad (59)$$

Finally, according to equation 9, we see that  $\sum_{k=0}^{\infty} \gamma_k^{(i,j)} / a^{k+1}$  is convergent and

$$\left| \sum_{k=0}^{\infty} \frac{\gamma_k^{(i,j)}}{a^{k+1}} \right| \leq \sum_{k=0}^{\infty} \frac{\left| \gamma_k^{(i,j)} \right|}{a^{k+1}} \leq \frac{C\varepsilon}{a} \sum_{k=0}^{\infty} \left( \frac{(a + nC\varepsilon)^2}{a} \right)^k = \frac{C\varepsilon}{a - (a + nC\varepsilon)^2} < 1. \quad (60)$$

This lemma is then proved.  $\square$

**Lemma 5:** If  $\mathbf{x}_0 \in S$ , then for all  $k \geq 0$ ,

$$\sum_{i=2}^n |x_{k,i}| \leq \sqrt{n}\varepsilon(b + 2nC\varepsilon)^k, \quad (61)$$

where  $S$ ,  $\varepsilon$ ,  $b$  and  $C$  are defined in equation 11, equation 9, equation 5, and Lemma 1 respectively.

*Proof.* Clearly, by Lemma 1, we know that  $\mathbf{x}_k \in B(\mathbf{0}, \varepsilon)$  since  $\mathbf{x}_0 \in S \subset B(\mathbf{0}, \varepsilon)$ . Then, by equation 5, equation 6, and Lemma 1,

$$\sum_{i=2}^n |x_{k,i}| \leq b \sum_{i=2}^n |x_{k-1,i}| + (n-1)C \|\mathbf{x}_{k-1}\|^2 \leq b \sum_{i=2}^n |x_{k-1,i}| + nC\varepsilon \|\mathbf{x}_{k-1}\|. \quad (62)$$

On the other hand, by the definition of  $S$ ,

$$\|\mathbf{x}_k\| \leq \sum_{i=1}^n |x_{k,i}| \leq 2 \sum_{i=2}^n |x_{k,i}|. \quad (63)$$

Then, by equation 62

$$\sum_{i=2}^n |x_{k,i}| \leq (b + 2nC\varepsilon) \sum_{i=2}^n |x_{k-1,i}|. \quad (64)$$

By induction, we have,

$$\sum_{i=2}^n |x_{k,i}| \leq (b + 2nC\varepsilon)^k \sum_{i=2}^n |x_{0,i}| \leq (b + 2nC\varepsilon)^k \sqrt{n} \|\mathbf{x}_0\| \leq \sqrt{n}\varepsilon(b + 2nC\varepsilon)^k. \quad (65)$$

The proof of this lemma is then finished.  $\square$

**Lemma 6:** If  $x_0 \in S$ , then

$$\sum_{k=0}^{\infty} \frac{g_{k,1}}{a^{k+1}} + x_{0,1} = 0. \quad (66)$$

where  $S$ ,  $a$ ,  $g_{k,1}$ , are defined in equation 11, equation 5 and equation 7 respectively.

*Proof.* Suppose that

$$\sum_{k=0}^{\infty} \frac{g_{k,1}}{a^{k+1}} + x_{0,1} = \theta. \quad (67)$$

By equation 6, equation 7 and induction, we have, for all  $l \geq 0$ ,

$$\sum_{k=l}^{\infty} \frac{g_{k,1}}{a^{k+1-l}} + x_{l,1} = a^l \theta. \quad (68)$$

Then,

$$\theta = \frac{x_{l,1}}{a^l} + \sum_{k=l}^{\infty} \frac{g_{k,1}}{a^{k+1}}. \quad (69)$$

As a result,

$$|\theta| \leq \left| \frac{x_{l,1}}{a^l} \right| + \sum_{k=l}^{\infty} \frac{|g_{k,1}|}{a^{k+1}}. \quad (70)$$

By Lemma 5,

$$|x_{l,1}| \leq \sum_{i=2}^n |x_{l,i}| \leq \sqrt{n} \varepsilon (b + 2nC\varepsilon)^l, \quad (71)$$

then for all  $l \geq 0$ ,

$$|\theta| \leq \sqrt{n} \varepsilon \left( \frac{b + 2nC\varepsilon}{a} \right)^l + \sum_{k=l}^{\infty} \frac{|g_{k,1}|}{a^{k+1}}. \quad (72)$$

By equation 9 and Lemma 3, let  $l \rightarrow \infty$ , we have  $\theta = 0$ . We then finish the proof of this lemma.  $\square$

Next, for proving  $S$  is measure zero, we need the following classical theorems.

