Understanding the Unstable Convergence of Gradient Descent

Kwangjun Ahn [1 2] **Jingzhao Zhang** [3] **Suvrit Sra** [1]

Abstract

Most existing analyses of (stochastic) gradient descent rely on the condition that for L-smooth costs, the step size is less than $2/L$. However, many works have observed that in machine learning applications step sizes often do not fulfill this condition, yet (stochastic) gradient descent still converges, albeit in an unstable manner. We investigate this unstable convergence phenomenon from first principles, and discuss key causes behind it. We also identify its main characteristics, and how they interrelate based on both theory and experiments, offering a principled view toward understanding the phenomenon.

1. Introduction

Gradient descent (GD) runs the iteration

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta \nabla f(\boldsymbol{\theta}^t)\,,$$

seeking to optimize a cost function f. It also provides a conceptual foundation for stochastic gradient descent (SGD), one of the key algorithms in modern machine learning. A vast body of literature that analyzes (S)GD assumes that the cost f is L-smooth (we say f is L-smooth if $\|\nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}')\| \leq L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}'$), and subsequently exploits the associated "*descent lemma*":

$$f(\boldsymbol{\theta}^{t+1}) \leq f(\boldsymbol{\theta}^t) - \eta\big(1 - L\frac{\eta}{2}\big)\left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2 . \tag{1.1}$$

To ensure descent via inequality (1.1), the condition

$$L < \frac{2}{\eta}, \tag{1.2}$$

is imposed. This condition ensure that GD decreases the cost f at each iteration. Whenever condition (1.2) holds, we call it the **"stable" regime** in this paper.

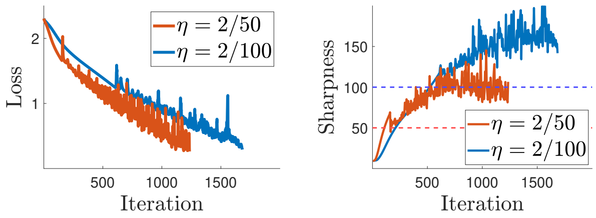


Figure 1: **Example of unstable convergence** for training CIFAR-10 with GD. We follow the experimental setup of (Cohen et al., 2021); see Experiment 1 for details. We use a ReLU network. Here, condition (1.2) fails, but the training loss still (non-monotonically) decreases in the long run.

When the cost is quadratic, condition (1.2) is in fact necessary for stability: if $\eta > \frac{2}{L}$, then GD diverges (see Fact 1). This observation carries over to most convex optimization settings and also neural networks when using the neural tangent kernel approximations (Jacot et al., 2018; Li and Liang, 2018; Lee et al., 2019). Thus, it is reasonable to assume condition (1.2) for those analyses. However, for general nonconvex costs, it is not clear whether the stable regime condition (1.2) is required or even reasonable.

Recently, it has been observed that GD on neural networks often violates condition (1.2). More specifically, Cohen et al. (2021) observe that when we run GD to train a neural network, the condition (1.2) fails, but contrary to the common wisdom from convex optimization, the training loss still (non-monotonically) decreases in the long run. See Figure 1 for an example of this phenomenon. We call this phenomenon **"unstable" convergence**.

Unfortunately, very little is known about unstable convergence. The causes and implications of this phenomenon have not been explored in the literature. More importantly, the main features as well as the scope of this phenomenon have not been discussed. Characterizing the main features is important because it not only furnishes better understanding of this bizarre phenomenon, but also lays a foundation for future theoretical studies; especially, the main characteristics of this phenomenon will help build a more practical theory of the neural network optimization.

Contributions. In light of the above motivation, the main contributions of this paper are as follows:

[1] **Department of EECS, MIT**, Cambridge, MA, USA [2] Part of this work was done while Kwangjun Ahn was visiting the Simons Institute for the Theory of Computing, Berkeley, CA, USA. [3] **IIIS, Tsinghua University**, Beijing, China. Correspondence to: Kwangjun Ahn <kjahn@mit.edu>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

Unstable Convergence

What are its causes (Section 3):

- Lack of (flat) stationary points near GD trajectory
- Forward-invariance ($F(S) \subseteq S$) of the GD dynamics

What are its main features (Section 4):

Object	Quantity	Behavior
Loss (§4.1)	$\mathsf{RP}(\boldsymbol{\theta}^t)$	oscillates near 0
Iterates (§4.2)	$L(\boldsymbol{\theta}^t; \eta\nabla f(\boldsymbol{\theta}^t))$	oscillates near $2/\eta$
Sharpness (§4.4)	$\lambda_{\max}(\nabla^2 f(\boldsymbol{\theta}^t))$	oscillates $\frac{\text{near}}{\text{above}}$ $2/\eta$

- **Relative Progress:** $\mathsf{RP}(\boldsymbol{\theta}) := \frac{f(\boldsymbol{\theta}-\eta\nabla f(\boldsymbol{\theta}))-f(\boldsymbol{\theta})}{\eta\|\nabla f(\boldsymbol{\theta})\|^2}$
- **Directional smoothness:** $L(\boldsymbol{\theta}; \mathbf{v}) := \frac{\langle \mathbf{v}, \nabla f(\boldsymbol{\theta})-\nabla f(\boldsymbol{\theta}-\mathbf{v})\rangle}{\|\mathbf{v}\|^2}$

- Loss behavior ⇔ Iterates behavior (Subsection 4.3)
- Loss behaviour ⇒ Sharpness behavior (Subsection 4.4)

Extenson to SGD (Section 5)

Figure 2: Overview/summary of results.

1. We discuss the main causes driving the unstable convergence phenomenon (Section 3).
2. We identify the main features that characterize unstable convergence in terms of how loss, iterates, and sharpness[1] evolve with GD updates. Moreover, we investigate and clarify the relations between them. Our characterizations demonstrate that the features of unstable convergence are in stark contrast with those of traditional stable convergence, suggesting that their optimization mechanisms are significantly different.
3. In particular, the main features considered in this work provide alternative ways to identify unstable convergence in practice.

Figure 2 provides a more technical overview of our main findings, along with their interpretations.

1.1. Related Work

Under various contexts, several recents works have observed the unstable convergence phenomenon in training neural networks with (S)GD (Wu et al., 2018; Xing et al., 2018; Lewkowycz et al., 2020; Jastrzkebski et al., 2017; 2018). We refer readers to the related work section of Cohen et al. (2021) for greater context.

The unstable convergence phenomenon is first formally identified by Cohen et al. (2021), and in their paper it is named *edge of stability*. More specifically, they observe a more refined version of the unstable convergence: when training a neural network with GD, the sharpness at the iterate goes beyond the threshold $\eta/2$, and often saturates right at (or above) the threshold. In Section 5, we will explore the relations between our main features and their observed phenomenon.

Concurrent works. Recently, Ma et al. (2022) also investigate the causes of unstable convergence based on their empirical observations. Their main observation is that unstable convergence might be due to the landscape of loss function where the loss grows slower than a quadratic near the local minima. As we will see in Subsection 3.2, their main finding is consistent with our explanation. They also demonstrate through examples that such "sub-quadratic" growth near the minima is caused by the heterogeneity of data; see their Section 6 for details.

Another work by Arora et al. (2022) identifies a setting in which one can prove the unstable convergence phenomenon theoretically. More specifically, they show that the normalized gradient descent dynamics of form $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta\nabla f(\boldsymbol{\theta}^t)/\left\|\nabla f(\boldsymbol{\theta}^t)\right\|$ can provably exhibit the unstable convergence phenomenon near the minima under some suitable assumptions; see their Section 4 for details.

