Journal of Machine Learning Research 26 (2025) 1-68 Submitted 12/23; Revised 11/25; Published 11/25

Good regularity creates large learning rate implicit biases: edge of stability, balancing, and catapult

Yuqing Wang YWAN1050@JH.EDU
Georgia Institute of Technology & Johns Hopkins University

Zhenghao Xu ZHENGHAOXU@GATECH.EDU
Georgia Institute of Technology

Tuo Zhao TOURZHAO@GATECH.EDU
Georgia Institute of Technology

Molei Tao MTAO@GATECH.EDU
Georgia Institute of Technology

Editor: Mladen Kolar

Abstract

Large learning rates, when applied to gradient descent for nonconvex optimization, yield various implicit biases including the edge of stability (Cohen et al., 2021), balancing (Wang et al., 2022a), and catapult (Lewkowycz et al., 2020). These phenomena cannot be well explained by classical optimization theory. Though significant theoretical progress has been made in understanding these implicit biases, it remains unclear for which objective functions they are more likely to occur — more precisely, for which functions there exists a larger set of initial conditions that lead to these phenomena? This paper provides an initial step in answering this question and also shows that these implicit biases are in fact various tips of the same iceberg. To establish these results, we develop a global convergence theory under large learning rates, for a family of nonconvex functions without globally Lipschitz continuous gradient, which was typically assumed in existing convergence analysis. Specifically, these phenomena are more likely to occur when the optimization objective function has good regularity. This regularity, together with gradient descent using a large learning rate that favors flatter regions, results in these nontrivial dynamical behaviors. Another corollary is the first non-asymptotic convergence rate bound for large-learning-rate gradient descent optimization of nonconvex functions. Although our theory only applies to specific functions so far, the possibility of extrapolating it to neural networks is also experimentally validated, for which different choices of loss, activation functions, and other techniques such as batch normalization can all affect regularity significantly and lead to very different training dynamics.

Keywords: large learning rate, implicit bias, convergence of gradient descent, nonconvex optimization, non-globally-Lipschitz gradient

1. Introduction

Large learning rates are often employed in deep learning practices, which is believed to improve training efficiency and generalization (Seong et al., 2018; Lewkowycz et al., 2020; Yue et al., 2023; Smith, 2018). However, a comprehensive theoretical foundation supporting such a choice has been largely absent, as a large body of the theory literature primarily fo-

cused on small or infinitesimal learning rates (gradient flow), leaving the empirical successes achieved with large learning rates unexplained. Only recently have researchers redirected their efforts toward investigating the benefits of larger learning rates.

A prevalent conjecture posits that large learning rates result in a preference toward 'flat minima' which in turn leads to better generalization. This notion has served as a catalyst for many novel insights examining the 'sharpness' of the solution under large learning rates, i.e., the largest eigenvalue of the Hessian matrix associated with the objective function. A notable example can be found in Cohen et al. (2021), where the phenomenon of the *Edge of Stability* (EoS) is examined in detail (see Section 4). Specifically, they demonstrate that during the training of neural networks using Gradient Descent (GD), a stage of progressive sharpening transpires prior to the stabilization of sharpness near $2/h$, where h is the Learning Rate. Furthermore, another significant example comes from the study by Wang et al. (2022a), unveiling a *balancing* phenomenon (see Section 5). This phenomenon pertains to the convergence of weights in two-layer linear diagonal networks, showcasing a tendency for the weights to approach each other as compared to their initial values. This is in stark contrast to the behavior observed under small learning rates, where weight discrepancies remain relatively constant (Du et al.; Ye and Du, 2021; Ma et al., 2021). Consequently, large learning rates bias GD towards flat minima, even when the optimization is initiated arbitrarily close to a sharp minimum (see Corollary 3.3 in Wang et al. (2022a) and Section 5). It is worth noting that the initial phase of the balancing effect gives rise to *loss catapulting* (Lewkowycz et al., 2020), another intriguing empirical observation—during this phase, the loss experiences an initial increase (GD's escaping from the neighborhood of a sharp minimum) before subsequently decreasing. Collectively, these phenomena will be referred to as *large learning rate phenomena* thereafter.

Remarkable theoretical studies have been conducted to analyze EoS and balancing concerning specific objective functions (for example, Chen and Bruna, 2023; MacDonald et al., 2023; Ahn et al., 2023; Damian et al., 2022; Zhu et al., 2023; Wang et al., 2022b; Kreisler et al., 2023; Arora et al., 2022; Ahn et al., 2022; Lyu et al., 2022; Song and Yun, 2023). Nevertheless, the following questions remain not yet fully resolved:

When and why do EoS and other large learning rate phenomena occur?

These questions serve as the primary motivation for our paper. Specifically, we generalize the objectives examined in recent theoretical work, for example, Wang et al. (2022a); Zhu et al. (2023); Ahn et al. (2023), and construct a family of functions that can be compared as fairly as possible, despite their different regularities. This allows us to investigate the question in a comparative fashion, understanding which functions admit a larger set of initial conditions that lead to EoS. Beyond that, we also use these functions to demonstrate that many large learning rate phenomena are implicit biases rooted in a unifying mechanism. While still being simplified models, they are designed to encapsulate key features of practical neural networks, including the composition of loss and activation functions, and are made amenable to quantitative analysis. Within this family, we then prove that the combined influence of good regularity and the preferences of large-learning-rate GD toward flatter minima leads to balancing, and makes EoS more likely (i.e. there is a larger set of initial conditions leading to EoS). These rigorous results lead us to conjecture that regularity

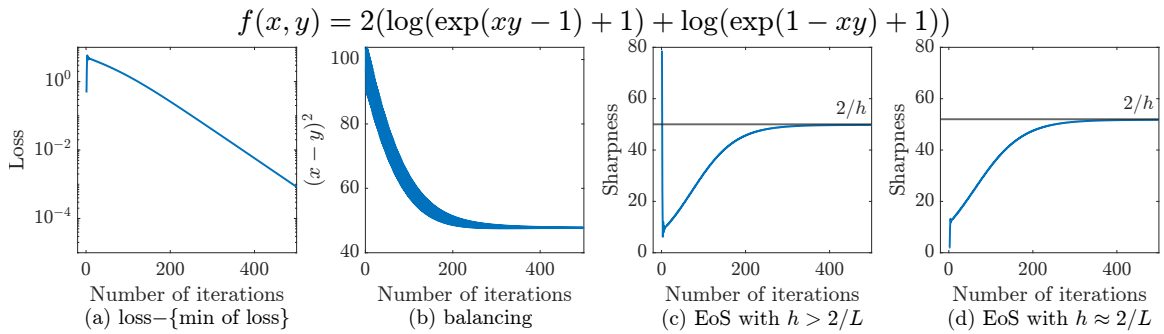


Figure 1: Large learning rate phenomena. The learning rate is chosen to be $h = 4/(x_0^2 + y_0^2)$ for all the figures. (a)(b)(c) share the same initial condition $x = 0.2, y = 10$, while for (d), $x = 2, y = 10$. (a) shows the decay of loss, i.e., $f - \min f$, and exhibits a catapult phenomenon at the beginning of the iterations. (b) shows the balancing phenomenon. (c) and (d) illustrate EoS under different initial conditions: if GD starts at large sharpness (c), the learning rate $h > 2/L$, where L is the local Lipschitz constant of the gradient in a bounded region where the GD trajectory belongs to, and there is a sharp decay at the beginning of the iterations before progressive sharpening; if it starts at small sharpness, the learning rate $h \approx 2/L$, and there is no de-sharpening stage (see more explanation in Section 4).

will similarly impact large learning rate phenomena for more general functions, and this conjecture is tested positively on empirical experiments with neural networks.

