On the Unstable Convergence Regime of Gradient Descent

Shuo Chen[1], **Jiaying Peng**[2], **Xiaolong Li**[1*], **Yao Zhao**[1]

[1]Institute of Information Science, Beijing Jiaotong University, Beijing, 100044, China
[2]School of Mathematical Sciences, Capital Normal University, Beijing, 100048, China
schen1307@foxmail.com, jiayingpeng.math@foxmail.com, lixl@bjtu.edu.cn, yzhao@bjtu.edu.cn

Abstract

Traditional gradient descent (GD) has been fully investigated for convex or L-smoothness functions, and it is widely utilized in current neural network optimization. The classical descent lemma ensures that for a function with L-smoothness, the GD trajectory converges stably towards the minimum when the learning rate is below $2/L$. This convergence is marked by a consistent reduction in the loss function throughout the iterations. However, recent experimental studies have demonstrated that even when the L-smoothness condition is not met, or if the learning rate is increased leading to oscillations in the loss function during iterations, the GD trajectory still exhibits convergence over the long run. This phenomenon is referred to as the unstable convergence regime of GD. In this paper, we present a theoretical perspective to offer a qualitative analysis of this phenomenon. The unstable convergence is in fact an inherent property of GD for general twice differentiable functions. Specifically, the forward-invariance of GD is established, i.e., it ensures that any point within a local region will always remain within this region under GD iteration. Then, based on the forward-invariance, for the initialization outside an open set containing the local minimum, the loss function will oscillate at the first several iterations and then become monotonely decreasing after the GD trajectory jumped into the open set. This work theoretically clarifies the unstable convergence phenomenon of GD discussed in previous experimental works. The unstable convergence of GD mainly depends on the selection of the initialization, and it is actually inevitable due to the complex nature of loss function.

Introduction

Gradient descent (GD) has been extensively studied in the literature, especially for convex or L-smoothness functions. This technique serves as a basis for current neural networks optimization. Given a loss function $f \in C^1(\mathcal{D})$ and an initialization $\boldsymbol{\theta}_0 \in \mathcal{D}$, where $\mathcal{D}$ is a subset of $\mathbb{R}^n$, the GD trajectory $\{\boldsymbol{\theta}_k\}_{k\geq 0}$ is iteratively generated as

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla f(\boldsymbol{\theta}_k) \tag{1}$$

where $\eta > 0$ is a pre-selected learning rate (LR). The function f is called L-smoothness, if there exists a constant L such that the following Lipschitz condition holds, for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{D}$,

$$\|\nabla f(\boldsymbol{\theta}_1) - \nabla f(\boldsymbol{\theta}_2)\| \leq L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|. \tag{2}$$

Besides the convexity or strong convexity, the L-smoothness condition is the most common assumption in traditional optimization (Nesterov 2003; Schmidt 2014; Bottou, Curtis, and Nocedal 2018; Martens 2020). Such assumption is also supposed to be valid for the loss function of neural networks. With the L-smoothness assumption, we have the following classical descent lemma,

$$f(\boldsymbol{\theta}_{k+1}) \leq f(\boldsymbol{\theta}_k) - \eta\left(1 - \frac{L\eta}{2}\right)\|\nabla f(\boldsymbol{\theta}_k)\|^2. \tag{3}$$

In this situation, we have $f(\boldsymbol{\theta}_{k+1}) \leq f(\boldsymbol{\theta}_k)$ if LR satisfies

$$\eta < 2/L. \tag{4}$$

It establishes a sufficient condition for loss reduction during each iteration of GD.This criterion leads to a "stable" convergence phase in GD, as discussed in (Ahn, Zhang, and Sra 2022). Additionally, for quadratic functions, condition (4) becomes indispensable to ensure the stability of the GD trajectory. In fact, when $\eta > 2/L$, the GD trajectory diverges for quadratic functions.The sufficient condition (4) considering L-smooth globally characterizes the loss function. Most

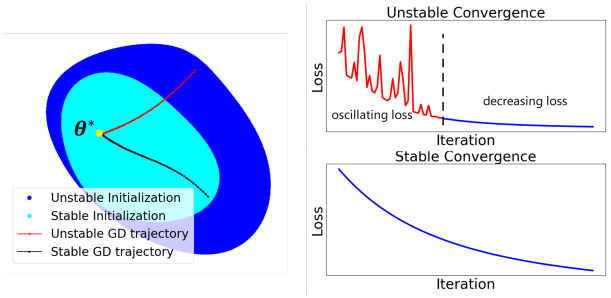

Figure 1: For loss function with a strict local minimum $\boldsymbol{\theta}^*$, all stable initializations constitute an open set marked in cyan. For the initialization outside the open set, its GD trajectory (red dots in the left figure) will ultimately jump into the open set and then stable convergence occurs, while the loss function will oscillate for the first several GD updates.

*Corresponding Author

existing studies determine LR based on the global Lipschitz constant, which theoretically results in a relatively conservative LR. However, recent experimental discoveries in deep learning have demonstrated that increasing LR can expedite the training of neural networks, alongside the convergence of the GD trajectory (Yaida 2018; Bjorck et al. 2018; Li, Wei, and Ma 2019; Zhu et al. 2019; Lewkowycz et al. 2020; Smith, Elsen, and De 2020; Thomas et al. 2020; Arora, Li, and Panigrahi 2022).

Earlier investigations focused on full-batch GD revealed this interesting observation, stating that in neural network training, the descent becomes unstable once the sharpness exceeds $2/\eta$ (Xing et al. 2018; Wu, Ma, and E 2018). Here, "sharpness" means the maximum eigenvalue of the Hessian at the present iteration. This instability manifests as escalating oscillations along the path of greatest curvature, while refraining from inducing complete divergence or stagnation. Subsequently, the concept of "Edge of Stability" (EoS) is introduced (Cohen et al. 2021), uncovering a captivating irregularity during neural network training. Specifically, while training a neural network with GD when the condition (4) falters, contrary to prevailing wisdom derived from convex optimization or the L-smoothness assumption, although the training loss may oscillate in the first several iterations, it can ultimately decrease and converge over time. The unstable convergence in GD is further investigated in (Ahn, Zhang, and Sra 2022). This study illuminates the triggers of unstable convergence, delineates the core characteristics that signify it in terms of the evolution of loss, iterations, and sharpness through GD updates. Additionally, from an alternate vantage point, (Lee and Jang 2023) introduces a novel sharpness metric known as interaction-aware sharpness (IAS). This metric captures the interplay between stochastic GD (SGD) and the loss landscape. The investigation defines stable and unstable sectors within the parameter space and dissects the progression from stable to unstable regions, elucidating the mechanics behind escaping the unstable realm. Besides the aforementioned works, the EoS phenomenon has been extensively studied recently (Arora, Li, and Panigrahi 2022; Wang, Li, and Li 2022; Zhu et al. 2023; Lee and Jang 2023), and more detailed introduction will be given later in the next section.