2. Unstable GD Can't Reach Stationary Points

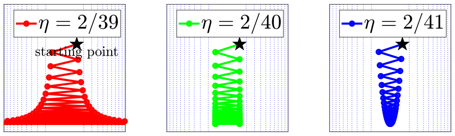
In this section, we build intuitions about what the unstable regime $\eta > 2/L$ suggests. First, note that the fixed points $\boldsymbol{\theta}^\infty$ of the GD dynamics $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta\nabla f(\boldsymbol{\theta}^t)$ are the stationary points, i.e., points such that $\nabla f(\boldsymbol{\theta}^\infty) = 0$. Hence, the GD dynamics will eventually approach one of the stationary points, and in order to understand the unstable regime, we first need to understand the behavior of the dynamics near the stationary points whose sharpness is greater than $2/\eta$.

As a warm up, we first consider the simplest setting of quadratic costs where the sharpness is constant globally. We begin with the following well-known fact.

Fact 1. *On a quadratic cost $f(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^\top P\boldsymbol{\theta} + \mathbf{q}^\top\boldsymbol{\theta} + r$, GD will diverge if any eigenvalue of P exceeds the threshold $2/\eta$. For convex quadratics, this condition is "iff."* □

Below we quickly illustrate this fact through an example.

Example 1. *Consider optimizing a quadratic cost $f(\theta_1, \theta_2) = 20\theta_1^2 + \theta_2^2$. Note that in this case $L = 40$. Let us run GD on this cost with $\eta = 2/39,\ 2/40,\ 2/41$.*



As shown in the above plots, GD converges to the optimum if $\eta < 2/L$ and it diverges if $\eta > 2/L$. □

Due to the above fact, one can build the following intuition:

[1] In this paper, following (Cohen et al., 2021), sharpness means the maximum eigenvalue of the loss Hessian, i.e., $\lambda_{\max}(\nabla^2 f(\boldsymbol{\theta}^t))$.

when a stationary point has sharpness greater than $2/\eta$, the GD dynamics cannot converge to the stationary point.

We first formalize this intuition. In particular, we show that GD cannot converge to a stationary point that has sharpness greater than $2/\eta$. We make the following assumptions about the spectrum of Hessian at stationary points and the non-degeneracy of the GD dynamics.

Assumption 1. *Let $F(\boldsymbol{\theta}) = \boldsymbol{\theta} - \eta\nabla f(\boldsymbol{\theta})$ and assume that for any subset S of measure zero, $F^{-1}(S)$ is of measure zero. Moreover, each stationary point $\boldsymbol{p}$ satisfies $\frac{1}{\eta} \notin \lambda(\nabla^2 f(\boldsymbol{p}))$, where $\lambda(\nabla^2 f(\boldsymbol{p}))$ denotes the set of eigenvalues of the Hessian of f at $\boldsymbol{p}$.*

Theorem 1. *For a given subset $\mathcal{X}$ of the domain of parameter $\boldsymbol{\theta}$, assume that f is C^2 in $\mathcal{X}$. Suppose that for each stationary point $\boldsymbol{p} \in \mathcal{X}$, it holds that either $\lambda_{\min}(\nabla^2 f(\boldsymbol{p})) < 0$ or $\lambda_{\max}(\nabla^2 f(\boldsymbol{p})) > \frac{2}{\eta}$. Then under Assumption 1, there is a measure-zero subset $\mathcal{N}$ s.t. for all initializations $\boldsymbol{\theta}^0 \in \mathcal{X} \setminus \mathcal{N}$, the GD dynamics $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta\nabla f(\boldsymbol{\theta}^t)$ do not converge to any of the stationary points in $\mathcal{X}$.*

Proof of Theorem 1. The proof is inspired by those of (Lee et al., 2016; Panageas and Piliouras, 2017). First, we recall Stable Manifold Theorem (Shub, 2013, Thm III.7).

Lemma 1. *Let $\mathbf{p}$ be a fixed point for the C^r local diffeomorphism $h : U \to \mathbb{R}^n$ where U is an open neighborhood of $\mathbf{p}$ in $\mathbb{R}^n$ and $r \geq 1$. Let $E^{\mathsf{s}} \oplus E^{\mathsf{c}} \oplus E^{\mathsf{u}}$ be the invariant spliiting of $\mathbb{R}^n$ into the generalized eigenspaces of $Dh(\mathbf{p})$ corresponding to eigenvalues of absolute value less than one, equal to one, and greater than one. To the $Dh(\mathbf{p})$-invariant subspace $E^{\mathsf{s}} \oplus E^{\mathsf{c}}$ there is an associated local h-invariant C^r embedded disc $W^{\mathsf{s}\oplus\mathsf{c}}_{\mathsf{loc}}$ of dimension $\dim(E^{\mathsf{s}} \oplus E^{\mathsf{c}})$ and ball B around $\mathbf{p}$ such that $h(W^{\mathsf{s}\oplus\mathsf{c}}_{\mathsf{loc}}) \cap B \subset W^{\mathsf{s}\oplus\mathsf{c}}_{\mathsf{loc}}$. Moreover, if $h^n(\mathbf{x}) \in B$ for all $n \geq 0$, then $\mathbf{x} \in W^{\mathsf{s}\oplus\mathsf{c}}_{\mathsf{loc}}$.*

To apply Lemma 1, we first show that the GD dynamics F is a local diffeomorphism at each stationary point $\boldsymbol{p}$ satisfying $\frac{1}{\eta} \notin \lambda(\nabla^2 f(\boldsymbol{p}))$. This follows from the inverse function theorem: (i) Note that F is a C^1 vector field since f is C^2, (ii) the Jacobian of F is equal to $DF(\boldsymbol{p}) = I - \eta\nabla^2 f(\boldsymbol{p})$, and since $\frac{1}{\eta} \notin \lambda(\nabla^2 f(\boldsymbol{p}))$, the Jacobian is invertible. Hence, by inverse function theorem, we conclude that F is a local diffeomorphism around $\boldsymbol{p}$.

Hence for each stationary point $\boldsymbol{p}$ satisfying $1/\eta \notin \lambda(\nabla^2 f(\boldsymbol{p}))$, we can apply Lemma 1 at $\boldsymbol{p}$. Let $B_{\boldsymbol{p}}$ be the open ball due to Lemma 1. Let $\mathcal{S}$ be the set of stationary points. Consider the following open cover

$$\mathcal{C} := \bigcup_{\substack{\boldsymbol{p}\text{: stationary point} \\ \frac{1}{\eta} \notin \lambda(\nabla^2 f(\boldsymbol{p}))}} B_{\boldsymbol{p}}\,. \tag{2.1}$$

Then from Assumption 1, it follows that $\mathcal{S} \subset \mathcal{C}$ and hence $\mathcal{C}$ is an open cover of $\mathcal{S}$. Thus, Lindelöf's lemma guarantees that there exists a countable subcover of $\mathcal{C}$, i.e., there exist $\boldsymbol{p}_1, \boldsymbol{p}_2, \dots$ s.t. $\mathcal{C} = \cup_{i=1}^{\infty} B_{\boldsymbol{p}_i}$. If GD converges to a stationary point $\boldsymbol{p}$, there must exist t_0 and i such that $F^t(\boldsymbol{p}_0) \in B_{\boldsymbol{p}_i}$ for all $t \geq t_0$. From Lemma 1, we conclude that $F^t(\boldsymbol{\theta}_0) \in W^{\mathsf{s}\oplus\mathsf{c}}_{\mathsf{loc}}(\boldsymbol{p}_i)$. In other words, we have $\boldsymbol{\theta}_0 \in F^{-t}(W^{\mathsf{s}\oplus\mathsf{c}}_{\mathsf{loc}}(\boldsymbol{p}_i))$ for all $t \geq t_0$. Hence the set of initial points in $\mathcal{X}$ for which GD converges to a stationary point is a subset of

$$\mathcal{N} := \bigcup_{i=1}^{\infty}\bigcup_{t=0}^{\infty} F^{-t}(W^{\mathsf{s}\oplus\mathsf{c}}_{\mathsf{loc}}(\boldsymbol{p}_i))\,.$$