**Theorem 3:** (Inverse Function Theorem, Munkres (1997)). Let  $F : U \rightarrow \mathbb{R}^n$  be a  $\mathcal{C}^k$  mapping with  $k \geq 1$ , where  $U$  is an open subset of  $\mathbb{R}^n$ . Suppose that  $F$  is injective and  $\det DF(\mathbf{x}) \neq 0$  for all  $\mathbf{x} \in U$ , then  $F$  is a  $\mathcal{C}^r$ -diffeomorphism from  $U$  to the open set  $F(U)$ . Here,  $DF(\mathbf{x})$  is the Jacobian of  $F$  at  $\mathbf{x}$ .

**Theorem 4:** (Sard's Theorem, Lee (2012)) Let  $F : U \rightarrow \mathbb{R}^n$  be a  $\mathcal{C}^k$  mapping with  $k \geq 1$ , where  $U$  is an open subset of  $\mathbb{R}^m$ . Define  $A = \{\mathbf{x} \in U \mid \mathbf{x} \text{ is a critical point of } F\}$ , then the image  $F(A)$  has measure zero in  $\mathbb{R}^n$ . Here,  $\mathbf{x}$  is a critical point means that  $\text{rank}(DF(\mathbf{x})) < n$ .

By the above theorems and lemmas, we now prove that the measure of  $S$  is zero. Define  $\sigma : B(\mathbf{0}, \varepsilon) \rightarrow \mathbb{R}$  as, for  $\mathbf{x} = (x_1, \dots, x_n) \in B(\mathbf{0}, \varepsilon)$ ,

$$\sigma(\mathbf{x}) = \sum_{k=0}^{\infty} \frac{g_{k,1}(\mathbf{x})}{a^{k+1}}, \quad (73)$$

and  $H : B(\mathbf{0}, \varepsilon) \rightarrow \mathbb{R}^n$  by

$$H(\mathbf{x}) = (\sigma(\mathbf{x}) + x_1, x_2, \dots, x_n). \quad (74)$$

By Lemma 3, since  $\sum_{k=0}^{\infty} g_{k,1}(\mathbf{x})/a^{k+1}$  is convergent,  $\sigma(\mathbf{x})$  is well defined on  $B(\mathbf{0}, \varepsilon)$ , which also implies  $H(\mathbf{x})$  is well defined on  $B(\mathbf{0}, \varepsilon)$ . By Lemma 4, as  $\sum_{k=0}^{\infty} \gamma_k^{(1,j)}/a^{k+1}$  is convergent for all  $1 \leq j \leq n$ , then  $\sigma(\mathbf{x})$  is  $\mathcal{C}^1$  on  $B(\mathbf{0}, \varepsilon)$  with (consider here  $\mathbf{x} = \mathbf{x}_0$  in definition of  $\gamma_{1,j}$ )

$$\partial_j \sigma = \sum_{k=0}^{\infty} \frac{\gamma_k^{(1,j)}}{a^{k+1}}. \quad (75)$$

We will prove that  $H$  is an injection on  $B(\mathbf{0}, \varepsilon)$ . By contradiction, suppose that  $H$  is not injective, then there exists  $\mathbf{x} \neq \mathbf{y}$  such that  $H(\mathbf{x}) = H(\mathbf{y})$ . Thus,  $x_i = y_i$  for all  $2 \leq i \leq n$  and

$$\frac{\sigma(\mathbf{x}) - \sigma(\mathbf{y})}{x_1 - y_1} = -1. \quad (76)$$

Then there exists  $\xi$  between  $x_1$  and  $y_1$  such that

$$\partial_1 \sigma(\xi, x_2, \dots, x_n) = -1, \quad (77)$$

which is contradict to equation 53 given by Lemma 4, as  $\|(\xi, x_2, \dots, x_n)\| \leq \max\{\|\mathbf{x}\|, \|\mathbf{y}\|\} < \varepsilon$  and by taking  $j = 1$  in equation 75. Moreover, as

$$DH = \begin{pmatrix} \partial_1 \sigma + 1 & \partial_2 \sigma & \cdots & \partial_n \sigma \\ \mathbf{0} & & I_{n-1} & \end{pmatrix}. \quad (78)$$

Clearly, by equation 53 and equation 75, we have  $\det(DH) = \partial_1 \sigma + 1 > 0$ . Thus, according to Theorem 3,  $H$  is a diffeomorphism from  $B(\mathbf{0}, \varepsilon)$  to  $H(B(\mathbf{0}, \varepsilon))$ . Define

$$W = \{\mathbf{z} = (z_1, \dots, z_{n-1}) \in \mathbb{R}^{n-1} : (0, \mathbf{z}) \in H(B(\mathbf{0}, \varepsilon))\}. \quad (79)$$