Before further explanation of our results, let us begin with some visualization of our claims. Figure 1 illustrates multiple instances of the aforementioned large learning rate phenomena, including the balancing phenomenon (Figure 1(b)), the catapult phenomenon in the loss decay (Figure 1(a)), and the edge of stability (EoS) shown in Figure 1(c) and (d). It is important to note that large learning rates do not always lead to these phenomena. Figure 2 depicts a scenario where we use an objective function with the same global minima but worse regularity compared to Figure 1. In this case, neither EoS nor balancing (including catapult) occurs, despite the use of large learning rates. The landscapes of these two functions are visualized in Figure 3.

Now we are ready to provide slightly more detailed, but still high-level answers to the two key questions:

- **About 'when'.** Within our designed family of functions, we rigorously demonstrate that large learning rate phenomena depend on the regularity of the objective function. Roughly speaking, when the objective has good regularity, i.e., small degrees of regularity (See Definition 2 in Section 3), EoS and balancing are more prone to appear (in this paper, we focus on the EoS phenomenon appearing in the limiting behavior of GD). However, when the function has bad regularity (large degrees of regularity), these two phenomena are more likely to vanish. More precisely, whether there exists EoS and/or balancing is not a yes/no question. Instead, it is a quantitative one, focusing on how large is the set of initial conditions that exhibit such large learning rate

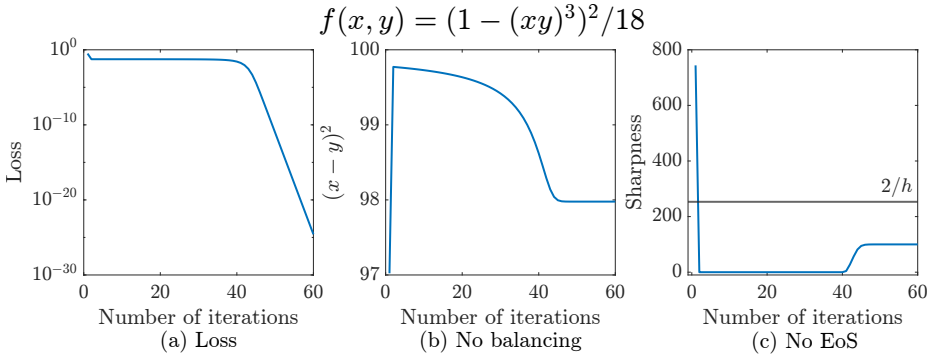
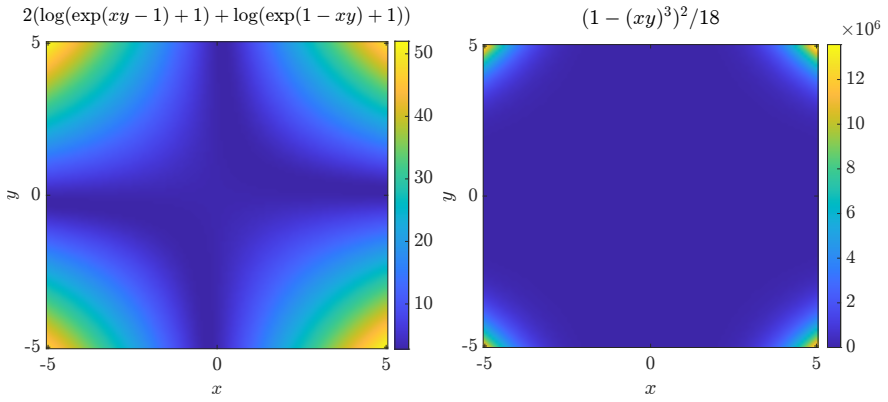


Figure 2: Disappearance of large learning rate phenomena. All the three figures share the same initial condition $x_0 = 0.15, y_0 = 10$ with the learning rate to be $h = 4/(x_0^2 + y_0^2)/(x_0 y_0)^{2b-2}$. It can be seen from (c) that the learning rate is indeed large, $> 2/L$, where L is the local Lipschitz constant of this trajectory. However, none of the large learning rate phenomena appear, i.e., no EoS, no balancing, and no catapult of loss.

phenomena. As the degree of regularity increases, the set of the initial conditions that does not lead to these phenomena will expand, whose boundary moves in accordance with $2^{\frac{1}{b-1}}$, where $2b$ is the degree of regularity of the objective function.

- **About 'why'.** Our theoretical analysis reveals that a crucial characteristic of these phenomena is the ability of large learning rates to steer GD toward flatter regions. In other words, the sharpness of the optimization landscape along GD iterations is influenced and controlled by how large the learning rate is. In fact, upon the convergence of GD, which is nontrivial due to the large learning rate but proved as a groundwork for our theory, the limiting maximum Hessian eigenvalue is at most $\tilde{\mathcal{O}}(h)$ below $2/h$ for functions of good regularities, i.e., the stability limit of sharpness $\leq 2/h$ is actually attained. But for bad regularity functions, this limit will be bounded by $1/h$ instead of the standard stability limit $2/h$. We name this nontrivial phenomenon as *one-sided stability*. Moreover, when the objective function has good regularity, GD is guaranteed to enter a region with sharpness less than the stability limit (the entry marks the onset of progressive sharpening), and then 'crawl' up in sharpness till around $2/h$ (i.e., edge of stability).

Furthermore, we demonstrate the practical implications of our theory in deep learning through neural network models. Different choices of loss functions and/or activation functions will lead to different regularities of the training objective functions. This will result in very different training dynamics with or without EoS and balancing, which may also lead to different generalization performances. Additionally, we consider a practical training technique—batch normalization (Ioffe and Szegedy, 2015) and show its capability of enhancing model regularity. When applied to bad regularity cases, batch normalization leads to the reemergence of the large learning rate phenomena. Collectively, these experi-



(a) values of a good regularity function (b) values of a bad regularity function

(c) sharpnesses of the good regularity function (d) sharpnesses of the bad regularity function

Figure 3: The landscapes and sharpnesses of good and bad regularity functions; different colors indicate different values. (a)(b) are landscapes of good and bad regularity functions. For (a), the value of the good regularity function changes slowly, resulting in a small range on the color bar. In contrast, the value of a bad regularity function (b) changes significantly in the region away from the minima, while the changes near the minima can be ignored. As a result, the color bar exhibits a very large range, making it difficult to distinguish most of the regions on this heat map. (c)(d) are sharpnesses of the two functions. For (c), the sharpness of the good regularity function concentrates on the minima $xy = 1$ while for other regions, the sharpness is very small. In contrast, for (d), the sharpness varies a lot outside the neighborhood of the minima.

ments consistently reinforce the significance of regularity in accordance with our theoretical framework.

We also remark on the nontriviality of our convergence analysis under large learning rates for a family of non-convex functions without globally Lipschitz continuous gradients. Especially, to the best of our knowledge, our work contains the first non-asymptotic result to quantify the rate of convergence of GD under large learning rates for nonconvex and non-Lipschitz-gradient functions (see Theorem 15).

★ **Main contributions** of our paper are summarized in the following, except the technical contributions, which will be deferred to Section 8.

- We demonstrate conditions for the occurrence of Edge of Stability and balancing within a class of functions.
- For bad regularity functions, we demonstrate the disappearance of EoS and a new, alternative phenomenon called *one-sided stability*; we also quantify the disappearance of balancing. This new phenomenon is different from the results of stability theory.
- We unify these phenomena and show they are different aspects of the same dynamical mechanism associated with the large learning rate.
- We provide the first rigorous bound of the global convergence rate of large-learning-rate GD for the family of functions defined in this paper (2), which are non-convex without globally Lipschitz gradients.