Now from the assumption that either $\lambda_{\min}(\nabla^2 f(\boldsymbol{p})) < 0$ or $\lambda_{\max}(\nabla^2 f(\boldsymbol{p})) > \frac{2}{\eta}$, it follows that $I - \eta\nabla^2 f(\boldsymbol{p})$ has an eigenvalue whose absolute value is greater than 1. Hence, for each stationary point $\boldsymbol{p}$, $\dim(E^{\mathsf{u}}) \geq 1$. This implies that each $W^{\mathsf{s}\oplus\mathsf{c}}_{\mathsf{loc}}(\boldsymbol{p})$ has measure zero, and from the assumption that $F^{-1}(\mathcal{X})$ is of measure zero if $\mathcal{X}$ is of measure zero, it holds that each $F^{-t}(W^{\mathsf{s}\oplus\mathsf{c}}_{\mathsf{loc}}(\boldsymbol{p}_i))$ is of measure zero. Thus, being a countable union of measure zero sets, $\mathcal{N}$ is measure zero. It follows that for initialization $\boldsymbol{\theta}_0 \in \mathcal{X} \setminus \mathcal{N}$, the GD dynamics never converge to a stationary point in $\mathcal{X}$. □

Remark 1. Note that Theorem 1 applies to the case when stationary points are not isolated. Moreover, the condition that every stationary point $\boldsymbol{p}$ satisfies $\frac{1}{\eta} \notin \lambda(\nabla^2 f(\boldsymbol{p}))$ can be relaxed to the condition that the open cover $\mathcal{C}$ in (2.1) covers the entire set of stationary points $\mathcal{S}$.

The main takeaway of Theorem 1 is that for almost all initializations, GD cannot converge to the stationary point whose sharpness is larger than $2/\eta$ even when there is only a single eigenvector whose eigenvalue exceeds the threshold $2/\eta$. Having this intuition, we next discuss how "convergence" could happen under the unstable regime.

3. How Can Unstable GD "Converge"?

In the previous section, we saw that when stationary points have large sharpness relative to the step size, GD cannot converge to those stationary points. However, as we saw in Figure 1, GD can still "converge" under the unstable regime; GD still manages to (non-monotonically) decrease the training loss in the long run. In this section, we understand this bizzare co-occurrence. We first discuss some possible causes for the unstable regime.

3.1. What Causes the Unstable Regime

One possible cause for the unstable regime is that the landscape has only "trivial" stationary points; we will formalize the meaning of "trivial" shortly. This situation turns out to be quite common for neural networks as illustrated by the following result.

Proposition 1. *Assume the loss of neural network parametrized $\boldsymbol{\theta}$ contains a weight decay term as follows,*

$$\ell(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i, \boldsymbol{\theta}) + \gamma\|\boldsymbol{\theta}\|_2^2.$$

If we partition the network parameter $\boldsymbol{\theta} = [\boldsymbol{\xi}; \boldsymbol{\zeta}]$ such that a subset of the network parameters $\boldsymbol{\zeta}$ is positive homogeneous, i.e. for any input data x_i and positive number $c > 0$

$$f(\boldsymbol{x}_i, [\boldsymbol{\xi}, \boldsymbol{\zeta}]) = f(x_i, [\boldsymbol{\xi}, c\boldsymbol{\zeta}]),$$

Then the loss $\ell(\boldsymbol{\theta})$ has no stationary point if $\boldsymbol{\zeta} \neq 0$.

Proof of Proposition 1. This statement follows by a simple observation that from positive homogeneity,

$$\langle \nabla_{\boldsymbol{\zeta}} f(\boldsymbol{x}_i, [\boldsymbol{\xi}, \boldsymbol{\zeta}]), \boldsymbol{\zeta}\rangle = 0.$$

Therefore, if $\nabla_{\boldsymbol{\zeta}}\gamma\|\boldsymbol{\theta}\|_2^2 \neq 0$, we have

$$\nabla_{\boldsymbol{\zeta}} f(\boldsymbol{x}_i, [\boldsymbol{\xi}, \boldsymbol{\zeta}]) + \nabla_{\boldsymbol{\zeta}}\gamma\|\boldsymbol{\theta}\|_2^2 \neq 0,$$

which concludes the proof. □

Notice that the positive homogeneity parameters exist in many networks such as ResNet or Transformer when normalization layers exist ($\boldsymbol{a}_{L+1} = \boldsymbol{a}_L/\|\boldsymbol{a}_L\|$, where $\boldsymbol{a}_L$ denote the input to layer L, and $\|\cdot\|$ denotes a norm of choice).

Note that in practical settings, Proposition 1 also suggests that there could be lack of flat minima near the GD trajectory. In the above example of ResNet or Transformer, the networks often add a small ϵ term to the normalization $\boldsymbol{a}_{L+1} = \boldsymbol{a}_L/(\epsilon + \|\boldsymbol{a}_L\|)$ to avoid the loss being undefined at $\boldsymbol{a}_L = 0$. However, the stationary points only exist when $\|\boldsymbol{a}_L\| \approx \epsilon$, in which case the sharpness of the stationary point is very large (on the order of $\sim 1/\epsilon$).

In fact, it has been extensively observed in the literature that the sharpness around GD with practical stepsize choices often goes beyond the threshold $2/\eta$. This claim is verified through a comprehensive set of experiments and called **progressive sharpening** in (Cohen et al., 2021); we refer readers to their Section 3.1 for details. For instance, the sharpness curve in Figure 1 shows this phenomenon. Moreover, a similar phenomenon was observed in (Wu et al., 2018), and they speculated that the density of sharp minima is much larger than the density of flat minima in the neural network landscape. See their Section 4.1 for details.

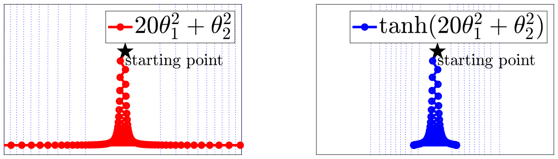
We summarize our discussion regarding the causes of unstable regime as follows.

Takeaway 1. *For practical stepisze choices, lack of (flat) non-trivial stationary points near the GD trajectory can cause GD to enter the unstable regime.*

3.2. Causes for Convergence

As we discussed in Fact 1, for quadratic costs (or more generally for most convex costs), GD being in the unstable regime implies that GD will diverge entirely. However, as demonstrated by (Cohen et al., 2021) through a comprehensive set of experiments, in neural network training, this situation no longer holds. In this section, we discuss how in the unstable regime "convergence" could happen through examples. As a warm-up, let us revisit the quadratic cost considered in Fact 1, but this time with some modifications.