Now we prove that  $W$  is an open set in  $\mathbb{R}^{n-1}$ . Actually, for any  $\mathbf{z} \in W$ , as  $(0, \mathbf{z}) \in H(B(\mathbf{0}, \varepsilon))$ , there exists  $\delta > 0$  such that  $B((0, \mathbf{z}), \delta) \subset H(B(\mathbf{0}, \varepsilon))$  since  $H(B(\mathbf{0}, \varepsilon))$  is open. Therefore, for  $\mathbf{z}' \in B(\mathbf{z}, \delta)$ , as  $\|(0, \mathbf{z}) - (0, \mathbf{z}')\| < \delta$ , we know that  $(0, \mathbf{z}') \in B((0, \mathbf{z}), \delta) \subset H(B(\mathbf{0}, \varepsilon))$ . This implies  $\mathbf{z}' \in W$  according to the definition of  $W$  in equation 79. Lastly, define  $\varphi : W \rightarrow B(\mathbf{0}, \varepsilon)$  as,

$$\varphi(\mathbf{z}) = H^{-1}(0, \mathbf{z}). \quad (80)$$

Since  $H$  is  $C^1$ , we have  $\varphi \in C^1$  and  $\text{rank}(D\varphi) \leq n - 1$ . As a result, by Theorem 4,  $\varphi(W)$  has measure zero. For all  $\mathbf{x}_0 \in S$ , by Lemma 6, we have

$$\sigma(\mathbf{x}_0) + x_{0,1} = 0. \quad (81)$$

Therefore,  $H(\mathbf{x}_0) = (0, x_{0,2}, \dots, x_{0,n})$ , and thus

$$\mathbf{x}_0 = H^{-1}(0, x_{0,2}, \dots, x_{0,n}) = \varphi(x_{0,2}, \dots, x_{0,n}) \in \varphi(W). \quad (82)$$

Hence,  $S \subset \varphi(W)$  has measure zero. This finishes the proof.

## APPENDIX E: PROOF OF THEOREM 1 WITH $2/(\lambda_1 + \lambda_n) < \eta < 2/\lambda_n$

In this appendix, we provide the proof of Theorem 1 with  $2/(\lambda_1 + \lambda_n) < \eta < 2/\lambda_n$ . The proof of this case is similar to that of small learning rate, and only main steps are provided as below. We still assume  $\mathbf{x}^* = \mathbf{0}$  and the Hessian matrix  $\nabla^2 f(0)$  is diagonal. Define

$$a = |1 - \eta\lambda_n|, \quad b = \max_{1 \leq i \leq n-1} \{|1 - \eta\lambda_i|\}. \quad (83)$$

Since  $2/(\lambda_1 + \lambda_n) < \eta < 2/\lambda_n$ , it follows that  $a > b \geq 0$ . Notice that Lemma 1 still holds for this case. Then we select a new  $\varepsilon > 0$  satisfying

$$B(\mathbf{0}, \varepsilon) \subset \Omega, \quad C\varepsilon + (a + nC\varepsilon)^2 < a, \quad \varepsilon \leq \frac{a-b}{3n^2C}, \quad (84)$$

where  $a$  and  $b$  are defined in equation 83,  $C$  and  $\Omega$  is given by Lemma 1. Moreover, with such defined  $\varepsilon$ , we know that in this case, the following forward-invariance properties holds for  $k \geq 0$ ,

$$\mathbf{x}_0 \in B(\mathbf{0}, \varepsilon) \implies \mathbf{x}_k \in B(\mathbf{0}, \varepsilon). \quad (85)$$

Similarly, we redefine, for  $\varepsilon$  defined in equation 84,

$$S = \left\{ \mathbf{x}_0 \in B(\mathbf{0}, \varepsilon) : \forall k \geq 0, |x_{k,n}| < \sum_{i=1}^{n-1} |x_{k,i}| \right\}. \quad (86)$$