In this paper, following the work (Ahn, Zhang, and Sra 2022), we present a qualitative analysis for the unstable convergence regime of GD from a theoretical viewpoint. We prove that the unstable convergence is actually an inherent property of GD for general twice differentiable functions. Our main contributions can be summarized as follows (see Figure 1 for an illustration):

- The forward-invariance of GD is established. Specifically, for a strict local minimum $\boldsymbol{\theta}^*$ of $f \in C^2(\mathbb{R}^n)$, there exists an open set U containing $\boldsymbol{\theta}^*$ such that the GD iteration is invariant in U, i.e., $\boldsymbol{\theta} - \eta \nabla f(\boldsymbol{\theta}) \in U$ holds for any $\boldsymbol{\theta} \in U$, where η is a given LR leading to GD trajectory convergence to $\boldsymbol{\theta}^*$. This property provides a theoretical basis for our study.
- We further prove that, all initializations leading to GD trajectory convergence to $\boldsymbol{\theta}^*$ in a stable way (i.e., with decreasing loss at each GD iteration) constitute an open set denoted U_η. Once a GD trajectory converges to $\boldsymbol{\theta}^*$ with an initialization outside U_η, it will ultimately jumped into U_η and then the loss will permanently decrease (stable convergence occurs), while the loss is oscillating for the first several GD iterations. That is to say, the unstable convergence occurs just for the GD trajectory outside U_η. It clarifies the unstable convergence regime of GD theoretically, and provides a new insight for EoS. The unstable convergence of GD mainly depends on the selection of the initialization, and it is in fact inevitable due to the complex nature of loss function.

Related Work

The optimization dynamics of neural networks during training has garnered significant research attention. Prior research has explored the interaction between gradient distribution and loss landscape geometry, particularly concerning topics like escaping efficiency, stationarity, and convergence (Yaida 2018; Zhu et al. 2019; Thomas et al. 2020). They delve into this interplay in relation to escaping efficiency, stationarity, and convergence, respectively. However, these studies rely on specific assumptions like stochastic differential equation (SDE) approximations and the presence of stationary-state distributions. Notably, (Cohen et al. 2021) unveils a captivating phenomenon in full-batch GD, which is noteworthy and introduced by a concept EoS. Meanwhile, they also find that the sharpness of the loss landscape, quantified by the principal eigenvalue of the Hessian matrix, undergoes an incremental rise before stabilizing beyond a specific threshold through abundant experiments. These discoveries challenge conventional assumptions linking learning rates to convergence (Nesterov 2003; Schmidt 2014; Bottou, Curtis, and Nocedal 2018; Martens 2020). However, there are still uncharted areas in this domain. For example, the dynamics of sharpness and the consistent occurrence of stable optimization warrant further investigation. Additionally, the extension of this phenomenon to mini-batch SGD remains an open question worth exploring.

One of the later works, (Arora, Li, and Panigrahi 2022) demonstrates EoS with modified loss or normalized GD, affecting sharpness around the manifold. Besides, they also examine the implicit bias of sharpness near the minimum manifold with batch normal layers. It showcases the gradient trajectory into EoS and its impact on spherical sharpness. Finally, another work from (Lee and Jang 2023), distinct in its simplicity and lack of normalization, complements these studies. In contrast, their study introduces insights into the interaction without requiring these underlying conditions. Rather than focusing solely on full-batch GD scenarios through, they investigate dynamic behavior as IAS increases early in training, leading to an unstable phase. Moreover, their analysis extends to SGD settings. While prior work has analyzed convergence in full-batch GD, considering conditions such as L-smoothness and learning rate constraints (Nesterov 2003; Schmidt 2014; Bottou, Curtis, and Nocedal 2018; Martens 2020), these analyses may overlook crucial nuances in the interplay between optimization dynamics and loss landscape. This observation challenges

established stability assumptions in optimization, aligning with often-observed instability in practical settings.

Another perspective of research by (Lewkowycz et al. 2020; Wang et al. 2021) has focused on the implicit bias arising from high learning rates. They introduce a "catapult phase" akin to EoS, where loss remains bounded above sharpness thresholds. The work (Wang et al. 2021) analyzes matrix factorization with large learning rates, revealing two phases: oscillation and eventual monotonic loss decrease. This work also examine sharpness in a linear network with assumptions (Wang, Li, and Li 2022). This observation also intersects with two pivotal research avenues in neural network training: generalization and optimization. First, enhanced generalization has been associated with augmented learning rates in GD (Bjorck et al. 2018; Li, Wei, and Ma 2019; Lewkowycz et al. 2020; Smith, Elsen, and De 2020). Second, accumulating evidence suggests that minima with lower sharpness tend to exhibit superior generalization capabilities (Hochreiter and Schmidhuber 1997; Keskar et al. 2017). As a crucial aspect of neural network training, it also has been linked to factors like batch gradient distribution and loss landscape sharpness (Hochreiter and Schmidhuber 1997; Keskar et al. 2017; Hoffer, Hubara, and Soudry 2017; Jastrzębski et al. 2017; Smith et al. 2017; Zhu et al. 2019). Lee and Jang also contributes by establishing a novel link between batch gradient distribution and sharpness characteristics (Lee and Jang 2023). Prior studies have highlighted SGD's implicit bias toward improved generalization (Neyshabur et al. 2017), with factors including batch gradient distribution and loss landscape sharpness influencing generalization. This insight illuminates the mechanism driving SGD's ability to achieve enhanced generalization.

In order to provide an access to research as much as possible, many research on some specific examples has been proposed. For example, (Zhu et al. 2023) highlights the significance of sharpness concentration, adaptivity, and model architecture by connecting them with established areas of study. The work builds upon prior investigations into optimization algorithm behaviors, adaptive learning rates, and the trajectory evolution of over-parameterized models. The exploration of sharpness adaptivity and concentration offers novel insights into the intricacies of gradient descent during training. Examining model degrees, especially comparing degree-4 to degree-2, enhances understanding of how model complexity affects convergence patterns. The identification of fractal behavior and chaotic dynamics in optimization trajectory underscores the non-linear nature of training processes. Empirical observations linking the degree-4 model to real-world deep networks validate the theoretical findings' relevance. The open questions raises in the conclusion signal exciting directions for future research, addressing hidden dynamics, automatic coupling mechanisms, and leveraging fractal behaviors for global dynamics analysis. This synthesis forms a comprehensive foundation for the unique contributions of this study in advancing our understanding of optimization in deep learning. Another similar idea that focus on the local property of loss landscape is also contributive. Another work (Ma et al. 2022) extends the existing literature on the optimization of neural network loss functions by addressing the limitations of the quadratic approximation and emphasizing the importance of the multiscale structure. Their work contributes to the field by empirically demonstrating the subquadratic growth and separate scales structure, offering explanations for intriguing training phenomena. Additionally, they explore multiscale structure origins, revealing that non-convex models and uneven training data contribute to these complexities.