Example 2 ("Flattened" quadratic cost)**.** *For the same quadratic cost as in Fact 1, we chose the same diverging step size $\eta = 2/39 > 2/L$, but this time we change the cost a bit by applying $\tanh(\cdot)$ on top of the quadratic cost. More formally, we consider the cost $\tanh(20\cdot\theta_1^2 + \theta_2^2)$. Due to the fact $\tanh \approx x$ near zero, this transformation wouldn't change the geometry near the global minimum. We run GD on the modified cost, and the result looks as follows (we include the result for the original quadratic cost on the left for comparison):*



As one can see from the above plot, for the transformed cost, GD does not diverge in the unstable regime. □

The above toy example illustrates that indeed for nonconvex costs, being in the unstable regime does not necessarily imply complete divergence. For the above example, this was possible because of $\tanh(\cdot)$, which 'flattens" out the landscape of the quadratic cost away from the minimum.

More formally, let us denote the GD dynamics by $F(\boldsymbol{\theta}) := \boldsymbol{\theta} - \eta\nabla f(\boldsymbol{\theta})$. Then the role of $\tanh(\cdot)$ in the above example is that it creates a compact subset near the minimum that is *forward-invariant*: we say S is forward-invariant with respect to the dynamics F if $F(S) \subseteq S$. Because the gradient of $\tan(\text{quadratic})$ vanishes as the point gets farther away from the minimum, there exists a forward-invariant compact subset $\mathcal{X}$ near the minimum.

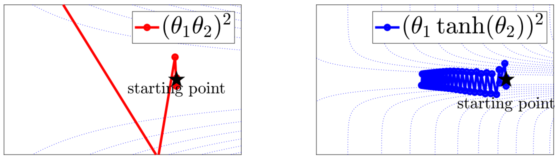
Remark 2. In a very recent concurrent work by (Ma et al., 2022), this phenomenon is discussed in a more principled manner using the *subquadratic growth property*. More specifically, they observed that for practical neural network settings, the loss landscape near the minima exhibits growth that is slower that quadratic, in which case the GD dynamics do not diverge entirely even in the unstable regime. See their Section 4 for details.

We demonstrate this point for neural network examples. We first consider the simplest neural network example, namely a single hidden neuron network.

Example 3 (Single neuron networks)**.** *We consider a trivial task of fitting the data* $(1, 0)$ *with a single hidden neuron neural network. Formally, we consider two types of networks:*

- *linear network:* $f(\theta_1, \theta_2) = (\theta_1 \cdot (1 \cdot \theta_2) - 0)^2$.
- $\tanh$ *network:* $f(\theta_1, \theta_2) = (\theta_1 \cdot \tanh(1 \cdot \theta_2) - 0)^2$.

We initialize both networks at $\boldsymbol{\theta}^0 = (13, 0.01)$ *choose step size* $\eta = 2/150$ *to train them.*

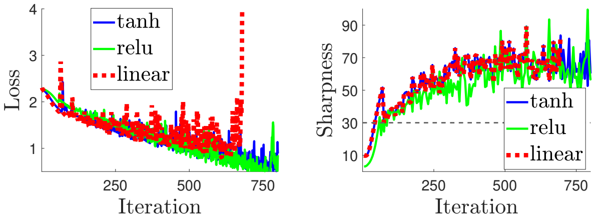


As one can see from the above plots, for a linear network, the iterate quickly diverges, while for the $\tanh$ *network, the iterate does not diverge and converges to a minimum (whose sharpness is indeed approximately equal to* $2/\eta$*).* □

Example 3 illustrates that the use of activation function like $\tanh$ can create a compact forward-invariant subset near the minima, which helps GD not diverge in the unstable regime. In fact, the above example suggests that GD indeed exhibits some convergence behaviour where while being in the unstable regime, GD travels along the valley of minima until it finds a flat enough minimum where it can stabilize.

We now consider more practical neural network examples inspired by the settings considered in (Cohen et al., 2021).

Experiment 1 (CIFAR-10 experiment)**.** *For this example, we follow the setting of the main experiment (Cohen et al., 2021) in their Section 3. Specifically, we use (full-batch) GD to train a neural network on* $5,000$ *examples from CIFAR-10 with the CrossEntropy loss, and the network is a fully-connected architecture with two hidden layers of width* 200. *Under this common setting, we consider three types of networks: (i) linear network without activations; (ii)* $\tanh$ *activations; (ii) ReLU activations. We choose the step size* $\eta = 2/30$ *and the results are as follows:*



As one can see from the above plot, GD converges for the networks with activation functions, while GD diverges without activation functions. □

We summarize our discussion regarding the causes for convergence as follows.

Takeaway 2. *Ingredients of neural networks such as activation functions create a compact forward-invariant set near the minima, which helps GD (non-monotonically) converge in the unstable regime.*

In this section, we have discussed the causes of unstable convergence and explain how the intuitions differ from those of conventional convex optimization. We next move on to study the main characteristics of unstable convergence. For instance, we observe that under the unstable convergence phenomenon, the loss is very non-monotonic. Can we understand the behavior of loss in a more principled way?

4. Characteristics of the Unstable Convergence

In this section, we aim to quantify unstable convergence through several quantities that can be computed during the training. In particular, we will characterize the unstable convergence in terms of the loss behavior and the iterate behavior. We will later demonstrate that the two different behaviors are interconnected with each other.

4.1. Characteristics in Loss Behavior

We first investigate what happens to the loss under unstable convergence. As a warm-up, we first consider the loss behavior under **stable** convergence.

4.1.1. Warm-up: The Stable Regime

Recall from the descent lemma (1.1) that when GD is in the stable regime, then we have $f(\boldsymbol{\theta}^{t+1}) - f(\boldsymbol{\theta}^t) \leq -c\eta \left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2$ for some constant $c > 0$. Putting it differently, we have

$$\frac{f(\boldsymbol{\theta}^{t+1}) - f(\boldsymbol{\theta}^t)}{\eta \left\|\nabla f(\boldsymbol{\theta}^t)\right\|^2} \leq -\text{const.}$$

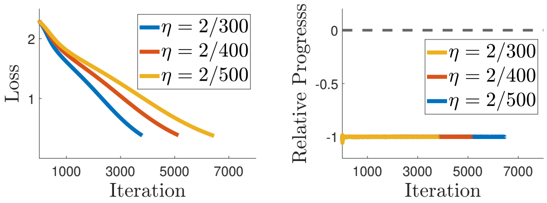
Let us give the ratio on the LHS a name for convenience:

Definition 1 (Relative progress ratio)**.** *We define*

$$\mathsf{RP}(\boldsymbol{\theta}) := \frac{f(\boldsymbol{\theta} - \eta\nabla f(\boldsymbol{\theta})) - f(\boldsymbol{\theta})}{\eta \left\|\nabla f(\boldsymbol{\theta})\right\|^2}.$$

Let us revisit Experiment 1 and verify that for smaller step sizes the relative progress ratio is indeed a negative number.

Experiment 2 (CIFAR-10; stable regime)**.** *We use the same setting as Experiment 1, which follows the setting of the main experiment in (Cohen et al., 2021). For activations, we choose* $\tanh$ *following (Cohen et al., 2021). We choose much smaller step sizes so that GD is in the stable regime. We plot the loss and the relative progress ratio until the training accuracy hits* 95%.