The same as the case of small learning rate, one can prove that the measure of the set  $S$  is zero. Then, consider the initial point  $\mathbf{x}_0 \in B(\mathbf{0}, \varepsilon) \setminus S$ , i.e.,  $\mathbf{x}_0 \in B(\mathbf{0}, \varepsilon)$  satisfies that there exists  $k^* \geq 0$ ,

$$|x_{k^*,n}| \geq \sum_{i=1}^{n-1} |x_{k^*,i}|. \quad (87)$$

By induction, we can prove that for all  $k \geq k^*$ ,

$$|x_{k,n}| \geq \sum_{i=1}^{n-1} |x_{k,i}|. \quad (88)$$

Moreover, we know that  $x_{k,n} \neq 0$  for all  $k \geq k^*$ , if  $x_{k^*,n} \neq 0$ . And, for all  $1 \leq i \leq n-1$ , if  $x_{k^*,i} \neq 0$ ,

$$\lim_{k \rightarrow \infty} \left| \frac{x_{k,i}}{x_{k,n}} \right| = 0. \quad (89)$$

Therefore, on one hand, if  $x_{k^*,n} \neq 0$ ,

$$\lim_{k \rightarrow \infty} (-1)^k \frac{\mathbf{x}_k}{\|\mathbf{x}_k\|} = (0, \dots, 0, \text{sign}(x_{k^*,n})). \quad (90)$$

On the other hand, if  $x_{k^*,n} = 0$ , we see that  $\mathbf{x}_{k^*} = \mathbf{0}$  since  $\mathbf{x}_0 \notin S$ . Define

$$W_k = \{\mathbf{x}_0 \in B(\mathbf{0}, \varepsilon) : \mathbf{x}_k = \mathbf{0}\}, \quad (91)$$

then  $W_k$  is a zero measure set and  $\mathbf{x}_0 \in W_{k^*}$ . This finishes the proof for Theorem 1 with  $2/(\lambda_1 + \lambda_n) < \eta < 2/\lambda_n$  for almost all  $\mathbf{x}_0 \in B(\mathbf{0}, \varepsilon)$  since the set  $S \cup (\cup_{k=0}^{\infty} W_k)$  has zero measure.

Finally, by the same argument as in the case of small learning rate, the Theorem 1 with large learning rate can be proved in the same way.

## APPENDIX F: EXPERIMENT DETAILS IN SECTION 3.2

In this appendix, we provide additional figures to further illustrate the GD dynamics and sharpness behavior discussed in Section 3.2. We begin with the function introduced in the main text,

$$f(x, y) = x^2 + \frac{y^2}{2} + x^2y + x^3, \quad (92)$$

which admits a local minimum at  $(0, 0)$  with Hessian eigenvalues  $\lambda_1 = 1$  and  $\lambda_2 = 2$ . Its GD dynamics and sharpness evolution under different learning rates are shown in Figure 5 and Figure 6. For small learning rates, the GD trajectory converges with a clear direction to the local minimum, whose convergence direction aligns with the eigenvector associated with  $\lambda_1$ , and the maximum eigenvalue  $\lambda_2(\mathbf{x}_k)$  varies monotonically to its limit  $\lambda_2 = 2$ , in agreement with Theorem 1. On the other hand, for larger learning rates, with  $2/(\lambda_1 + \lambda_2) = 2/3 < \eta < 2/\lambda_2 = 1$ , the GD trajectory exhibits oscillations along the sharpest direction. Moreover, under some cases for large learning rates, the eigenvalue  $\lambda_2(\mathbf{x}_k)$  fluctuates around the asymptotic value and occasionally exceeds the theoretical threshold  $2/\eta$ , shown as the red dashed line. These fluctuations exhibit characteristics of convergence and supports the theoretical conclusion.

Moreover, we also provide some experimental results for the function

$$f(x, y) = x^2 + \frac{y^2}{2} + xy^2 + x^3. \quad (93)$$

The motivation for selecting this function is that  $C_\eta$  may be 0 under some settings. This makes their behavior particularly interesting for understanding the GD dynamics and sharpness evolution. This function has a local minimum  $(0, 0)$  with Hessian eigenvalues  $\lambda_1 = 1$  and  $\lambda_2 = 2$ . Here, we observe that the results on this function is similar with that of  $f$  defined in equation 92, but  $C_\eta = 0$  for some settings (See Figure 7 for details).

## APPENDIX G: EXPERIMENT DETAILS IN SECTION 3.3

To examine the emergence of convergence directions in modern optimizers, we design controlled experiments using a convolutional neural network trained on a subset of the CIFAR-10 dataset. This appendix provides a brief account of our implementation.

**Network Architecture:** We use a lightweight convolutional neural network consisting of two convolutional layers followed by a fully connected output layer. The architecture is shown in the following figure:

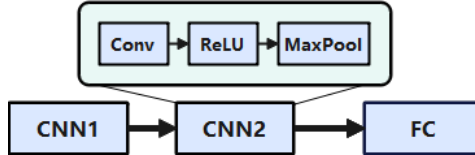


Figure 4: Network architecture for experiments performed in Section 3.3

**Dataset and Loss Function:** We use the CIFAR-10 dataset, which contains 50,000 training and 10,000 test images across 10 balanced classes. In addition, the criterion for the training is cross entropy loss.