Main Results

Before presenting our main results, we first introduce the following theorem proved in (Ahn, Zhang, and Sra 2022). For clarity, throughout the following context, let $\lambda_{\max}(\boldsymbol{\theta})$ and $\lambda_{\min}(\boldsymbol{\theta})$ be the maximum and minimum eigenvalue of $\nabla^2 f(\boldsymbol{\theta})$, respectively. Moreover, as convention, $\boldsymbol{\theta}$ is called a stationary point if $\nabla f(\boldsymbol{\theta}) = 0$.

Theorem 1 (Theorem 1, Ahn, Zhang, and Sra 2022)**.** *Consider $f \in C^2(\mathcal{D})$, where $\mathcal{D}$ is a subset of $\mathbb{R}^n$. Suppose that for any subset $\mathcal{S} \subset \mathcal{D}$ with zero measure, $F^{-1}(\mathcal{S})$ is measure zero as well, where F is defined as*

$$F(\boldsymbol{\theta}) = \boldsymbol{\theta} - \eta \nabla f(\boldsymbol{\theta}) \tag{5}$$

which means that the forward GD update. Besides, for all stationary point $\boldsymbol{\theta}^*$, suppose that $1/\eta$ is not the eigenvalue of $\nabla^2 f(\boldsymbol{\theta}^*)$, and, $\lambda_{\max}(\boldsymbol{\theta}^*) > 2/\eta$ or $\lambda_{\min}(\boldsymbol{\theta}^*) < 0$. Then, there exists a measure zero subset $\mathcal{N} \subset \mathcal{D}$ such that for all initialization $\boldsymbol{\theta} \in \mathcal{D} - \mathcal{N}$, the GD trajectory will not converge to any of the stationary points in $\mathcal{D}$.

According to this theorem, since the convergence behavior of GD is our focus, for a considered local minimum $\boldsymbol{\theta}^*$, we adopt the following assumption that LR satisfies,

$$0 < \lambda_{\max}(\boldsymbol{\theta}^*) < 2/\eta \tag{6}$$

which is a necessary condition for the GD trajectory convergence.

Forward-Invariance of GD

Let us first introduce some notations and assumptions. Suppose that $\boldsymbol{\theta}^* \in \mathbb{R}^n$ is a strict local minimum of $f \in C^2(\mathbb{R}^n)$. That is, there exists an open ball $B(\boldsymbol{\theta}^*, t)$ centered at $\boldsymbol{\theta}^*$ with radius $t > 0$, such that $f(\boldsymbol{\theta}) > f(\boldsymbol{\theta}^*)$ holds for all $\boldsymbol{\theta} \in B(\boldsymbol{\theta}^*, t) \backslash \{\boldsymbol{\theta}^*\}$. Moreover, suppose that $\boldsymbol{\theta}^*$ is the only stationary point in $\overline{B(\boldsymbol{\theta}^*, t)}$, i.e., $\nabla f(\boldsymbol{\theta}) \neq 0$ for any $\boldsymbol{\theta} \in \overline{B(\boldsymbol{\theta}^*, t)} \backslash \{\boldsymbol{\theta}^*\}$. This setup ensures the existence of a region in the vicinity of $\boldsymbol{\theta}^*$ where we can analyze the behavior of GD, and no further conditions such as the convexity or L-smoothness are required. We try to understand the GD behavior with as few as possible conditions to derive general theoretical results.

The central pillar of our work is the establishment of the forward-invariance property of GD around $\boldsymbol{\theta}^*$. We summarize our result in the following theorem.

Theorem 2. *Consider the above defined f, $\boldsymbol{\theta}^*$ and η. There exists an open set $U \subset B(\boldsymbol{\theta}^*, t)$ containing $\boldsymbol{\theta}^*$, such that the following properties hold:*

- ***(forward-invariance)**: $\boldsymbol{\theta} \in U \Longrightarrow \boldsymbol{\theta} - \eta \nabla f(\boldsymbol{\theta}) \in U$,*

- ***(descent property)**: there exists a constant $C > 0$ such that for each $\boldsymbol{\theta} \in U$,*

$$f(\boldsymbol{\theta} - \eta \nabla f(\boldsymbol{\theta})) - f(\boldsymbol{\theta}) \leq -C\eta \|\nabla f(\boldsymbol{\theta})\|^2. \quad (7)$$

This theorem has significant implications for understanding the unstable convergence of GD. When a GD trajectory $\{\boldsymbol{\theta}_k\}_{k\geq 0}$ converges to $\boldsymbol{\theta}^*$, as U is open, it is guaranteed that there exists $K \geq 0$ such that $\boldsymbol{\theta}_K \in U$. Once the GD trajectory enters U starting from $\boldsymbol{\theta}_K$, it remains in U for all $k \geq K$. This leads to the inequality (7), indicating a decrease in the loss function with each iteration. Whereas, generally speaking, the loss oscillation may occur only for the sub-trajectory $(\boldsymbol{\theta}_0, ..., \boldsymbol{\theta}_{K-1})$. Moreover, the importance of the forward-invariance is empirically observed in (Ahn, Zhang, and Sra 2022), and it is one main cause for the EoS phenomenon of neural networks. Here, in a general setup, we provide a rigorous proof for the existence of the forward-invariance near the local minimum. Besides, in previous works such as (Panageas and Piliouras 2016), without theoretical guarantee, the forward-invariance is considered as an assumption and only verified through experiments.

To prove Theorem 2, we first establish a crucial Lemma.

Lemma 1. *Consider the above defined f, $\boldsymbol{\theta}^*$ and η. There exist a constant $C > 0$ and a parameter $0 < t' < t$ such that for all $\boldsymbol{\theta} \in B(\boldsymbol{\theta}^*, t')$ and $\alpha \in [0, \eta]$,*

$$f(\boldsymbol{\theta} - \alpha \nabla f(\boldsymbol{\theta})) - f(\boldsymbol{\theta}) \leq -C\alpha \|\nabla f(\boldsymbol{\theta})\|^2. \quad (8)$$

Proof. The core step is applying the following second-order Taylor expansion,

$$\begin{aligned} &f(\boldsymbol{\theta} - \alpha \nabla f(\boldsymbol{\theta})) = f(\boldsymbol{\theta}) - \alpha \|\nabla f(\boldsymbol{\theta})\|^2 + \\ &\alpha^2 \int_0^1 (1-x) \nabla f(\boldsymbol{\theta})^t \nabla^2 f(\boldsymbol{\theta} - x\alpha \nabla f(\boldsymbol{\theta})) \nabla f(\boldsymbol{\theta}) dx. \end{aligned} \quad (9)$$

We first prove that

$$g(s) \triangleq \max_{\|\boldsymbol{e}\|=1, \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq s} \boldsymbol{e}^t \nabla^2 f(\boldsymbol{\theta}) \boldsymbol{e} \quad (10)$$

is continuous on $[0, t]$. Obviously, g is increasing and

$$g(0) = \max_{\|\boldsymbol{e}\|=1} \boldsymbol{e}^t \nabla^2 f(\boldsymbol{\theta}^*) \boldsymbol{e} = \lambda_{\max}(\boldsymbol{\theta}^*). \quad (11)$$