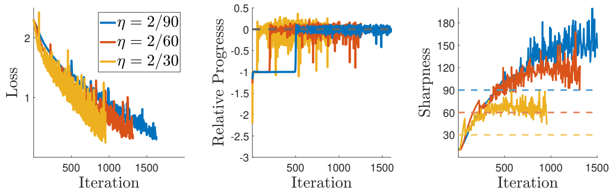
From the above plots, one can see that the relative progress ratio stays negative for all iterations. Moreover, there is no non-monotonic behavior in the loss curve. □

Remark 3. Given the result above, one might wonder why the relative progress saturates around -1. Although we do not have a clear explanation, we suspect that this happens because the trajectory of GD quickly converges to a single direction. We will quickly revisit this later this section. See Remark 6.

4.1.2. Relative Progress Ratio under Unstable Convergence

Given that relative progress is strictly negative number in the stable regime, we now investigate how relative progress ratio behaves in the case of unstable convergence.

Experiment 3 (CIFAR-10; unstable regime)**.** *We use the same setting as Experiment 2. This time we choose step sizes larger so that GD operates in the unstable convergence regime. We plot the loss and the relative progress ratio until the training accuracy hits* 95%.



The above experiment shows that in the the unstable regime, the relative progress ratio saturates around 0 *unlike the stable regime.* □

Remark 4. One curious aspect of the above results is that the optimization seems to get faster as we choose larger step sizes. This is in fact one of the main observations in (Cohen et al., 2021), suggesting that the unstable convergence is preferred in practice for its faster optimization. However, that does not mean one can increase the step size too large. For example, in the above experiment, we observe that the training loss diverges for step size $\eta = 2/10$.

Based on Experiment 3, we raise the following question:

Q. why does $\mathsf{RP}(\boldsymbol{\theta}^t)$ oscillate around 0 under unstable convergence?

We begin with explaining why $\mathsf{RP}(\boldsymbol{\theta}^t)$ cannot stay above 0. Since the loss is converging in a long term, it cannot be that $\mathsf{RP}(\boldsymbol{\theta}^t) > 0$ for many iterations; otherwise, the loss will keep increasing, contradicting the convergence.

More curious part is the fact that $\mathsf{RP}(\boldsymbol{\theta}^t)$ cannot stay below zero, which directly contrasts with the stable regime. To understand this phenomenon, we begin with some intuition.

We have seen that when GD encounters sharp minima, it oscillates near the minima because it cannot stabilize to the minima (due to Theorem 1). In other words, the loss change $f(\boldsymbol{\theta}^{t+1}) - f(\boldsymbol{\theta}^t)$ would be much smaller compared to $\left\|\eta^2 \nabla f(\boldsymbol{\theta}^t)\right\|^2$ the square of the distance that GD travels. Hence, intuitively, one might expect that relative progress cannot be too negative under the unstable convergence regime. We would like to formalize this intuition next.

4.2. Characteristics in Iterates Movement

To that end, let us formally define what it means for GD to oscillate. More generally, consider the situation where $\boldsymbol{\theta}$ is updated by moving along the vector $-\mathbf{v}$. Then this update is oscillatory if the directional derivative at the updated parameter $\boldsymbol{\theta} - \mathbf{v}$ is nearly negative of that at $\boldsymbol{\theta}$, i.e.,

$$\langle \mathbf{v},\, \nabla f(\boldsymbol{\theta} - \mathbf{v}) \rangle \approx -\langle \mathbf{v},\, \nabla f(\boldsymbol{\theta}) \rangle \;.$$

Inspired by this, we consider the following definition.

Definition 2 (Directional smoothness)**.** *For an update vector* $\mathbf{v}$*, we define*

$$L(\boldsymbol{\theta};\mathbf{v}) := \frac{1}{\|\mathbf{v}\|^2} \left\langle \mathbf{v},\, \nabla f(\boldsymbol{\theta}) - \nabla f\big(\boldsymbol{\theta} - \mathbf{v}\big) \right\rangle .$$

Now coming back to the gradient descent where the update vector is $\mathbf{v} = \eta \nabla f(\boldsymbol{\theta})$, we have

$$L(\boldsymbol{\theta};\eta\nabla f(\boldsymbol{\theta})) = \frac{\left\langle \nabla f(\boldsymbol{\theta}),\, \nabla f(\boldsymbol{\theta}) - \nabla f\big(\boldsymbol{\theta} - \eta \nabla f(\boldsymbol{\theta})\big) \right\rangle}{\eta \left\|\nabla f(\boldsymbol{\theta})\right\|^2} .$$

When GD is exhibiting oscillatory behaviour, we would have

$$\langle \nabla f(\boldsymbol{\theta}),\, \nabla f(\boldsymbol{\theta} - \mathbf{v}) \rangle \approx -\langle \nabla f(\boldsymbol{\theta}),\, \nabla f(\boldsymbol{\theta}) \rangle \;,$$

in which case, it holds that

$$L(\boldsymbol{\theta};\eta\nabla f(\boldsymbol{\theta})) \approx \frac{2}{\eta} \quad \text{(when GD iterates oscillate).} \tag{4.1}$$

For intuition, let us quickly verify (4.1) for quadratic costs.

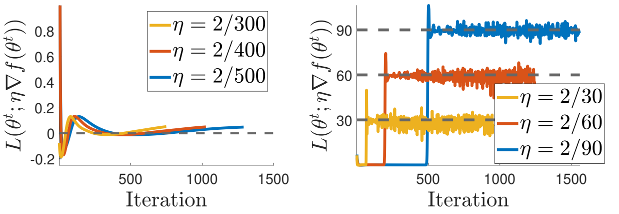
Example 4 (Quadratics)**.** *Consider a quadratic loss function* $f(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top P \boldsymbol{\theta}$ *with* $P \succeq 0$*. Then, the GD update reads* $\boldsymbol{\theta}^{t+1} = (I - \eta P)\boldsymbol{\theta}^t$*. For an eigenvector/eigenvalue pair* $(\mathbf{q}, \lambda)$ *of* P*, the quantity* $\langle \mathbf{q}_{\max},\, \boldsymbol{\theta}^t \rangle$ *evolves as*

$$\begin{aligned}\left\langle \mathbf{q}_{\max},\, \boldsymbol{\theta}^t \right\rangle &= \mathbf{q}^\top (I - \eta P) \boldsymbol{\theta}^{t-1} = (1 - \eta\lambda) \left\langle \mathbf{q}_{\max},\, \boldsymbol{\theta}^{t-1} \right\rangle \\ &= (1 - \eta\lambda)^t \left\langle \mathbf{q}_{\max},\, \boldsymbol{\theta}^0 \right\rangle .\end{aligned}$$

This implies that if $\lambda < 2/\eta$, then $\mathbf{q}^\top \boldsymbol{\theta}^t \to 0$. Hence, if $\eta = 2/\lambda_{\max}(P)$, then after sufficiently large iterations t, we have $\boldsymbol{\theta}^t \approx (-1)^t \left\langle \mathbf{q}_{\max}, \boldsymbol{\theta}^0 \right\rangle \mathbf{q}_{\max}$, in which case $L(\boldsymbol{\theta}; \eta\nabla f(\boldsymbol{\theta})) \approx \frac{2}{\eta}$. □

Given the above view on "oscillating" iterates, we now measure directional smoothness under unstable convergence.