**Training Configuration:** The configuration of our neural network are:

**CNN1:**  $3 \rightarrow 16$  filters, kernel size  $3 \times 3$ , padding 1, followed by ReLU and  $2 \times 2$  max pooling.

**CNN2:**  $16 \rightarrow 32$  filters, kernel size  $3 \times 3$ , padding 1, followed by ReLU and  $2 \times 2$  max pooling.

**FC:** Linear layer with input size  $32 \times 8 \times 8 = 2,048$  and output size 10.

For GD and SGD experiments, the model is trained using the SGD optimizer with a fixed learning rate of 0.1 and 0.01 respectively, and both of them has a momentum with  $\beta = 0.9$ . For Adam experiment, the model is trained using the Adam optimizer with a fixed learning rate of 0.01,  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . All code is implemented using PyTorch.

Moreover, to further illustrate the emergence of convergence directions of the three optimizers on classical networks, we also design controlled experiments using Inception-v3 and ResNet-18, which are trained on the same subset of the CIFAR-10 dataset. For both experiments on Inception-v3 and ResNet-18, we use the built-in architectures in PyTorch. We also apply the same settings of loss function, learning rates and momentum parameters as the previous experiment. One can see in Figure 8 and 9 such that on these two networks, all of three optimizers emerge the convergence direction as the fully connected network, which shows the correctness of our result.