Since $\boldsymbol{\theta}^*$ is a local minimum, $\nabla^2 f(\boldsymbol{\theta}^*)$ is semi-positive definite, hence

$$g(0) = \lambda_{\max}(\boldsymbol{\theta}^*) \geq 0. \quad (12)$$

In the following proof, without loss of generality, we assume $\boldsymbol{\theta}^* = 0$ for the sake of simplicity Noticed that, for all $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{e}$ with $\|\boldsymbol{e}\| = 1$, based on the Cauchy-Schwarz inequality, we have

$$|\boldsymbol{e}^t \nabla^2 f(\boldsymbol{\theta}_1) \boldsymbol{e} - \boldsymbol{e}^t \nabla^2 f(\boldsymbol{\theta}_2) \boldsymbol{e}| \leq \|\nabla^2 f(\boldsymbol{\theta}_1) - \nabla^2 f(\boldsymbol{\theta}_2)\|. \quad (13)$$

Here, in the right side of the above equation, the matrix norm is defined as the Frobenius norm. Then, since $\nabla^2 f$ is continuous on the compact set $\overline{B(\boldsymbol{\theta}^*, t)}$, $\nabla^2 f$ is uniformly continuous on this closed ball. That is to say, for a given $\varepsilon > 0$, there exists $\delta > 0$ such that for all $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ satisfies $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \leq \delta$,

$$\|\nabla^2 f(\boldsymbol{\theta}_1) - \nabla^2 f(\boldsymbol{\theta}_2)\| \leq \varepsilon. \quad (14)$$

As a result of (13) and (14), we have, for all $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \leq \delta$ and $\|\boldsymbol{e}\| = 1$,

$$\boldsymbol{e}^t \nabla^2 f(\boldsymbol{\theta}_1) \boldsymbol{e} \leq \boldsymbol{e}^t \nabla^2 f(\boldsymbol{\theta}_2) \boldsymbol{e} + \varepsilon. \quad (15)$$

Next, take

$$\boldsymbol{\theta}_2 = \frac{s}{s+\delta} \boldsymbol{\theta}_1. \quad (16)$$

It is then easy to verify that $\|\boldsymbol{\theta}_2\| \leq s$ and $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \leq \delta$ hold if $\|\boldsymbol{\theta}_1\| \leq s + \delta$. By applying (15) with (16), we have that for any $\|\boldsymbol{\theta}_1\| \leq s + \delta$ and $\|\boldsymbol{e}\| = 1$,

$$\boldsymbol{e}^t \nabla^2 f(\boldsymbol{\theta}_1) \boldsymbol{e} \leq g(s) + \varepsilon. \quad (17)$$

This yields, for all $s' \leq s + \delta$,

$$g(s') \leq g(s) + \varepsilon. \quad (18)$$

Notice that (18) is independent of s, and then the monotonicity of g promised its continuity on $[0, t]$. At this juncture, we select a constant M satisfying

$$\eta < \frac{2}{M} < \frac{2}{\lambda_{\max}(\boldsymbol{\theta}^*)}. \quad (19)$$

Since $M > \lambda_{\max}(\boldsymbol{\theta}^*) = g(0)$, then based on the continuity of g, there exists $t_1 \in (0, t)$ satisfying $g(t_1) \leq M$. Moreover, by the continuity of ∇f and the following estimation

$$\|\boldsymbol{\theta} - \alpha \nabla f(\boldsymbol{\theta})\| \leq \|\boldsymbol{\theta}\| + \alpha \|\nabla f(\boldsymbol{\theta})\| \quad (20)$$

there exists $t_2 \in (0, t_1)$ such that $\boldsymbol{\theta} - \alpha \nabla f(\boldsymbol{\theta}) \in \overline{B(\boldsymbol{\theta}^*, t_1)}$ holds for all $\boldsymbol{\theta} \in \overline{B(\boldsymbol{\theta}^*, t_2)}$ and $\alpha \in [0, \eta]$. Finally, revisiting (9), we have, for all $\boldsymbol{\theta} \in \overline{B(\boldsymbol{\theta}^*, t_2)}$, $\alpha \in [0, \eta]$ and $x \in [0, 1]$,

$$\nabla f(\boldsymbol{\theta})^t \nabla^2 f(\boldsymbol{\theta} - x\alpha \nabla f(\boldsymbol{\theta})) \nabla f(\boldsymbol{\theta}) \leq g(t_1) \|\nabla f(\boldsymbol{\theta})\|^2 \quad (21)$$

and thus

$$\begin{aligned} &f(\boldsymbol{\theta} - \alpha \nabla f(\boldsymbol{\theta})) - f(\boldsymbol{\theta}) \\ &\leq -\alpha \|\nabla f(\boldsymbol{\theta})\|^2 + \alpha^2 \|\nabla f(\boldsymbol{\theta})\|^2 \int_0^1 (1-x) g(t_1) dx \\ &\leq \alpha \|\nabla f(\boldsymbol{\theta})\|^2 \left(-1 + \eta \frac{M}{2} \right) = -C\alpha \|\nabla f(\boldsymbol{\theta})\|^2 \end{aligned} \quad (22)$$

where $C = 1 - \eta M/2$ is a constant. Take then $t' = t_2$, the theorem is finally proved. □

We now prove Lemma 2.

Proof of Theorem 2. For t' defined in Lemma 1, let

$$f^* = \min_{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| = t'} f(\boldsymbol{\theta}). \quad (23)$$

By the assumption of $\boldsymbol{\theta}^*$, we have $f^* > f(\boldsymbol{\theta}^*)$. We then define

$$U = \{\boldsymbol{\theta} \in \mathbb{R}^n : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < t',\ f(\boldsymbol{\theta}) < f^*\}. \quad (24)$$

Obviously, $U \subset B(\boldsymbol{\theta}^*, t')$ is an open set containing $\boldsymbol{\theta}^*$. We now prove that the set U meets the condition of Theorem 2 . For a given $\boldsymbol{\theta} \in U$, consider here

$$A_{\boldsymbol{\theta}} = \{\alpha > 0 : \boldsymbol{\theta} - \beta \nabla f(\boldsymbol{\theta}) \in U \text{ for all } \beta \in [0, \alpha]\} \quad (25)$$

and define

$$\alpha^* = \sup_{\alpha \in A_{\boldsymbol{\theta}}} \alpha, \quad \widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta} - \alpha^* \nabla f(\boldsymbol{\theta}). \tag{26}$$

As U is an open set, $\alpha^* > 0$. Moreover, by the definition of $A_{\boldsymbol{\theta}}$, the following conditions hold for each $\alpha \in A_{\boldsymbol{\theta}}$,

$$\|(\boldsymbol{\theta} - \alpha \nabla f(\boldsymbol{\theta})) - \boldsymbol{\theta}^*\| < t', \ f(\boldsymbol{\theta} - \alpha \nabla f(\boldsymbol{\theta})) < f^*. \tag{27}$$