Experiment 4 (Directional smoothness in stable and unstable regimes)**.** *Under the same setting as Experiments 2 and 3, we measure the value $L(\boldsymbol{\theta}^t; \eta\nabla f(\boldsymbol{\theta}^t))$ at each iteration.*



Indeed, one can see that for the unstable regime $L(\boldsymbol{\theta}^t; \eta\nabla f(\boldsymbol{\theta}^t))$ saturates around $2/\eta$, indicating that GD is exhibiting an oscillating behavior. □

Experiment 4 verifies that GD is indeed showing an oscillating behavior under unstable convergence. We remark that a similar conclusion is made in (Xing et al., 2018) as well as the recent concurrent works (Ma et al., 2022; Arora et al., 2022). Now coming back to our original question: *can we show a formal relation between the directional smoothness and the relative progress ratio?*

4.3. Relation between Relative Progress Ratio and Directionl Smoothness

Theorem 2 formalizes our intuition that under the oscillating behavior of GD, $\mathsf{RP}(\boldsymbol{\theta}^t)$ cannot stay below zero.

Theorem 2. *The following identity holds:*

$$\mathsf{RP}(\boldsymbol{\theta}) = -1 + \frac{\eta}{2} \cdot 2\int_0^1 \tau \cdot L(\boldsymbol{\theta}; \eta\tau\nabla f(\boldsymbol{\theta}))\, \mathrm{d}\tau\,. \quad (4.2)$$

Proof. See Appendix A. □

Theorem 2 implies that if the weighted average of $L(\boldsymbol{\theta}; \eta\tau\nabla f(\boldsymbol{\theta}))$ is close to $2/\eta$, namely

$$2\int_0^1 \tau \cdot L(\boldsymbol{\theta}; \eta\tau\nabla f(\boldsymbol{\theta}))\, \mathrm{d}\tau \approx \frac{2}{\eta}\,,$$

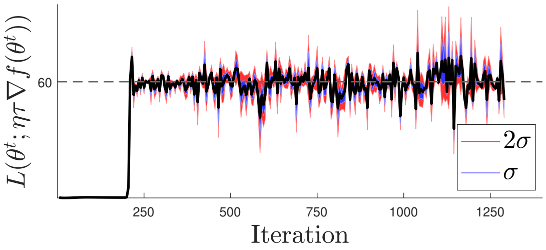
then $\mathsf{RP}(\boldsymbol{\theta})$ is indeed approximately equal to zero. This formally justifies that when GD shows an oscillating behavior, $\mathsf{RP}(\boldsymbol{\theta}^t)$ cannot stay below zero.

In our last experiment of this subsection, we verify that the above weighted average is approximately equal to the single value $L(\boldsymbol{\theta}; \eta\nabla f(\boldsymbol{\theta}))$, building a stronger relation between the directional smoothness and the relative progress ratio.

Experiment 5. *In the same setting as Experiment 3, we choose step size $\eta = 2/60$ and in every 5 iterations, we compute the following values:*

$$L(\boldsymbol{\theta}^t; \eta\tau\nabla f(\boldsymbol{\theta}^t)) \quad \textit{for } \tau \in \{0.01, 0.02, \dots, 1\}.$$

In the plot below, we report the mean of $L(\boldsymbol{\theta}^t; \eta\tau\nabla f(\boldsymbol{\theta}^t))$ among $\tau \in \{0.01, 0.02, \dots, 1\}$ together with the shades which indicate the standard deviations.



This experiment verifies that $L(\boldsymbol{\theta}^t; \eta\tau\nabla f(\boldsymbol{\theta}^t))$ does vary too much across $\tau \in [0, 1]$. Hence, the single value $L(\boldsymbol{\theta}; \eta\nabla f(\boldsymbol{\theta}))$ well represents the weighted average in Theorem 2. □

Hence, Experiment 5 justifies the relation

$$\boxed{\mathsf{RP}(\boldsymbol{\theta}) \approx -1 + \frac{\eta}{2} \cdot L(\boldsymbol{\theta}; \eta\nabla f(\boldsymbol{\theta}))\,,} \quad (4.3)$$

which precisely explains how the oscillatory behavior of GD results in a small relative progress ratio.

Remark 5. Interestingly, the validity of equation (4.3) and Experiment 5 suggests that even though the gradient Lipschitzness is not a good assumption for neural networks, some form of Hessian Lipschitzness is valid along the GD trajectory.

We summarize the finding in this section as follows.

Takeaway 3. *Under the unstable convergence regime, $\mathsf{RP}(\boldsymbol{\theta}^t)$ oscillates near 0 for the following two reasons:*

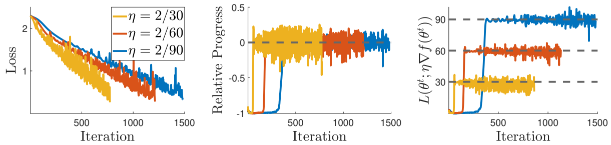
- *$\mathsf{RP}(\boldsymbol{\theta}^t)$ can't stay above 0 because otherwise the loss would not decrease in the long run.*
- *$\mathsf{RP}(\boldsymbol{\theta}^t)$ can't stay below 0 due to the oscillating behavior of GD iterates. This is formalized via (4.3).*

Remark 6. Given (4.3), one can have a better explanation for Remark 3 regarding why $\mathsf{RP}(\boldsymbol{\theta}^t)$ saturates around -1. In the second result of Experiment 4, the directional smoothness remains very small in the stable regime. Based on (4.3), this implies that $\mathsf{RP}(\boldsymbol{\theta}^t)$ is close to -1, which was indeed the case in Experiment 2.

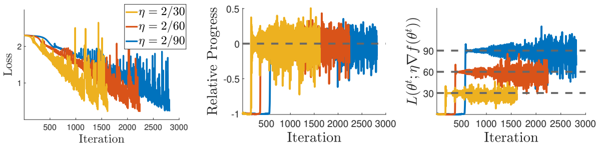
4.3.1. Additional Experiments

In this subsection, we verify the relation (4.3) for other experimental settings.

Experiment 6 (CIFAR-10; ReLU networks). *Under the same setting as Experiment 2 (the setting of the main experiments in (Cohen et al., 2021)), this time we choose ReLU as activation functions.*



In the next set of experiments, we put 2 more hidden layers of width 200 (total 4 hidden layers of width 200).



The results are largely similar to those for tanh *activations, and the relation* (4.3) *holds for all cases.* □

4.4. Implications for Sharpness

In the previous subsection, we saw that one characteristics of unstable convergence is that $\mathsf{RP}(\boldsymbol{\theta}^t)$ saturates around zero. In this section, we investigate implications of this characteristics for sharpness. In particular, we discuss some relations to a curious phenomenon called *edge of stability* (EoS) recently observed in (Cohen et al., 2021). The gist of their observation is that for GD on neural networks often satisfies the following properties:

A. $\lambda_{\max}(\nabla^2 f(\boldsymbol{\theta}^t)) > 2/\eta$ for most iterates.
B. In fact, in many cases $\lambda_{\max}(\nabla^2 f(\boldsymbol{\theta}^t))$ saturates right at (or slightly above) $2/\eta$.

To that end, we begin with the following consequence of Theorem 2.