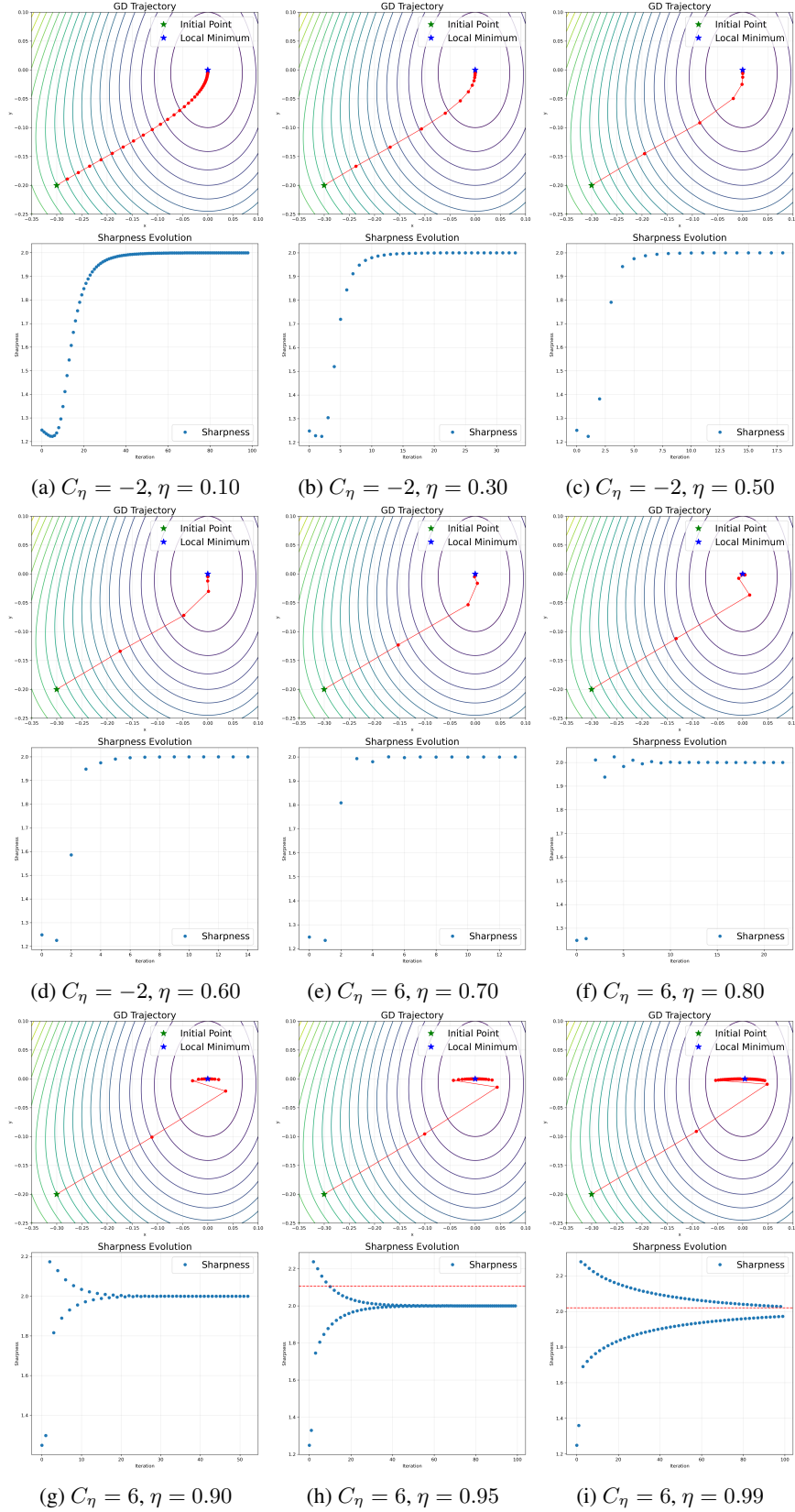


Figure 5: GD behavior on  $f(x, y) = x^2 + y^2/2 + x^2y + x^3$  under different learning rates with initial point  $(-0.3, -0.2)$ . The top row shows the GD trajectories, and the bottom row shows the evolution of the sharpness along GD trajectory.