This paper is organized as follows: Section 2 contains the related works, basic notations, and what is a large learning rate regime. Section 3 includes the notion degree of regularity, and the family of objective functions considered in Section 3.1, as well as a discussion of these concepts for neural networks with and without batch normalization in Section 3.2. Section 4 and 5 explain, respectively, the reasons for the occurrence of EoS and balancing, together with intuitions of GD's mechanisms influenced by both the regularity and the large learning rate in Section 4. Section 6 is about the asymptotic and non-asymptotic convergence results of GD under large learning rates for the family of non-convex functions. Section 7 contains the validation of our theory in neural network applications. Section 8 contains additional discussions such as the limitations and the innovations of our theory.

2. Background

We first briefly review related works and then introduce notations and some preliminaries on gradient descent.

2.1 Related works on large learning rate phenomena in training/optimization

★ *Balancing.* Prior to the theory of the balancing phenomenon in Wang et al. (2022a), there were already a considerable amount of interesting theoretical results studying the balanced manifold under gradient flow. For example, in deep fully connected feedforward networks, the difference between the squares of Frobenius norms of weight matrices at different layers is preserved if they are 'trained' by gradient flow, which is the zero learning rate limit of

deterministic gradient descent (i.e., no minibatch, no momentum) (Du et al.; Arora et al., 2018). In order to gain further theoretical understanding, the community then simplified the model to the matrix factorization problem, i.e., $\min_{X,Y\in\mathbb{R}^{n\times m}} \|A - XY^\top\|^2$. Du et al. proved the convergence of GD under diminishing learning rate, which is an approximation of gradient flow, and thus $\|X\|_{\mathrm{F}}^2 - \|Y\|_{\mathrm{F}}^2$ is almost preserved throughout the GD iteration. Ye and Du (2021) later extended the convergence result to the constant small learning rate.

In contrast to the regime of these infinitesimal or small learning rates, Wang et al. (2022a) proved that large learning rate ($> 2/L$ and can be approximately $4/L$, where L is the local Lipschitz constant of the gradient) can disrupt such preservation and steer the GD trajectory towards a more balanced minimum. They coined this phenomenon as '*balancing*'. Later on, Chen and Bruna (2023) extended the scope of the large learning rate, and proved, by assuming all trajectories are positive, the convergence to an orbit of period 2 for scalar case, where almost perfect balancing can be achieved. They also explored periodic orbits in various settings, including a two-layer single-neuron network and the general matrix factorization problem. Balancing will also be a central focus of this work; however, we will not view it merely as a large learning rate effect anymore; instead, we will elucidate that another factor is also important, namely the regularity of the objective function. This way, balancing can be understood for a larger class of objective functions.

★ *Edge of Stability.* In a seminal paper where stability analysis of (S)GD was conducted, Wu et al. also observed that GD (SGD with full batch gradient) in certain experiments approximately achieves its largest stable sharpness, i.e., $2/h$. A comprehensive empirical characterization of this phenomenon, termed Edge of Stability (EoS), was first made by Cohen et al. (2021) with full batch gradient descent. Later, EoS was also observed in, for example, adaptive gradient methods with sufficiently large batch size (Cohen et al., 2022), and reinforcement learning (Iordan et al., 2023).

The intricacy and importance of EoS led to, in addition to empirical results, a rapid accumulation of theoretical investigations, e.g., (Chen and Bruna, 2023; MacDonald et al., 2023; Ahn et al., 2023; Damian et al., 2022; Zhu et al., 2023; Wang et al., 2022b; Kreisler et al., 2023; Arora et al., 2022; Ahn et al., 2022; Lyu et al., 2022; Wu et al., 2023; Song and Yun, 2023). Zhu et al. (2023) for example theoretically prove the $\approx 2/h$ limiting sharpness based on the convergence of GD for a specific problem. More precisely, they considered an objective function, which in the settings of this paper corresponds to a relatively large regularity ($b = 2$ and $\mathrm{dor} = 4$; see Equation 2 and Definition 2), and initialized very close to the set of minimizers. Their setup successfully reproduced the later stage of EoS, i.e., the sharpness converging to $2/h$, but not necessarily the earlier stage, progressive sharpening. Note that such initialization is the 'good' region of the bad (large) regularity function, which can still lead to EoS (see more details in Section 8) and our theory does not contradict their result. Wang et al. (2022b) adopted a two-layer linear network model, and used the norm of the output layers as an indicator for sharpness. They analyzed a four-phase[1] behavior of the output norm, which corresponds to a detailed division of progressive sharpening and limiting stabilization of EoS, including the increase/decrease of sharpness above $2/h$.

1. Their four phases correspond to more detailed divisions of the two stages of progressive sharpening and stabilization at $2/h$, and thus are different from our three stages of the EoS phenomenon (no de-sharpening, i.e., stage 1) and the two-phase convergence under large learning rates (no phase 1). See Section 4.

Damian et al. (2022) studied a similar four-phase behavior in EoS by leveraging third-order derivative information of the loss function. Ahn et al. (2023) analyzed a class of loss functions, corresponding to our setting with dor = 1, and provided a detailed estimate of the order of difference ($\mathcal{O}(h^{1/(\beta-1)})$ with $\beta > 1$) between limiting sharpness and $2/h$ for their loss functions. In this paper, we focus more on the reason why EoS occurs and design the objective functions such that this difference is upper bounded by $\mathcal{O}(h)$ when EoS appears, and $\mathcal{O}(h^{-1})$ otherwise. A follow-up work is Song and Yun (2023), which studied two-layer linear network and nonlinear network using ℓ^2 loss and activation with dor = 0 (output data is 0) from the bifurcation perspective, i.e., GD converging to a periodic orbit, and proved the limiting sharpness ($\approx 2/h$) as well as a later part of progressive sharpening. Kreisler et al. (2023) considered a local convergence of deep linear networks (scalar products) and proposed an approach to analyze GD via evolving gradient flow at each step. Lyu et al. (2022) studied scale-invariant loss (dor = 0), analyzed a local convergence of GD to the $\approx 2/h$ limit of its sharpness under normalization layers and weight decay, and showed that GD approximates a sharpness-reduction flow. While these results are undoubtedly remarkable, the effects of the large learning rate, especially in the early stage of iterations, and how it globally creates the three stages altogether, are not fully demonstrated. This will be one of the main focuses of this paper and a major factor contributing to the occurrence of EoS.

★ *More results related to large learning rate.* Apart from balancing and EoS, there is another phenomenon known as catapult (Lewkowycz et al., 2020) that will also be considered here. There, GD with $> 2/L$ learning rate perceives an early phase catapulting of the loss. In addition to these three phenomena, the impact of a large learning rate manifests in numerous other studies as well. Examples of empirical works include: Andriushchenko et al. (2023b) observed that a large learning rate will bias SGD towards sparse solutions; Jastrzebski et al. (2021) showed that in the early phase, SGD with a large learning rate penalizes the trace of the Fisher Information Matrix, which helps generalization.

Large learning rates can lead to complicated training dynamics, very different from that of gradient flow which only characterizes the infinitesimal learning rate regime. The effects of small-to-intermediate learning rates can be partially characterized via the tool of modified equation (Li et al., 2019). However, as the learning rate continues to grow, the modified equation will stop to provide a good approximation (Kong and Tao, 2020), and interesting implicit biases such as those discussed here can manifest. When the learning rate becomes even bigger, the optimization dynamics can begin to invalidate the naive understanding that GD iterations either converge to a (fixed) point or diverge, even if the learning rate is fixed as a constant. The iterations may converge to higher-order dynamical structures beyond a fixed point. For example, Kong and Tao (2020) studied the case where the objective function is $L = \mathcal{O}(1/\epsilon)$-smooth but the learning rate is $o(1) \gg 2/L$, for which the convergence to a chaotic attractor is proved, via a route of period-doubling bifurcations, i.e., GD first convergent to 1-periodic orbit (i.e., fixed point), then 2-periodic orbit, 4-, 8-, and all the way to ∞-period; based on this chaotic convergence pattern they constructed a quantitative theory of local minima escape without stochastic gradients. Chen and Bruna (2023) proved the convergence to a 2-periodic orbit with a larger learning rate for specific 1D and 2D functions under specific assumptions. Song and Yun (2023) also studied the convergence to a 2-periodic orbit.