Then,

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \leq t' \quad \text{and} \quad f(\widetilde{\boldsymbol{\theta}}) \leq f^*. \tag{28}$$

If $f(\widetilde{\boldsymbol{\theta}}) < f^*$, by the definition of f^*, we see that $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| < t'$ must hold. In this situation, there exists $\alpha' > \alpha^*$ with

$$\|(\boldsymbol{\theta} - \alpha' \nabla f(\boldsymbol{\theta})) - \boldsymbol{\theta}^*\| < t', \ f(\boldsymbol{\theta} - \alpha' \nabla f(\boldsymbol{\theta})) < f^* \tag{29}$$

which contradicts to the definition of α^*. Then, $f(\widetilde{\boldsymbol{\theta}}) = f^*$ must hold. Moreover, based on Lemma 1, we have $f(\widetilde{\boldsymbol{\theta}}) - f(\boldsymbol{\theta}) \leq 0$ if $\eta \geq \alpha^*$. That is to say, $f(\boldsymbol{\theta}) \geq f(\widetilde{\boldsymbol{\theta}}) = f^*$, which contradicts to the definition of U. We then arrived the result that $\eta < \alpha^*$, and thus $\boldsymbol{\theta} - \eta \nabla f(\boldsymbol{\theta}) \in U$ is proved by the definition of α^*. Here we finished the proof for the forward-invariance of U. In addition, clearly, the descent property of U is a direct result of Lemma 1, and thus the theorem is finally proved. □

Stable and Unstable Convergence of GD

Based on the foundational results of Theorem 2, we explore the convergence behavior of GD. We first define the stable initializations as the following set

$$\begin{aligned} U_\eta = \{&\boldsymbol{\theta}_0 \in \mathbb{R}^n : \lim_{k \to \infty} \boldsymbol{\theta}_k = \boldsymbol{\theta}^*, \\ &f(\boldsymbol{\theta}_k) < f(\boldsymbol{\theta}_{k+1}) \text{ for all } k \in \mathbb{N} \text{ if } \boldsymbol{\theta}_k \neq \boldsymbol{\theta}^*\}. \end{aligned} \tag{30}$$

That is to say, U_η is composed of all initializations such that the GD trajectory converges stably towards $\boldsymbol{\theta}^*$, meaning that the loss function will monotonically decrease in each GD iteration. Particularly, if $\boldsymbol{\theta}_k = \boldsymbol{\theta}^*$, then $\{\boldsymbol{\theta}_k\}_{k \geq K}$ remains fixed at $\boldsymbol{\theta}^*$, rendering the GD trajectory a finite sequence. This set defines the stable convergence regime of GD. In addition, we also define all initializations leading to GD trajectory convergence as the following set

$$V_\eta = \left\{ \boldsymbol{\theta}_0 \in \mathbb{R}^n : \lim_{k \to \infty} \boldsymbol{\theta}_k = \boldsymbol{\theta}^* \right\}. \tag{31}$$

Clearly, $\boldsymbol{\theta}^* \in U_\eta \subset V_\eta$, and our main result is the following theorem.

Theorem 3. *The set U_η is an open set.*

Based on this theorem, the unstable convergence of GD is clear. For any initialization $\boldsymbol{\theta}_0 \in V_\eta \backslash U_\eta$, as U_η is open containing $\boldsymbol{\theta}^*$, there exists $K \geq 0$ such that $\boldsymbol{\theta}_K \in U_\eta$, meaning that the GD trajectory will ultimately jumped into U_η. Once the GD trajectory enters U starting from $\boldsymbol{\theta}_K$, the loss decreases with each step, and the GD trajectory enters the stable convergence regime. Whereas, by definition, the loss oscillation occurs for the sub-trajectory $(\boldsymbol{\theta}_0, ..., \boldsymbol{\theta}_{K-1})$, indicating the unstable convergence regime.

To prove Theorem 3, we need the following lemma.

Lemma 2. *Consider the above defined f, $\boldsymbol{\theta}^*$ and η. For the open set U defined in Theorem 2, the GD trajectory $\{\boldsymbol{\theta}_k\}_{k \geq 0}$ converges stably to $\boldsymbol{\theta}^*$ if the initialization $\boldsymbol{\theta}_0 \in U$.*

Proof. According to Theorem 2, $\{\boldsymbol{\theta}_k\}_{k \geq 0}$ will not jump over U if $\boldsymbol{\theta}_0 \in U$. Then according to (7), we have

$$f(\boldsymbol{\theta}_{k+1}) - f(\boldsymbol{\theta}_k) \leq -C\eta \|\nabla f(\boldsymbol{\theta}_k)\|^2 \tag{32}$$

and thus, by summing up on k,

$$\sum_{k=0}^{N} (f(\boldsymbol{\theta}_{k+1}) - f(\boldsymbol{\theta}_k)) \leq -C\eta \sum_{k=0}^{N} \|\nabla f(\boldsymbol{\theta}_k)\|^2. \tag{33}$$

Consequently,

$$C\eta \sum_{k=0}^{N} \|\nabla f(\boldsymbol{\theta}_k)\|^2 \leq f(\boldsymbol{\theta}_0) - f(\boldsymbol{\theta}_{N+1}) \leq f(\boldsymbol{\theta}_0) - f(\boldsymbol{\theta}^*). \tag{34}$$

Therefore,

$$\lim_{k \to \infty} \|\nabla f(\boldsymbol{\theta}_k)\| = 0. \tag{35}$$

We now prove that $\lim_{k \to \infty} \boldsymbol{\theta}_k = \boldsymbol{\theta}^*$. Otherwise, there exists a subsequence $\{\boldsymbol{\theta}_{n_k}\}_{k \geq 0}$ such that $\|\boldsymbol{\theta}_{n_k} - \boldsymbol{\theta}^*\| \geq \varepsilon$ holds for all $k \in \mathbb{N}$, where $\varepsilon > 0$ is a given constant. As $\{\boldsymbol{\theta}_{n_k}\}_{k \geq 0}$ is contained in the compact set $\overline{B(\boldsymbol{\theta}^*, t)}$, it has a subsequence $\{\boldsymbol{\theta}_{m_k}\}_{k \geq 0}$ with $\lim_{k \to \infty} \boldsymbol{\theta}_{m_k} = \boldsymbol{\theta}' \in \overline{B(\boldsymbol{\theta}^*, t)}$. Recall (35), we know that $\|\nabla f(\boldsymbol{\theta}')\| = 0$. Thus, $\boldsymbol{\theta}' = \boldsymbol{\theta}^*$ must hold, which is a contradiction. In this way, we finally proved that $\lim_{k \to \infty} \boldsymbol{\theta}_k = \boldsymbol{\theta}^*$. □

By this lemma and the definition of U_η in (30), it is clear that $U \subset U_\eta$. It is the key issue for proving Theorem 3. We now introduce the proof as follows.