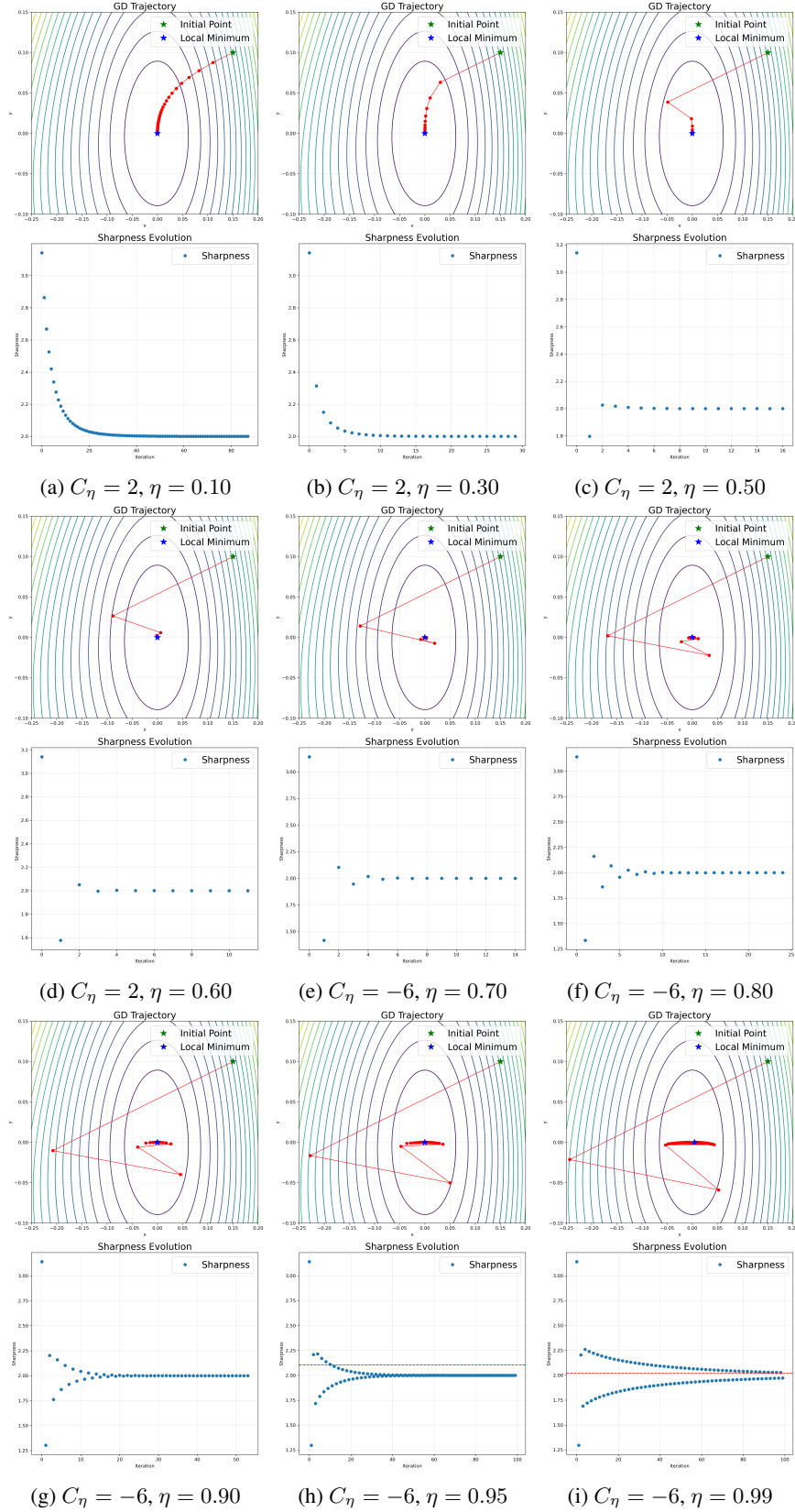


Figure 6: GD behavior on  $f(x, y) = x^2 + y^2/2 + x^2y + x^3$  under different learning rates with initial point  $(0.15, 0.1)$ . The top row shows the GD trajectories, and the bottom row shows the evolution of the sharpness along GD trajectory.

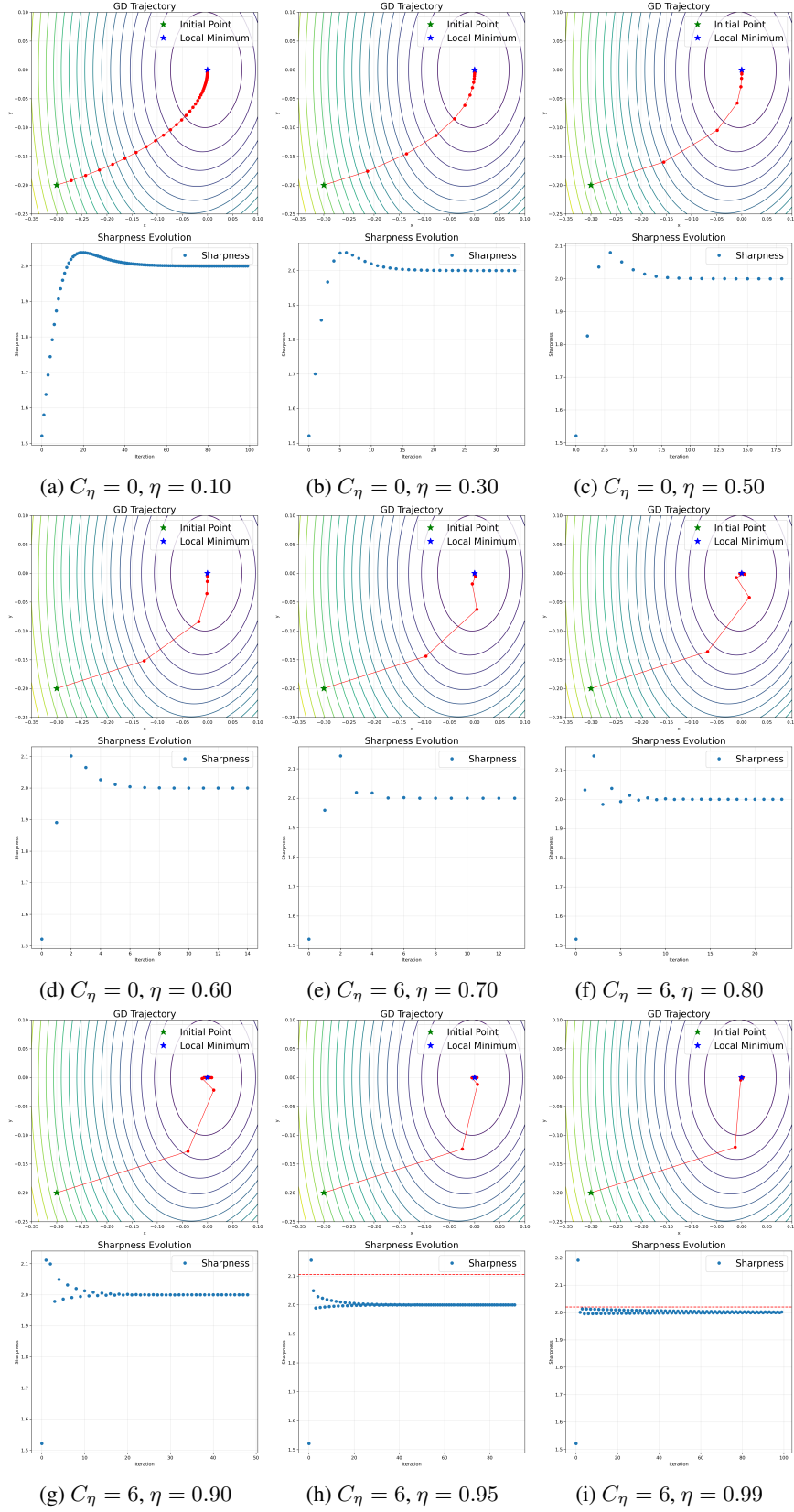


Figure 7: GD behavior on  $f(x, y) = x^2 + y^2/2 + xy^2 + x^3$  under different learning rates with initial point  $(-0.3, -0.2)$ . The top row shows the GD trajectories, and the bottom row shows the evolution of the sharpness along GD trajectory.