Apart from constant large learning rate, another line of research studies the acceleration of convergence to the minimum by using cyclic learning rates, which combine small learning rate leading to slow convergence with very large learning rate leading to divergence (e.g., Young, 1953; Altschuler, 2018; Oymak, 2021; Grimmer, 2024; Altschuler and Parrilo, 2025).

2.2 Related works on flatness and generalization

Training with large learning rate GD often empirically leads to better generalization. A popular belief is that this is a consequence of GD's preference for flat minima under large learning rates (e.g., Seong et al., 2018; Lewkowycz et al., 2020; Yue et al., 2023; Smith, 2018). The fact that large learning rates lead to flat minima has been theoretically confirmed and quantified, for example, by Wu et al.; Wang et al. (2022a). Theoretical explorations of flat minimum leading to better generalization have also been fruitful, including those based on PAC-Bayes bounds (Langford and Caruana, 2001; Dziugaite and Roy, 2017), information theory (Hinton and Van Camp, 1993; Hochreiter and Schmidhuber, 1997; Negrea et al., 2019), and Rademacher complexity (e.g., Nacson et al., 2022; Gatmiry et al., 2023), although alternative/more precise statements also exist (e.g., Dinh et al., 2017; Andriushchenko et al., 2023a; Wen et al., 2023). Also worth mentioning is the recently proposed but very popular approach of sharpness-aware minimization (Foret et al., 2021), which explicitly seeks smaller sharpness for better generalization.

2.3 Notations

Gradient descent (GD) iteration for solving a general nonconvex optimization problem $\min_u f(u)$ is given by $u_{k+1} = u_k - h\nabla f(u_k)$, where $h > 0$ is the learning rate (a.k.a. step size).

For problems considered in this article (see Equation 2 for more details), u contains two components, x and y, and thus we will also use (x_k, y_k) to denote the kth iterate of GD. Let S_k denote the largest eigenvalue magnitude of Hessian $\nabla^2 f$ evaluated at the kth iteration (x_k, y_k). It serves as a quantification of the local geometry of the landscape and will be referred to as the sharpness (value). We use (x_∞, y_∞) and S_∞ to denote the limiting point of GD iteration and the limiting sharpness respectively.

We denote L to be the Lipschitz constant of the gradient in the bounded region where the GD trajectory lies. We use L_{glo} to denote the global Lipzchitz constant of the gradient. We use L_0 and L_∞ to represent the local Lipschitz constant of the gradient in the neighborhood of the initial point (x_0, y_0) and the limiting point (x_∞, y_∞) respectively.

We follow the theoretical computer science convention and use $\mathcal{O}(\cdot)$ to indicate the order of a quantity and $\tilde{\mathcal{O}}(\cdot)$ for its order omitting logarithmic dependence. Although we consider large learning rate h throughout this paper, note that the term 'large' is a relative concept based on local geometry, and more precisely, h depends on x_0, y_0 (see detailed bounds of h in Section 6). Therefore the term $\mathcal{O}(h)$ is taken in the sense of $x_0^2 + y_0^2 \to \infty$. We use $\lesssim$ such that $g_1 \lesssim g_2$ means $g_1 \le g_2 + c\frac{\log(x_0^2+y_0^2)}{x_0^2+y_0^2}$ for some universal constant $c > 0$.

Finally, given two matrices A and B, $A \preceq B$ ($A \prec B$) means $B - A$ is positive semi-definite (positive definite). $\|\cdot\|_{\text{F}}$ denotes matrix Frobenius norm.

2.4 Learning rate in gradient descent

What learning rate should be used in gradient descent? 2nd-order derivative information of f, if available, is always helpful. More precisely, if the gradient ∇f is globally L_{glo}-Lipschitz, then L_{glo} plays an important role in determining/interpreting the learning rate to use. In fact, a well-known nonconvex optimization result states the following (see, e.g., more detailed discussion in Wang et al. (2022a) Appendix F)

Theorem 1 *If $f : \mathbb{R}^N \to \mathbb{R}$ is twice differentiable and ∇f is L_{glo}−Lipschitz, i.e., $-L_{\text{glo}} I \preceq \nabla^2 f(u) \preceq L_{\text{glo}} I$ for all $u \in \mathbb{R}^N$, then with $h < \frac{2}{L_{\text{glo}}}$, GD converges to a stationary point, i.e., a point where ∇f vanishes.*

This theorem is based on the global Lipschitzness of the gradient. However, objective functions in machine learning often lack *globally* Lipschitz continuous gradients. One strategy is to assume that the trajectory (i.e., $u_0, u_1, \cdots$) is located in a bounded region, in which the *local* Lipschitz constant (defined as the maximum eigenvalue magnitude of the Hessian $\nabla^2 f$ over the region) is upper bounded by some constant L. Altogether, it is most common to consider the following learning rate regime:

$$\text{Regular learning rate regime: } 0 < h < \frac{2}{L},$$

where the convergence of GD can be well understood. There are finer divisions of this regime with various behavior of GD, e.g., the infinitesimal learning rate regime which approximates the gradient flow, and a relatively large regime $\frac{1}{L} \leq h < \frac{2}{L}$, where GD may oscillate near the minima until convergence.

Nevertheless, Theorem 1 only demonstrates a sufficient condition for convergence under nonconvex settings. A learning rate larger than $2/L$ may also lead to the convergence of GD. This case will be the main focus of this paper and will be referred to as:

$$\text{Large learning rate regime: } \frac{2}{L} \lesssim h \lesssim \frac{C}{L} \text{ for some } C > 2.$$

There is in general no guarantee of convergence in this regime. In certain situations, using a very large learning rate may still result in convergence, but not necessarily to a stationary point (e.g., (Kong and Tao, 2020; Chen and Bruna, 2023)); however, this article considers cases where GD does converge to a minimizer (i.e., a stationary point), and we will show that there is actually such a regime, for which $C \approx 4$. This regime is where many of the interesting large learning rate phenomena occur.

3. Degree of regularity

Before we proceed with our main results, we first introduce a new family of functions and how we measure their regularities.

3.1 A family of functions with different degrees of regularities

Apart from the large learning rate, the regularity of the objective function also impacts large learning rate phenomena. Proposing and demonstrating this is one of the main contributions

of this article. To this end, we first define the following notion in order to quantify the effect of regularity later.