Corollary 1. *Let L_t be the maximum sharpness along the line segment between the iterates $\boldsymbol{\theta}^t$ and $\boldsymbol{\theta}^{t+1}$, i.e., $L_t := \sup_{\boldsymbol{\theta}\in\overline{\boldsymbol{\theta}^t\boldsymbol{\theta}^{t+1}}}\{\lambda_{\max}(\nabla^2 f(\boldsymbol{\theta}))\}$. Then, the following inequality holds:*

$$\frac{2}{\eta}\cdot(\mathsf{RP}(\boldsymbol{\theta})+1) \le L_t\,.$$

Proof. It follows from the fact that for each $\tau \in [0,1]$ $L(\boldsymbol{\theta}; \eta\tau\nabla f(\boldsymbol{\theta})) \le \sup\{\lambda_{\max}(\nabla^2 f(\boldsymbol{\theta})) \;:\; \boldsymbol{\theta}$ lies on the line segment between $\boldsymbol{\theta}^t$ and $\boldsymbol{\theta}^t - \eta\tau\nabla f(\boldsymbol{\theta})\}$. Clearly, the right hand side is upper bounded by L_t. □

Corollary 1 implies that when $\mathsf{RP}(\boldsymbol{\theta}^t)$ oscillates around zero, then L_t has to be frequently above the threshold $2/\eta$. One can actually refine this statement to understand the part A of EoS, given our results so far. In light of Experiment 5, if $L(\boldsymbol{\theta}^t; \eta\tau\nabla f(\boldsymbol{\theta}^t))$ does vary much across $\tau \in [0,1]$, then one can actually write

$$\begin{aligned}\frac{2}{\eta}(\mathsf{RP}(\boldsymbol{\theta}^t)+1) &\approx \lim_{\tau\to 0} L(\boldsymbol{\theta}^t; \eta\tau\nabla f(\boldsymbol{\theta}^t))\\ &\overset{\clubsuit}{\le} \lambda_{\max}(\nabla^2 f(\boldsymbol{\theta}^t))\,,\end{aligned}$$

which is the part A of EoS.

Moreover, let us for a moment additionally assume that $\nabla f(\boldsymbol{\theta}^t)$ is approximately parallel to the largest eigenvector of the Hessian $\nabla^2 f(\boldsymbol{\theta}^t)$. This might look stringent at first glance, but given the calculations in Example 4, this assumption is true for unstable GD on a quadratic cost. Also, recently, this behavior is theoretically proven for the normalized gradient descent dynamics (Arora et al., 2022). Under this assumption, one can further deduce that the inequality (♣) holds with approximate equality, and the part B of EoS would hold in that case.

5. Relative Progress for SGD

In this section, we extend our discussion to the stochastic gradient descent (SGD):

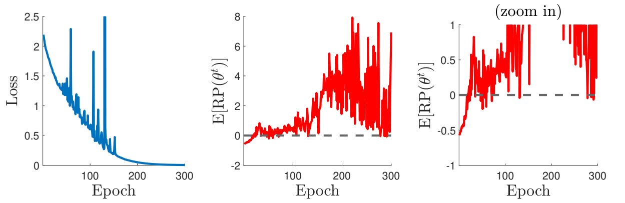
$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta g(\boldsymbol{\theta}^t), \;\; \text{where } \mathbb{E}[g(\boldsymbol{\theta}^t)] = \nabla f(\boldsymbol{\theta}^t)\,.$$

For the case of SGD, there is one obvious challenge. With SGD, the training loss does not decrease monotonically since SGD is a random algorithm. Hence, it is not clear how to precisely define what it means for SGD to be in the unstable regime. On the other hand, inspired by our discussion in Subsection 4.1, a more transparent way to define the unstable regime for SGD is via the relative progress ratio. In particular, we consider the following extension.

Definition 3 (Expected relative progress ratio)**.**

$$\mathbb{E}[\mathsf{RP}(\boldsymbol{\theta})] := \frac{\mathbb{E}f(\boldsymbol{\theta}-\eta g(\boldsymbol{\theta})) - f(\boldsymbol{\theta})}{\eta\left\|\nabla f(\boldsymbol{\theta})\right\|^2}\,.$$

Experiment 7 (CIFAR-10; SGD on ReLU networks). *Under the same setting as Experiment 6, this time we train the network with SGD with minibatch size of* 32 *and step size* $\eta = 2/100$. *We compute the full-batch loss and the expected relative progress ratio at the end of each epoch.*



Note that $\mathbb{E}[\mathsf{RP}(\boldsymbol{\theta}^t)]$ does not stay below zero. Based on our discussion in Subsection 4.1, this suggests that SGD is in the unstable regime. □

Remark 7 (Expected loss change is not negative?!). One very surprising aspect of the above results is that $\mathbb{E}[\mathsf{RP}(\boldsymbol{\theta}^t)]$ is not negative for a majority of iterations. This is rather counter-intuitive given that in the loss plot SGD decreases the loss in the long run. On the other hand, we note that this counter-intuitive phenomenon is also observed by (Cohen et al., 2021) in a comprehensive set of experiments. In particular, they mention "*what may be more surprising is that SGD is not even decreasing the training loss in expectation.*" See (Cohen et al., 2021, Appendix H) and (Cohen et al., 2021, Figures 25 and 26) for details.

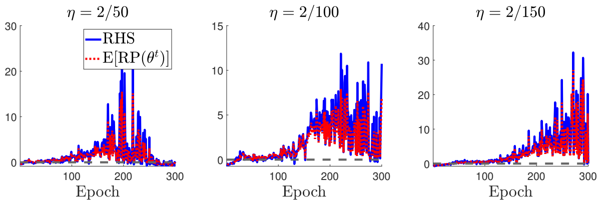
We now establish a relation analogous to (4.3) for SGD. Similarly to Theorem 2, one can prove the following:

$$\mathbb{E}[\mathsf{RP}(\boldsymbol{\theta})] = -1 + \frac{\eta}{2}\cdot 2\int_0^1 \tau\mathbb{E}_g\left[\frac{\|g(\boldsymbol{\theta})\|^2\cdot L(\boldsymbol{\theta};\eta\tau g(\boldsymbol{\theta}))}{\|\nabla f(\boldsymbol{\theta})\|^2}\right]\mathrm{d}\tau\,.$$

See Appendix A for derivation. Hence, the analogous relation to (4.3) is:

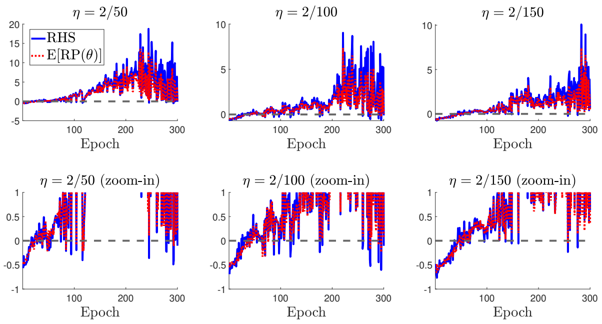
$$\mathbb{E}[\mathsf{RP}(\boldsymbol{\theta})] \approx -1 + \frac{\eta}{2}\cdot\mathbb{E}\left[\frac{\|g(\boldsymbol{\theta})\|^2}{\|\nabla f(\boldsymbol{\theta})\|^2}L(\boldsymbol{\theta};\eta g(\boldsymbol{\theta}))\right]. \quad (5.1)$$

Experiment 8 (CIFAR-10; verifying (5.1) for ReLU)**.** *Under the same setting as Experiment 7, we compute* $\mathbb{E}[\mathsf{RP}(\boldsymbol{\theta})]$ *and the RHS of* (5.1) *at the end of each epoch and compare those values. We choose step sizes* $\eta = 2/50,\ 2/100,\ 2/150.$



Note that the LHS and RHS of (5.1) *are very similar in all results, verifying the relation* (5.1). □

Experiment 9 (CIFAR-10; verifying (5.1) for tanh)**.** *We repeat Experiment 8 with* tanh *activations.*



The results are similar to those for ReLU networks. □

6. Discussion and Future Directions

This work demonstrated characteristics of unstable convergence that are in stark contrast with those of stable convergence. Consequently, this work leads to several interesting future directions.