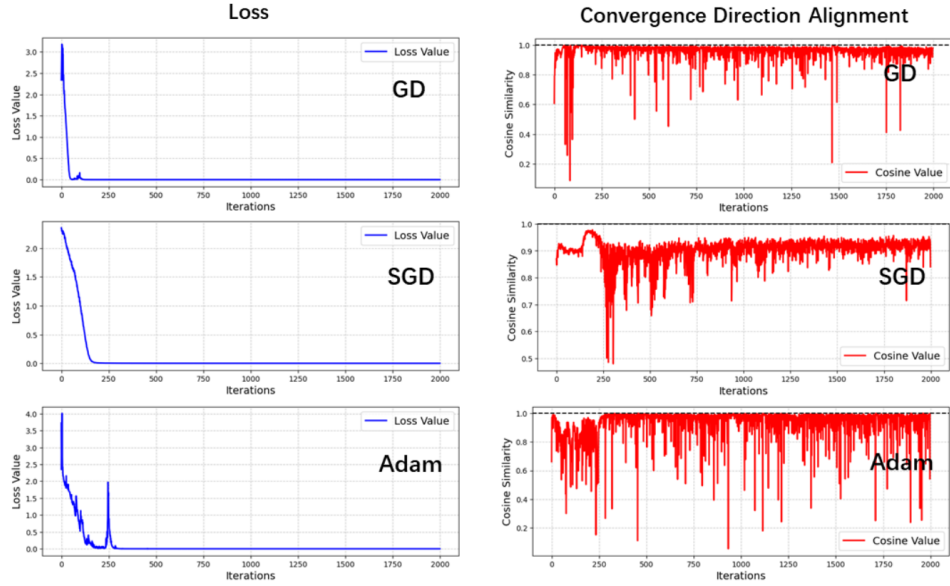


Figure 8: Convergence directions for GD, SGD, and Adam on Inception-v3 network. Each row presents the training loss (left) and the convergence direction alignment (right) for a different optimizer.

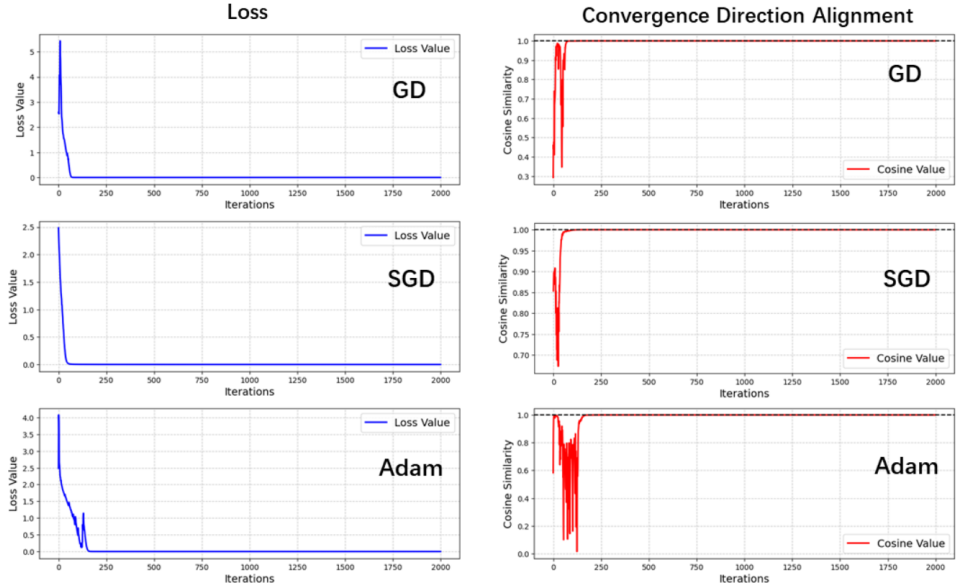


Figure 9: Convergence directions for GD, SGD, and Adam on ResNet-18 network. Each row presents the training loss (left) and the convergence direction alignment (right) for a different optimizer.