Definition 2 (Degree of regularity) *Given function $F(s) : \mathbb{R} \to \mathbb{R}$, its degree of regularity is*

$$\mathrm{dor}(F) = \inf\left\{n : |F(s)| \le C_1|s|^n, \text{ for } |s| \ge C_0 \text{ with some constant } C_0, C_1 > 0\right\}. \tag{1}$$

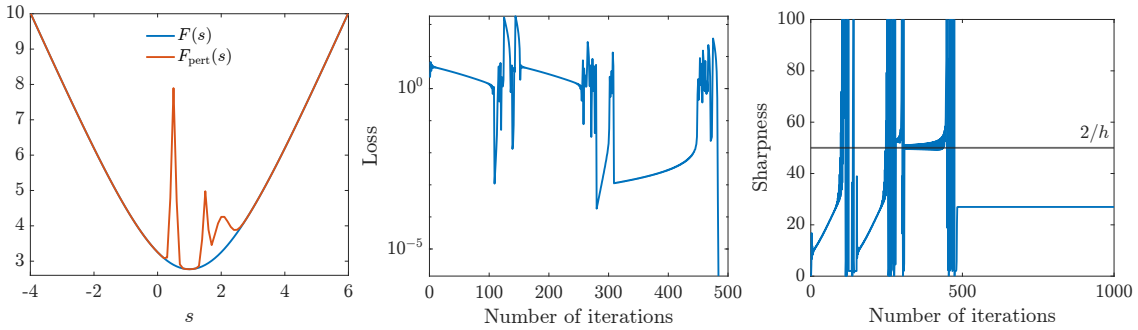
Remark 3 *One might wonder about the rationale behind our terminology. The intuition is that if a function F has a high dor, its $(k+1)$th derivative can possibly be much larger than its kth derivative at large enough s (e.g., think about a polynomial). This blowing-up phenomenon through differentiations can persist for $0 \le k \ll \mathrm{dor}(F)$, i.e., for many orders of derivatives. Therefore, the function has poor regularity. Similarly, low dor is a qualitative*[2] *indicator of good regularity.*

A preview of our results is that when $\mathrm{dor}(F)$ is small, it corresponds to good regularity functions in our function class, and we will show the large learning rate implicit biases of EoS and balancing are more likely to appear. In contrast, large $\mathrm{dor}(F)$ suggests bad regularity and may lead to the disappearance of these phenomena. See details in Section 4 and 5. The following remark clarifies the relation between dor, regularity, and large learning rate phenomena.

Remark 4 *We clarify that the above definition may not be the most general condition since we don't yet have a theory that can analyze any function. Degree of regularity (dor) will only serve as an initial attempt to understand how our specific family of functions used in* (2) *corresponds to implicit biases of large learning rate. It may appear to be more like a growth condition used in various fields, such as Banach algebra (Stampfli and Williams, 1968), optimization (Drusvyatskiy and Lewis, 2018), and partial differential equations (Marcellini, 2021). It also resembles the idea of generalized smoothness condition in non-convex optimization; see for example Li et al. (2023); Zhang et al. (2019). However, this is not the first time when growth condition is related to the regularity of the function; see, for example, Giaquinta (1987, 1988)*[3]*. Between growth condition and regularity, we conjecture that regularity (in the sense of Sobolev) is a more intrinsic factor of the implicit biases considered here.*

Note dor is a global property and it is our intention to have a global condition for regularity. This is because we conjecture that local behaviors of a function, such regularities near the minimizer or at infinity, are insufficient to guarantee large learning rate phenomena considered in this article. For example, Figure 4 shows a perturbed version of a good regularity function $f(x, y)$ with $a = 1$ in (2)*, denoted as $f_{\mathrm{pert}}(x, y)$. As is shown in the figure, when far away from or very close to the minima, it overlaps with the good regularity function, sharing exactly the same growth condition both globally towards infinity and*

2. Note: not quantitative, as our explanations were just intuitions. Rigorous minds can just consider 'dor' as a name.
3. Our definition of degree of regularity is still different from the relation between growth condition and regularity in these papers. Nevertheless, this is our initial attempt and is sufficient for our family of functions.



(a) $F(s)$ vs $F_{\text{pert}}(s)$ (b) loss-{min of loss} (c) no EoS

Figure 4: Perturbed version of $f(x,y) = 2(\log(\exp(xy-1)+1) + \log(\exp(1-xy)+1))$, the good regularity function. Let $s = xy$. In (a), the blue curve is $F(s) = 2(\log(\exp(s-1)+1) + \log(\exp(1-s)+1))$; the red curve is $F_{\text{pert}}(s) = e^{-10(s-2)^2} + 2e^{-100(s-1.5)^2} + 5e^{-100(s-0.5)^2} + F(s)$. In (b)(c), we apply GD on $f_{\text{pert}}(x,y) = F_{\text{pert}}(xy)$. The initial condition is $x_0 = 10, y_0 = 0.15$; the learning rate is $h = \frac{4}{x_0^2+y_0^2}$.

locally at the minimizer. Nevertheless, the Sobolev norm increases after the perturbation, i.e., $\|f(x,y)\|_{W^{k,p}} < \|f_{\text{pert}}(x,y)\|_{W^{k,p}}$ for Sobolev space $W^{k,p}$ with $k \in \mathbb{N}$ and $1 \le p < \infty$. This means the regularity of the function becomes worse. As numerically evidenced, EoS can still disappear.

In addition, the notion of EoS considered in this article refers to the entire process, including not only the final stage of sharpness stabilization around $2/h$, but also earlier preparational stages. We will use this terminology, unless EoS is otherwise specified to be the final stage itself. That is another reason why the global behavior of the objective function matters as GD can navigate across a large region throughout the full process.

In order to analyze the effect of regularity, we would like to design a family of $\mathcal{C}^2$ functions with various degrees of regularities. Wang et al. (2022a) studied a matrix factorization problem with the objective $\|A - XY^\top\|_{\text{F}}^2$. When A is a scalar, it becomes $(A - x^\top y)^2$ where x, y are vectors of the same dimension; if we let $s = x^\top y$, then $f(x,y) = F(s) = (A-s)^2$ and $\text{dor}(F) = 2$. Zhu et al. (2023) analyzed a minimalist[4] example for EoS, which is $f(x,y) = F(xy) = (1-(xy)^2)^2$ with $\text{dor}(F) = 4$ (see Section 8). Ahn et al. (2023) examine a class of Lipschitz functions on xy, i.e., dor = 1, including $F(xy) = \log(1+\exp(-xy)) + \log(1+\exp(xy))$ and $\sqrt{1+(xy)^2}$. Inspired by these designs, we define a more general class of functions, with dor ranging from 0 to $+\infty$ as follows

$$f(x,y) = F(xy) = \begin{cases} C_a(\log(e^{xy-1}+1) + \log(e^{1-xy}+1))^a \\ \qquad \text{with constant } C_a = \frac{1}{a2^{a-2}\log^{a-1}(2)} & \text{, for } 0 < a \le 1, a \in \mathbb{R} \\ C_b(1-(xy)^b)^2 \\ \qquad \text{with constant } C_b = \frac{1}{2b^2} & \text{, for } b = 2n-1, n \in \mathbb{Z}. \end{cases} \tag{2}$$

4. But it doesn't mean EoS cannot appear for simpler functions.

Note the 1st and 2nd lines of (2) respectively correspond to dor ≤ 1 and dor > 1. At first glance, the splitting of the definition seems unnatural, but in fact note that the 1st line can be seen as a smoothened version of the 2nd line: when $b < 1$, the gradient of $(1-(s)^b)^2$ at the minimum point $s = 0$ is singular; by smoothing this function using the idea inspired by Ahn et al. (2023) into that in the 1st line, its gradient becomes $\mathcal{C}^1$ for all the points.

In addition, the design is such that the locations of all minimizers are fixed, specifically at $xy = 1$, for all members of the function class. This way, it is fairer and more consistent to compare GD's convergence for this class. Note this choice of global minima introduces a more intricate landscape compared to the design in (Ahn et al., 2023) where the minima are $\{(x,y)\,|\,xy = 0\}$. The reason for complicating the functions in order to move their minimizers lies in the variation of topological structure, where $xy = 1$ exhibits two distinct branches while it degenerates into a single connected set in the case of $xy = 0$. Moreover, our results for global minima $xy = 1$ can be easily generalized to any non-zero value, not necessarily 1. An additional remark is that the global minima $xy = 1$ are also different from $xy = \pm 1$ used in Zhu et al. (2023), whose objective function shows similar landscape but additional symmetry, and that's why in this work we restrict b to odd values, ruling out the even. However, we conjecture that our analysis has the potential to be extended to the case with minima $xy = \pm 1$.