- *Better optimizer?* The characteristics based on the directional smoothness suggests that the adjacent iterates have nearly opposite gradients in the unstable regime. This obviates the efficacy of many efficient methods (e.g. variance reduced methods) which are designed based on the intuition that the adjacent iterates have similar gradients. Hence, it would be interesting to design an efficient optimizer that respects the new characteristics.

- *Faster training under unstable convergence?* As discussed in Remark 4, another striking feature of unstable convergence is that the training loss seems to converge faster. One interesting question is whether one can elucidate this faster optimization by further exploring our characterization of relative progress.

- *What assumptions are valid for neural network optimization?* It is clear that unstable convergence cannot be reasoned with the widespread condition of $\eta < \frac{2}{L}$. Then what assumptions would be valid for neural networks? As discussed in Remark 5, our Experiment 5 suggests that although the gradient Lipschitzness is not a good assumption for neural networks, some form of Hessian Lipschitzness might be a valid one.

- *Unstable regime for adaptive methods?* Our characterizations are limited to constant step size (S)GD, and it is not clear how these characterizations carry over to adaptive methods such as Adam and RMSProp. Investigating adaptive methods will help us understand how they differ from (S)GD for neural network optmization.

Acknowledgements

Kwangjun Ahn thanks **Pourya Habib Zadeh** for various discussions and his help with experiments during the initial stage of this work. Kwangjun Ahn also thanks **Stephen J. Wright** and **Sinho Chewi** for helpful discussions while visiting the Simons Institute for the Theory of Computing. Kwangjun Ahn also thanks **Daniel Soudry** for fruitful discussions regarding Theorem 1.

Kwangjun Ahn and Suvrit Sra acknowledge support from an NSF CAREER grant (1846088), and NSF CCF-2112665 (TILOS AI Research Institute). Kwangjun Ahn also acknowledges support from Kwanjeong Educational Foundation. Jingzhao Zhang acknowledges support from IIIS young scholar fellowship.

References

S. Arora, Z. Li, and A. Panigrahi. Understanding gradient descent on edge of stability in deep learning. *To appear at International Conference on Machine Learning (ICML) 2022 arXiv:2205.09745*, 2022.

J. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd Advances in Neural Information Processing Systems*, pages 8580–8589, 2018.

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

S. Jastrzkebski, Z. Kenton, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. On the relation between the sharpest directions of dnn loss and the sgd step length. *arXiv preprint arXiv:1807.05031*, 2018.

J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in Neural Information Processing Systems*, 32: 8572–8583, 2019.

J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257. PMLR, 2016.

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

Y. Li and Y. Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8168–8177, 2018.

C. Ma, D. Kunin, L. Wu, and L. Ying. The multiscale structure of neural network loss functions: The effect on optimization and origin. *arXiv preprint arXiv:2204.11326*, 2022.

I. Panageas and G. Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

M. Shub. *Global stability of dynamical systems*. Springer Science & Business Media, 2013.

L. Wu, C. Ma, et al. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31:8279–8288, 2018.

C. Xing, D. Arpit, C. Tsirigotis, and Y. Bengio. A walk with SGD. *arXiv preprint arXiv:1802.08770*, 2018.

A. Proof of Theorem 2

Using the fact

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left[f(\boldsymbol{\theta}-\eta\tau\nabla f(\boldsymbol{\theta})\right]=\left\langle -\eta\nabla f(\boldsymbol{\theta}),\,\nabla f\big(\boldsymbol{\theta}-\eta\tau\nabla f(\boldsymbol{\theta})\big)\right\rangle\,,$$

we obtain

$$\begin{aligned}f\big(\boldsymbol{\theta}-\eta\nabla f(\boldsymbol{\theta})\big)-f(\boldsymbol{\theta})&=-\eta\int_0^1\left\langle\nabla f(\boldsymbol{\theta}),\,\nabla f\big(\boldsymbol{\theta}-\eta\tau\nabla f(\boldsymbol{\theta})\big)\right\rangle\mathrm{d}\tau\\&=-\eta\left\|\nabla f(\boldsymbol{\theta})\right\|^2-\eta\int_0^1\left\langle\nabla f(\boldsymbol{\theta}),\,\nabla f\big(\boldsymbol{\theta}-\eta\tau\nabla f(\boldsymbol{\theta})\big)-\nabla f(\boldsymbol{\theta})\right\rangle\mathrm{d}\tau\,.\end{aligned}$$

Hence, after rearranging, we get

$$\frac{\eta}{2}\cdot 2\int_0^1\tau\cdot L(\boldsymbol{\theta};\eta\tau\nabla f(\boldsymbol{\theta}))\,\mathrm{d}\tau-1=\frac{f\big(\boldsymbol{\theta}-\eta\nabla f(\boldsymbol{\theta})\big)-f(\boldsymbol{\theta})}{\eta\cdot\left\|\nabla f(\boldsymbol{\theta})\right\|^2}=\mathsf{RP}(\boldsymbol{\theta})\,,$$

which is precisely the relation in Theorem 2.

Derivation for the SGD case. For the SGD case, the derivation is similar.

$$\begin{aligned}&f\big(\boldsymbol{\theta}-\eta g(\boldsymbol{\theta})\big)-f(\boldsymbol{\theta})\\&\overset{(a)}{=}-\eta\int_0^1\left\langle g(\boldsymbol{\theta}),\,\nabla f\big(\boldsymbol{\theta}-\eta\tau g(\boldsymbol{\theta})\big)\right\rangle\mathrm{d}\tau\\&=-\eta\left\langle g(\boldsymbol{\theta}),\,\nabla f(\boldsymbol{\theta})\right\rangle-\eta\int_0^1\left\langle g(\boldsymbol{\theta}),\,\nabla f\big(\boldsymbol{\theta}-\eta\tau g(\boldsymbol{\theta})\big)-\nabla f(\boldsymbol{\theta})\right\rangle\mathrm{d}\tau\end{aligned}$$

where (a) is due to the fact

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left[f(\boldsymbol{\theta}-\eta\tau g(\boldsymbol{\theta})\right]=\left\langle -\eta g(\boldsymbol{\theta}),\,\nabla f\big(\boldsymbol{\theta}-\eta\tau g(\boldsymbol{\theta})\big)\right\rangle\,.$$

After taking expectation over the randomness in the stochastic gradient, we obtain

$$\mathbb{E}f\big(\boldsymbol{\theta}-\eta g(\boldsymbol{\theta})\big)-f(\boldsymbol{\theta})=-\eta\left\|\nabla f(\boldsymbol{\theta})\right\|^2-\eta\int_0^1\mathbb{E}\left\langle g(\boldsymbol{\theta}),\,\nabla f\big(\boldsymbol{\theta}-\eta\tau g(\boldsymbol{\theta})\big)-\nabla f(\boldsymbol{\theta})\right\rangle\mathrm{d}\tau$$

Hence, after rearranging we obtain the desired equation:

$$\frac{\mathbb{E}f\big(\boldsymbol{\theta}-\eta g(\boldsymbol{\theta})\big)-f(\boldsymbol{\theta})}{\eta\left\|\nabla f(\boldsymbol{\theta})\right\|^2}=-1+\frac{\eta}{2}\cdot 2\int_0^1\tau\cdot\mathbb{E}_g\left[\frac{\left\|g(\boldsymbol{\theta})\right\|^2\cdot L(\boldsymbol{\theta};\eta\tau g(\boldsymbol{\theta}))}{\left\|\nabla f(\boldsymbol{\theta})\right\|^2}\right]\,\mathrm{d}\tau\,.$$

This completes the derivation.