Proof of Theorem 3. For any $\boldsymbol{\theta}_0 \in U_\eta$, we will prove that $B(\boldsymbol{\theta}_0, \varepsilon) \subset U_\eta$ holds for some $\varepsilon > 0$. This fact is obvious if $\boldsymbol{\theta}_0 \in U$. Otherwise, as $\{\boldsymbol{\theta}_k\}_{k \geq 0}$ converges to $\boldsymbol{\theta}^*$, there exists $K > 0$ such that $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{K-1} \notin U$ while $\boldsymbol{\theta}_K \in U$. We then define

$$\begin{cases} M = \min\limits_{0 \leq i \leq K-1} f(\boldsymbol{\theta}_i) - f(\boldsymbol{\theta}_{i+1}) \\ M_1 = \max\limits_{0 \leq i \leq K, \|\boldsymbol{\tau}\| \leq 1} \|f(\boldsymbol{\theta}_i + \boldsymbol{\tau})\| \\ M_2 = \max\limits_{0 \leq i \leq K, \|\boldsymbol{\tau}\| \leq 1} \|\nabla^2 f(\boldsymbol{\theta}_i + \boldsymbol{\tau})\| \end{cases}. \tag{36}$$

Obviously, M, M_1 and M_2 are strictly positive. Next, we claim that, for any $\boldsymbol{\theta}_0' \in B(\boldsymbol{\theta}_0, \varepsilon)$ with

$$\varepsilon \leq 1/(1 + \eta M_2)^K \tag{37}$$

the GD trajectory $\{\boldsymbol{\theta}_k'\}_{k \geq 0}$ with initialization $\boldsymbol{\theta}_0'$ satisfies, for each $0 \leq i \leq K$,

$$\|\boldsymbol{\theta}_i' - \boldsymbol{\theta}_i\| < \varepsilon (1 + \eta M_2)^i. \tag{38}$$

Actually, the above inequality can be proved by induction

with the following estimation, for $i < K$,

$$\begin{aligned}
&\|\boldsymbol{\theta}'_{i+1} - \boldsymbol{\theta}_{i+1}\| \\
&= \|\boldsymbol{\theta}'_i - \boldsymbol{\theta}_i - \eta(\nabla f(\boldsymbol{\theta}'_i) - \nabla f(\boldsymbol{\theta}_i))\| \\
&\le \|\boldsymbol{\theta}'_i - \boldsymbol{\theta}_i\| + \eta \left\| \int_0^1 \nabla^2 f(\boldsymbol{\theta}_i + x(\boldsymbol{\theta}'_i - \boldsymbol{\theta}_i))(\boldsymbol{\theta}'_i - \boldsymbol{\theta}_i) dx \right\| \\
&\le \|\boldsymbol{\theta}'_i - \boldsymbol{\theta}_i\| + \eta \int_0^1 \|\nabla^2 f(\boldsymbol{\theta}_i + x(\boldsymbol{\theta}'_i - \boldsymbol{\theta}_i))(\boldsymbol{\theta}'_i - \boldsymbol{\theta}_i)\| dx \\
&\le \|\boldsymbol{\theta}'_i - \boldsymbol{\theta}_i\| + \eta \int_0^1 \|\nabla^2 f(\boldsymbol{\theta}_i + x(\boldsymbol{\theta}'_i - \boldsymbol{\theta}_i))\| \|\boldsymbol{\theta}'_i - \boldsymbol{\theta}_i\| dx \\
&\le (1 + \eta M_2)\|\boldsymbol{\theta}'_i - \boldsymbol{\theta}_i\|.
\end{aligned} \tag{39}$$

Moreover, we have, for $i < K$,

$$\begin{aligned}
&f(\boldsymbol{\theta}'_{i+1}) - f(\boldsymbol{\theta}'_i) \\
&= f(\boldsymbol{\theta}'_{i+1}) - f(\boldsymbol{\theta}_{i+1}) + f(\boldsymbol{\theta}_{i+1}) - f(\boldsymbol{\theta}_i) + f(\boldsymbol{\theta}_i) - f(\boldsymbol{\theta}'_i) \\
&\le \int_0^1 |(\boldsymbol{\theta}'_{i+1} - \boldsymbol{\theta}_{i+1})^t \nabla f(\boldsymbol{\theta}_{i+1} + x(\boldsymbol{\theta}'_{i+1} - \boldsymbol{\theta}_{i+1}))| dx \\
&\qquad - M + \int_0^1 |(\boldsymbol{\theta}'_i - \boldsymbol{\theta}_i)^t \nabla f(\boldsymbol{\theta}_i + x(\boldsymbol{\theta}'_i - \boldsymbol{\theta}_i))| dx \\
&\le M_1(\|\boldsymbol{\theta}'_{i+1} - \boldsymbol{\theta}_{i+1}\| + \|\boldsymbol{\theta}'_i - \boldsymbol{\theta}_i\|) - M
\end{aligned} \tag{40}$$

which says that, according to (38),

$$f(\boldsymbol{\theta}'_{i+1}) - f(\boldsymbol{\theta}'_i) \le 2\varepsilon M_1(1 + \eta M_2)^K - M. \tag{41}$$

On the other hand, as $\boldsymbol{\theta}_K \in U$ with U open, there exists $\varepsilon' > 0$ such that $B(\boldsymbol{\theta}_K, \varepsilon') \subset U$. Then, as a result of (38) and (41), by taking sufficiently small $\varepsilon > 0$ which satisfies

$$\varepsilon(1 + \eta M_2)^K < \varepsilon' \quad \text{and} \quad 2\varepsilon M_1(1 + \eta M_2)^K < M \tag{42}$$

we know that for each initialization $\boldsymbol{\theta}'_0 \in B(\boldsymbol{\theta}_0, \varepsilon)$, the GD trajectory $\{\boldsymbol{\theta}'_k\}_{k \ge 0}$ enters U with $\boldsymbol{\theta}_K \in U$, and simultaneously, $f(\boldsymbol{\theta}'_{i+1}) < f(\boldsymbol{\theta}'_i)$ holds for each $i < K$. The theorem is then finally proved based on Lemma 2. □

Finally, before closing this subsection, we claim that the set V_η defined in (31) is also an open set. The proof is almost the same as that of Theorem 3.