Finally, note the constant coefficients are chosen such that the largest eigenvalue of Hessian (i.e., sharpness) at any minimizer is always $x^2 + y^2$, regardless of the value of a or b, which contributes to a fairer comparison as well.

We summarize the properties of this family of functions in the following.

Proposition 5 *This family of functions* (2) *have the following properties:*

- $f \in \mathcal{C}^2(\mathbb{R}^2)$*;*
- *All the minima of these functions are global minima, located at* $xy = 1$*;*
- *The sharpness at any minimizer* (x,y) *is* x^2+y^2*, for all these functions;*
- *The degree of regularity satisfies* $\operatorname{dor}(F) = \begin{cases} a, & \text{for } 0 < a \le 1, \\ 2b, & \text{for } b = 2n-1. \end{cases}$

3.2 (Degree of) regularity of neural network training objective

★ **Toy model.** Although still far from being realistic, the above functions (2) are designed to emulate the setup of a neural network, particularly combining both the loss function and the activation function. To supplement our theory, Section 7 will provide empirical experiments on practical neural networks. But for now, let us first discuss some analytical connections. For example, consider the neural network model defined in the following problem

$$\min_{W_1,\cdots,W_\ell} f(W_1,\cdots,W_\ell) = \sum_i \mathcal{L}(g(W_1,\cdots,W_\ell;u_i^{\text{input}}),u_i^{\text{output}}), \tag{3}$$

where $g = W_\ell\sigma_{\ell-1}(W_{\ell-1}\sigma_{\ell-2}(\cdots\sigma_1(W_1u_i^{\text{input}})))$; $\mathcal{L}(\cdot,\cdot)$ is the loss function, $\sigma_i(\cdot)$ is the activation function, $(u_i^{\text{input}},u_i^{\text{output}})$'s are data pairs, and W_i's are weight matrices. To better

understand the regularity in neural network model (3), we consider the following toy model of a 3-layer neural network trained on one data point $\{u^{\text{input}} = 1, u^{\text{output}} = 1\}$: the first layer with n linear neurons, the second layer with just 1 (nonlinear) neuron, and last layer fixed (weight assumed WLOG to be 1). Then the training objective function is

$$f(W_1, W_2) = \mathcal{L}(\sigma(W_2 W_1), 1), \text{ where } W_1, W_2^\top \in \mathbb{R}^n. \tag{4}$$

This function could again be rewritten as $F(W_1 W_2) = f(W_1, W_2)$ and we have under certain conditions,

$$\operatorname{dor}(F) = \operatorname{dor}(\mathcal{L}(\cdot, 1))\operatorname{dor}(\sigma). \tag{5}$$

More details including the proof are in Appendix B.

This means the regularity of this objective function depends on two parts: one is the neural network model g which is further based on the regularity of the activation function, and the other is the loss function $\mathcal{L}$.

★ **Basic components of neural network.** Now let us exemplify some loss functions and activation functions with various regularities.

Example 1 *Degrees of regularities of some loss functions*

- *ℓ^2 loss:* $\mathcal{L}(s, \cdot) = \frac{1}{2}(\cdot - s)^2$, $\operatorname{dor}(\mathcal{L}) = 2$;
- *Huber loss:* $\mathcal{L}(s, \cdot) = \begin{cases} \frac{1}{2}(\cdot - s)^2, & |\cdot - s| \le \delta \\ \delta(|\cdot - s| - \frac{1}{2}\delta), & \text{otherwise} \end{cases}$, $\operatorname{dor}(\mathcal{L}) = 1$;

Degrees of regularities of some activation functions

- *Hyperbolic tangent* tanh*:* $\sigma(s) = \frac{e^s - e^{-s}}{e^s + e^{-s}}$, $\operatorname{dor}(\sigma) = 0$.
- *(Leaky) ReLU function:* $\sigma(s) = \begin{cases} \alpha s, & s < 0 \\ s, & s \ge 0 \end{cases}$ *for some* $0 \le \alpha \le 1$, $\operatorname{dor}(\sigma) = 1$.
- *$ReLU^k$ function*[5]*:* $\sigma(s) = \max(0, s^k)$, $\operatorname{dor}(\sigma) = k$.

According to the above discussion, the degree of regularity of the neural network objective varies as either the loss or the activation changes. Moreover, large learning rate phenomena may not occur for certain choices of loss and activation (i.e., when dor is large) if no extra techniques are applied to help training. Empirical results on more practical network models are consistent with our theory as shown in Section 7, where loss and activation functions with different regularities are tested.

Next, we will show that a certain training method—batch normalization—can eliminate the effect of choices of loss and activation that otherwise lead to bad regularity.

★ **Batch normalization.** Batch normalization (Ioffe and Szegedy, 2015) is an efficient approach to enhance the robustness of the networks and improve both optimization and

5. Applications and analysis of this activation can be found in e.g. Yang and Zhou (2024); Luo and Yang (2024); Gao et al. (2023).

generalization. Existing works propose that batch normalization can smoothen the landscape, i.e., decrease the sharpness, with a larger learning rate (Lyu et al., 2022; Santurkar et al., 2018; Bjorck et al., 2018; Ghorbani et al., 2019; Karakida et al., 2019). We reinforce this perspective and suggest that batch normalization helps reduce the degree of regularity of the model. Consider the general objective function (3). If we add batch normalization to each layer and implement (full batch) GD, i.e.

$$g_{j+1} = \sigma_j(\mathrm{BN}(W_j g_j)), g_1 = (u_1^{\mathrm{input}}, \cdots, u_N^{\mathrm{input}}), \text{ where } N \text{ is the number of data}, \tag{6}$$

then we can consider $\tilde{\sigma}(\cdot) = \sigma(\mathrm{BN}(\cdot))$ to be batch normalized activation function for some activation σ. Given N inputs $\{v_i\}_{i=1}^N$ with $v_i = (v_i^{(1)}, \cdots, v_i^{(n)})^\top \in \mathbb{R}^n$, we have

$$\tilde{\sigma}(v_i^{(k)}) = \sigma\left(\frac{v_i^{(k)} - \mathrm{Mean}(v^{(k)})}{\sqrt{\mathrm{Var}(v^{(k)}) + \epsilon}}\right),$$

where $\mathrm{Mean}(v^{(k)}) = \frac{1}{N}\sum_{i=1}^N v_i^{(k)}$, $\mathrm{Var}(v^{(k)}) = \frac{1}{N}\sum_{i=1}^N (v_i^{(k)} - \mathrm{Mean}(v^{(k)}))^2$, and $\epsilon > 0$ is a very small hyperparameter. Then with high probability, $\tilde{\sigma}(v)$ is bounded. By Definition 2, we have $\mathrm{dor}(\tilde{\sigma}) \approx 0$, which is independent of $\mathrm{dor}(\sigma)$. If we consider the above 3-layer toy example, then

$$\mathrm{dor}(F) = \mathrm{dor}(\mathcal{L}(\cdot, 1))\mathrm{dor}(\tilde{\sigma}) \approx 0.$$

This shows that batch normalization can effectively drive the training objective function from bad (large) regularity to good (small) regularity. We also experimentally validate our theory on how batch normalization affects regularity and ramifications in Section 7, where large learning rate phenomena reappear after batch normalization is added to a model with bad regularity.

Despite all the discussions above, we emphasize again that, the theories constructed in this article are still not enough for analyzing general neural networks, although they can already work for a larger class of objective functions than those considered in the literature.