Numerical Experiments

In Theorem 2 and Theorem 3, we provide theoretical discussions concerning the forward-invariance, loss reduction, and convergence of GD. Especially near the minimum point, a distinct trend of decreasing convergence becomes prominent. The first row of Figure 2 (six figures) shows an example for the following function of two variables

$$f_1(x, y) = x^2 + y^2 + x^2 y + x^3 + y^3. \tag{43}$$

It has a strict local minimum $\boldsymbol{\theta}^* = (0, 0)$ with $\lambda_{\max}(\boldsymbol{\theta}^*) = 2$, and thus the maximum LR is 1. For the first figure of the six, it shows the landscape of f_1 near $\boldsymbol{\theta}$, and the remaining five demonstrate the convergence patterns of GD when initiated with varying initializations and LRs. Specifically, the set of stable initializations U_η is depicted in cyan, while the set of unstable initializations, i.e., $V_\eta \backslash U_\eta$, is shown in blue. When LR is chosen to be relatively small, GD is performed with all the points in the graph as initializations and the function values are monotonically decreasing, i.e., stable convergence. As LR increases, unstable convergence occurs, while the set of initializations leading to unstable convergence expands. The same behavior is observed for

$$f_2(x, y) = x^2 + y^2 + x^2 y + x^4 + y^4, \tag{44}$$

where it has a unique global minimum $\boldsymbol{\theta}^* = (0, 0)$ with $\lambda_{\max}(\boldsymbol{\theta}^*) = 2$ (and thus, the maximum LR is 1).

Notably, the unstable initialization configurations depicted in Figure 2 exhibit irregular and even disconnected patterns, implying that the dynamics within the analyzed open set U_η could be considerably intricate. The third and fourth rows of Figure 2 demonstrate for the other tow functions f_3 and f_4 with complex sets of initializations for stable and unstable convergence, where

$$f_3(x, y) = x^4 y^2 + x^2 y^4 - 3x^2 y^2 + 1 \tag{45}$$

and

$$f_4(x, y) = (1 - xy)^2 + \frac{1}{20}(x^2 + y^2). \tag{46}$$

Here, f_3 has four global minimums $(\pm 1, \pm 1)$ with the same $\lambda_{\max} = 12$ (and thus, the maximum LR is $1/6$). A distinctive feature of this function is that, each point lies in $x-$ or $y-$coordinate is stationary while not a minimum. For f_4, it has two global minimums $(\sqrt{19}/(2\sqrt{5}), \sqrt{19}/(2\sqrt{5}))$ and $(-\sqrt{19}/(2\sqrt{5}), -\sqrt{19}/(2\sqrt{5}))$ with the same $\lambda_{\max} = 19/5$ (and thus, the maximum LR is $10/19$). Notice that, in these cases with multiples minimums, all the initializations converge to any of the minimums are considered and marked.

An increasing LR maintains an analogous trend in convergence behavior. However, the region's configuration becomes progressively intricate. This intricacy underscores the profound complexity surrounding the issue of convergence states. Nevertheless, the existence and inevitability of the stable convergence area remain undeniable. After passing through the unstable convergence stage, the system enters a state of stable convergence. In Figure 3, both the GD trajectory and the loss changes are displayed for the two types of initializations. The irregular loss oscillations seen in the previous section ultimately dissipate with iterations. The trajectory transitions from an unstable area to a stable one, finally converging to a local minimum. The figure provides a visual representation of this convergence process, where the red line signifies instability and the black line signifies stability. By comprehensively analyzing these convergence types along with the loss curves, we gained valuable insights into the behavior of our optimization process for the chosen functions. This approach enhanced our understanding of how the learning rate affects convergence stability and illuminated the complexities inherent in non-convex loss landscapes.

Further details regarding the experiments can be found in the supplementary materials.

Extension

The above studies with a strictly local minimum can be extended to the case that the set of global minimums is compact. The detailed introduction can be found in the supplementary material due space limitation.

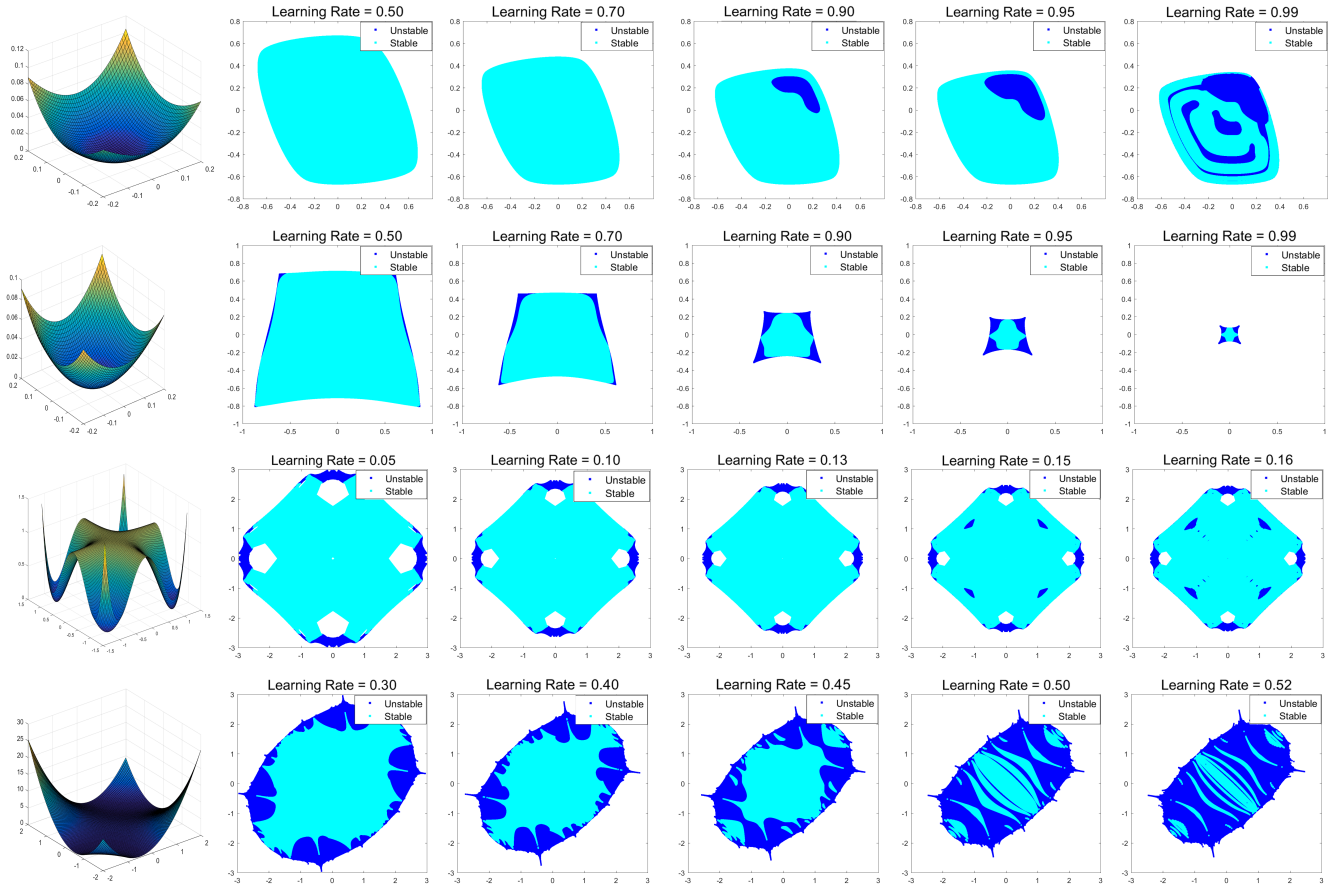

Figure 2: For each row of this figure, the left one shows the landscape of a given loss functions near local minimum(s), and the right five show the corresponding sets U_η (the cyan points, i.e., the stable initializations) and V_η (the cyan and blue points, i.e, the blue points are the unstable initializations) under gradually increasing LR reaching its maximum.