Now we are ready to quantify large learning rate phenomena under different regularities. See Section 4 and 5, and note results there will be based on the convergence theorems in Section 6.

4. Edge of stability

Section 2 mentioned that EoS is a large learning rate phenomenon, and we now provide more detailed explanations. Its original description (Cohen et al., 2021) contained two stages, namely progressive sharpening, and limiting sharpness stabilization around $2/h$. Ahn et al. (2023) later observed a third stage before progressive sharpening (which will be referred to as pre-EoS) where the sharpness will first decrease before increasing, and more empirical investigations of sharpness reduction in early training were then conducted by Kalra and Barkeshli (2023). A description of the full process is the following:

- **Pre-EoS (de-sharpening).** This stage characterizes the situation where at the very beginning of the iterations, the sharpness decreases sharply before the occurrence of the well-known EoS (see Figure 1(c) and Ahn et al. (2023)). It does not necessarily

come with all the EoS phenomenon (see Figure 1(c) and (d)) and only appears when the initial sharpness is very large. Nevertheless, it helps demonstrate the behavior of GD under large learning rates.

- **Progressive sharpening.** This stage is governed by increasing sharpness. Due to pre-EoS, GD is guaranteed to start in a relatively flat region (small sharpness) when this stage begins, even if it was initialized in a sharp region. The minimizer that GD will eventually converge to has larger sharpness than the majority of sharpness values along GD trajectory in this stage, and as GD progresses, the sharpness value 'crawls' up. Such behaviors stem from the good regularity of the objective function and do not necessarily appear in functions with bad regularities.

- **Limiting sharpness near** $2/h$**.** Stability theory of the GD dynamics (see Appendix A) guarantees that the limiting sharpness has to be not exceeding $2/h$, but not necessarily close to $2/h$. It is the good regularity of objective function that will drive the final sharpness towards $2/h$.

In this section, we establish a rigorous theoretical framework that elucidates the above three stages, and quantify our claim that EoS is more likely to occur when the objective function exhibits good regularity. In the context of the progressive sharpening in EoS, we will primarily focus on the concept of 'sharpening' rather than 'progressive' due to the subjective nature of this word: this stage can actually exhibit diverse sharpening speeds (see Section 7).

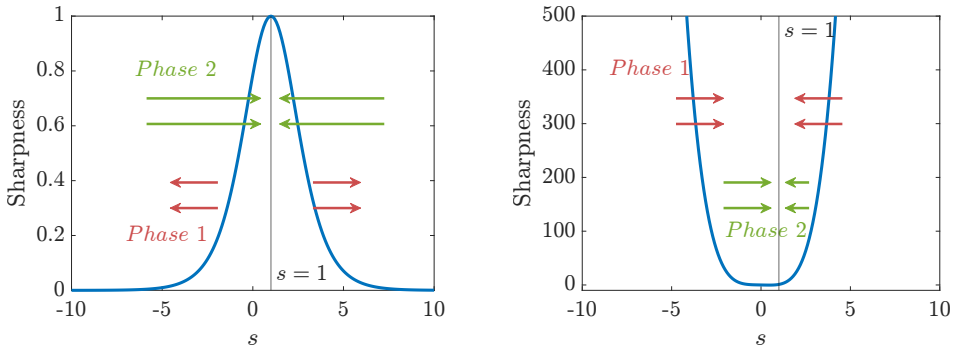
In order to better understand the insight of EoS, we commence by exploring the two-phase convergence of GD under large learning rates, as initially proved in Wang et al. (2022a) for matrix factorization problem:

- *Phase 1*, GD tries to escape from the attraction of sharp minima and enters a flat region, where the local Lipschitz constant of the gradient is upper bounded by $\approx 2/h$.

- *Phase 2*, GD converges inside the flat region.

This two-phase convergence pattern is also provably true for our family of functions under large learning rates. Detailed convergence analyses are deferred to Section 6.

★ **Intuition for de-sharpening and progressive sharpening.** To illustrate the intuition, let $s = xy$ and consider $F(s)$ in (2) (note x, y follow GD trajectories but s does not). Consider the sharpness of $F(s)$ (i.e., $F''(s)$) in two example cases, one with good regularity and the other with bad regularity. As shown in Figure 5, the sharpness of good regularity function concentrates near the minimum ($s = xy = 1$) and vanishes when far away from it (Figure 5(a)), while the situation with bad regularity function is the opposite (Figure 5(b)). Therefore, in *Phase 1* as GD escapes to a flatter region, GD travels away from the minimum for the good regularity function, while for the bad regularity function, GD moves closer to it. This corresponds to the **de-sharpening** stage that can occur in both cases. In *Phase 2*, GD starts to converge but results in different behaviors. In the good regularity case (Figure 5(a)), GD faces a longer path, requiring its sharpness to 'climb up the hill' to converge. This process allows sufficient time for sharpness to increase (**progressive sharpening**) and eventually reach its peak, namely **limiting sharpness near** $2/h$. In contrast, in the bad



(a) good regularity (b) bad regularity

Figure 5: The sharpness of different $F(s)$ with good (small) and bad (large) regularities. The blue curves are the sharpness, i.e., $F''(s)$; the red arrows denote the direction of GD in *Phase 1*; the green arrows denote the direction of GD in *Phase 2*.

regularity case, GD is already closer to the minimum at the end of *Phase 1* when compared to its initial state. Consequently, the sharpness will not undergo significant changes before GD converges. As a result, the typical phenomena observed in good regularity scenarios do not occur.

Note, however, that the actual landscape in which GD navigates is $f(x, y) = F(xy)$ instead of $F(s)$, and the behavior of GD is thus much more complicated. For example, since $s = xy$, each point in Figure 5 actually represents a continuous set of points. The function f has a rescaling symmetry of (x, y), known as homogenity (Wang et al., 2022a), meaning that $(x, y) \mapsto (cx, y/c)$ will not change f's value for any $c \neq 0$. This transformation can drastically change the eigenvalues of Hessian $\nabla^2 f$, especially when $|c|$ is large. Therefore, the fact that $f(x, y) = F(xy)$ creates a combination of flat and sharp regions. In particular, the global minima of $F(xy)$ form an unbounded set, as opposed to one single point in $F(s)$. Even if we just consider these global minima alone, their sharpnesses, which are given by $x^2 + y^2$, already assume any values in $[2, \infty)$, creating sharp and flat regions per se. The complexity of the actual landscape is the groundwork of diverse phenomena like EoS and makes the following results for understanding them rather nontrivial.

For functions with good regularities (i.e., small degrees), the occurrence of EoS is linked to a large range of initial conditions as is shown in the following theorem.

Theorem 6 (EoS) *Consider* $0 < a \leq 1$ *in* (2). *Assume the initial condition satisfies* $(x_0, y_0) \in \{(x, y) : 1 < xy \lesssim M_1(a), x^2 + y^2 \gtrsim 4c_1^{-4/3}\} \backslash \mathcal{B}_a$, *for some* $c_1 = c_1(a) > 0$ *and* $M_1(a) > 3$, *where* $\mathcal{B}_a$ *is a Lebesgue measure-0 set. Let the learning rate be* $h = \frac{C}{x_0^2 + y_0^2}$ *for* $2 < M_2(a, x_0, y_0) \leq C \leq 4$, *where* $M_2(a, x_0, y_0) \lesssim 2.6$. *See definitions of* M_1, M_2, c_1 *in Theorem 13. The sharpness at* k*th iteration* S_k *satisfies the following properties:*

- *Entering flat region (end of pre-EoS, and preparation for progressive sharpening): There exists $N \in \mathbb{N}$, s.t.,*

$$S_N \lesssim \frac{1}{4}(6-C)S_\infty,$$

and note $\frac{1}{4}(6-C)S_\infty < S_\infty$ for all C defined above. For example, if $C = 4$, then $S_N \lesssim \frac{1}{2}S_\infty$.