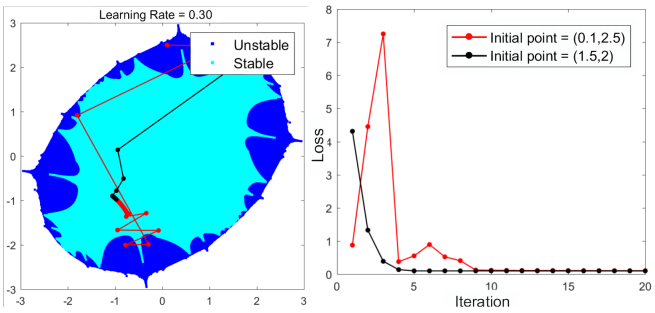

Figure 3: GD trajectory and the loss changes with an unstable initialization $(0.1, 2.5)$ and a stable initialization $(1.5, 2)$ for f_4 defined in (46). The red line represents unstable and the black line represents stable.

Conclusion

In summary, we present a qualitative analysis for the unstable convergence regime of GD from a theoretical viewpoint with forward-invariance. We now introduce several future directions:

- The main limitation of our work is the assumption for the minimums, it cannot handle the loss function with uncompact global minimums such as $f(x, y) = (1 - xy)^2$. The technical difficulty is proving the forward-invariance in subtle cases. Moreover, extending the proposed results to SGD is also a worthy future direction.
- Besides the unstable convergence, EoS focus on the sharpness of the GD trajectory, and found that the sharpness oscillates around the stability threshold $2/\eta$ while the loss still continues to decrease in the long run. An in-depth analysis on sharpness along the GD trajectory is very helpful for understanding the mechanism of GD.
- Committed to deriving general results with minimal assumptions on the loss function, we address diverse neural network architectures and activations. Future emphasis will be on analyzing how varied networks impact convergence and the EoS phenomenon.
- This work exclusively delves into qualitative analysis of GD convergence. Future emphasis should include quantitative aspects like convergence speed, demanding a thorough exploration of valid assumptions in neural network optimization.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 61972031, 62261160653 and 62372037, and in part by the Shandong Provincial Natural Science Foundation under Grant ZR2022LZH011.

References

Ahn, K.; Zhang, J.; and Sra, S. 2022. Understanding the unstable convergence of gradient descent. In *International Conference on Machine Learning*.

Arora, S.; Li, Z.; and Panigrahi, A. 2022. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, 948–1024.

Bjorck, N.; Gomes, C. P.; Selman, B.; and Weinberger, K. Q. 2018. Understanding batch normalization. *Advances in neural information processing systems*, 31.

Bottou, L.; Curtis, F. E.; and Nocedal, J. 2018. Optimization methods for large-scale machine learning. *SIAM review*, 60(2): 223–311.

Cohen, J.; Kaur, S.; Li, Y.; Kolter, J. Z.; and Talwalkar, A. 2021. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. In *International Conference on Learning Representations*.

Hochreiter, S.; and Schmidhuber, J. 1997. Flat minima. *Neural computation*, 9(1): 1–42.

Hoffer, E.; Hubara, I.; and Soudry, D. 2017. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30.

Jastrzębski, S.; Kenton, Z.; Arpit, D.; Ballas, N.; Fischer, A.; Bengio, Y.; and Storkey, A. 2017. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*.

Keskar, N. S.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; and Tang, P. T. P. 2017. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *International Conference on Learning Representations*.

Lee, S.; and Jang, C. 2023. A new characterization of the edge of stability based on a sharpness measure aware of batch gradient distribution. In *International Conference on Learning Representations*.

Lewkowycz, A.; Bahri, Y.; Dyer, E.; Sohl-Dickstein, J.; and Gur-Ari, G. 2020. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*.

Li, Y.; Wei, C.; and Ma, T. 2019. Towards explaining the regularization effect of initial large learning rate in training neural networks. *Advances in Neural Information Processing Systems*, 32.

Ma, C.; Kunin, D.; Wu, L.; and Ying, L. 2022. Beyond the quadratic approximation: the multiscale structure of neural network loss landscapes. *JMLR*.

Martens, J. 2020. New insights and perspectives on the natural gradient method. *The Journal of Machine Learning Research*, 21(1): 5776–5851.

Nesterov, Y. 2003. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

Neyshabur, B.; Bhojanapalli, S.; McAllester, D.; and Srebro, N. 2017. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30.

Panageas, I.; and Piliouras, G. 2016. Gradient Descent Only Converges to Minimizers: Non-Isolated Critical Points and Invariant Regions.

Schmidt, M. 2014. Convergence rate of stochastic gradient with constant step size.

Smith, S.; Elsen, E.; and De, S. 2020. On the generalization benefit of noise in stochastic gradient descent. In *International Conference on Machine Learning*, 9058–9067.

Smith, S. L.; Kindermans, P.-J.; Ying, C.; and Le, Q. V. 2017. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*.

Thomas, V.; Pedregosa, F.; Merriënboer, B.; Manzagol, P.-A.; Bengio, Y.; and Le Roux, N. 2020. On the interplay between noise and curvature and its effect on optimization and generalization. In *International Conference on Artificial Intelligence and Statistics*.

Wang, Y.; Chen, M.; Zhao, T.; and Tao, M. 2021. Large Learning Rate Tames Homogeneity: Convergence and Balancing Effect. In *International Conference on Learning Representations*.

Wang, Z.; Li, Z.; and Li, J. 2022. Analyzing sharpness along gd trajectory: Progressive sharpening and edge of stability. *Advances in Neural Information Processing Systems*, 35: 9983–9994.

Wu, L.; Ma, C.; and E, W. 2018. How SGD Selects the Global Minima in Over-parameterized Learning: A Dynamical Stability Perspective. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Xing, C.; Arpit, D.; Tsirigotis, C.; and Bengio, Y. 2018. A Walk with SGD. arXiv:1802.08770.

Yaida, S. 2018. Fluctuation-dissipation relations for stochastic gradient descent. In *International Conference on Learning Representations*.

Zhu, X.; Wang, Z.; Wang, X.; Zhou, M.; and Ge, R. 2023. Understanding Edge-of-Stability Training Dynamics with a Minimalist Example. In *International Conference on Learning Representations*.

Zhu, Z.; Wu, J.; Yu, B.; Wu, L.; and Ma, J. 2019. The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping from Sharp Minima and Regularization Effects. In *International Conference on Machine Learning*, 7654–7663.