- *Limiting sharpness: GD converges to a global minimum (x_∞, y_∞), and its sharpness S_∞ satisfies*

$$\frac{2}{h} - \tilde{\mathcal{O}}(h) \le S_\infty = x_\infty^2 + y_\infty^2 \le \frac{2}{h}.$$

Remark 7 *The above theorem embodies a detailed description of the whole EoS process. More precisely, there is **progressive sharpening**: since there is at least one point along the trajectory with small sharpness (smaller than limiting sharpness), the sharpness will eventually increase to the limiting sharpness. Note despite the growth of sharpness, GD still converges inside the flat region (i.e., in* Phase 2*) since the limiting sharpness is upper bounded by $2/h$. In the end, the **sharpness stabilizes near** $2/h$ within a distance of $\tilde{\mathcal{O}}(h)$. The **pre-EoS (de-sharpening)** can also occur if the initial condition (x_0, y_0) is close to the minima. In this case, the initial sharpness S_0 is approximately $x_0^2 + y_0^2$ and we have*

$$S_N \lesssim \frac{1}{4}(6-C)S_\infty \approx \frac{1}{4}(6-C)\frac{2}{h} = \left(\frac{3}{C} - \frac{1}{2}\right)(x_0^2 + y_0^2) \approx \left(\frac{3}{C} - \frac{1}{2}\right)S_0 < S_0, \text{ for } 2 < M_2 \le C \le 4$$

which means the sharpness will first decrease before GD enters the progressive sharpening stage. This also implies that the limiting sharpness is smaller than the initial sharpness, which corresponds to the balancing phenomenon (see Section 5).

Figure 6 shows a diagram of the proof idea for the limiting sharpness $\approx 2/h$. The complete proof is in Appendix D. Roughly speaking, good regularity and large learning rate control the two terms $\frac{F'(xy)}{xy-1}$ and $x^2 + y^2$ respectively, which result in a slow and oscillating convergence pattern of $xy - 1$. This convergence rate can be approximated by a function of sharpness $(1 - hS_k)$ at each step and thus eventually leads to the $\approx 2/h$ limiting sharpness.

Next, we will show that when the regularity is bad (large degree), there exists a large set of initial conditions that do not lead to EoS even though we use a large learning rate. Moreover, such a set expands as regularity gets worse (i.e., the degree of regularity becomes larger).

Theorem 8 (no EoS; one-sided stability) *Consider $b = 2n+1$ for $n \in \mathbb{Z}$ in (2). Assume the initial condition satisfies $(x_0, y_0) \in \{(x, y) : xy > 2^{\frac{1}{b-1}}, x^2 + y^2 \ge 4\} \backslash \mathcal{B}_b$, where $\mathcal{B}_b$ is a Lebesgue measure-0 set. Let the learning rate be $\frac{2}{(x_0^2+y_0^2+4)(x_0y_0)^{2b-2}} \le h \le \frac{M_3(b,x_0,y_0)}{(x_0^2+y_0^2+4)(x_0y_0)^{2b-2}}$, where $3 < M_3(b, x_0, y_0) \le 4$, and the precise definition of M_3 is given in Theorem 15. Then GD converges to a global minimum (x_∞, y_∞), and its sharpness S_∞ satisfies*

$$S_\infty = x_\infty^2 + y_\infty^2 \le \frac{1}{h}.$$

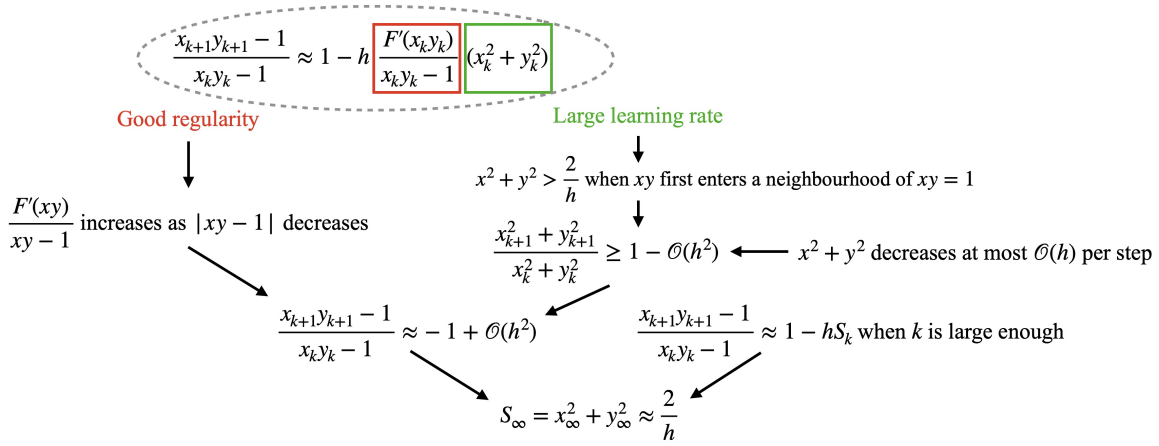


Figure 6: Proof idea of $S_\infty \approx \frac{2}{h}$ for good regularity functions. The formula inside the dashed ellipse is what we mainly consider. The term inside the red box is primarily affected by good regularity (also in red), while the one inside the green box is influenced by large learning rate (in green).

This theorem shows that for functions with bad regularities, the limiting sharpness is bounded by $1/h$, which is well below the stability limit of $2/h$. Therefore, not only is there no EoS, but also a different phenomenon occurs, which we refer to as **one-sided stability**. Note the lower bound of x_0y_0, i.e., $2^{\frac{1}{b-1}}$, decreases as b increases. This shows the inflation of the initial condition set that leads to the non-EoS phenomenon as the degree of regularity increases. See more explanation in 'Example implications of the general theory' at the end of this section.

Here is the intuition of why bad regularity gives $1/h$ which is very different from the $2/h$ stability limit. We denote $\tilde{L}$ to be the Lipschitz constant of the gradient in the bounded region of GD trajectory in *Phase 2*, the final converging phase, and L to be the Lipschitz constant of the gradient in the bounded region of whole GD trajectory. The key idea is that the sharpness of bad regularity functions varies far more than that of good regularity functions; this leads to $h \approx 2/\tilde{L}$ for good regularity but $h < 1/\tilde{L}$ for bad regularity. Before delving into further details, let us first clarify the difference between $h < 1/\tilde{L}$ and $1/\tilde{L} < h < 2/\tilde{L}$. As is shown in Figure 7, when $h < 1/\tilde{L}$, GD evolves along one side of the minimum point; when $1/\tilde{L} < h < 2/\tilde{L}$, GD jumps between the two sides. Suppose we start with a large learning rate h in a sharp region. The landscape along the GD trajectory will quickly become flat enough, i.e., the sharpness in *Phase 2* is small enough compared to those of the early iterations so that roughly $\tilde{L} < L/C$ for C defined in Theorem 8. Then, the large learning rate h, which is $> 2/L$, will eventually be $< 1/\tilde{L}$, and result in GD decreasing from only one side of the continuous set of minima thereafter, which is where the name 'one-sided stability' comes from. This $h < 1/\tilde{L}$ gives rise to the limiting sharpness bound $1/h$.

More precisely, Figure 8 sketches the proof idea of limiting sharpness $\leq \frac{1}{h}$. A complete proof is in Appendix F. For bad regularity functions, the sharpness varies a lot even if the function is just perturbed a little bit (see Figure 5). Together with large learning rate,

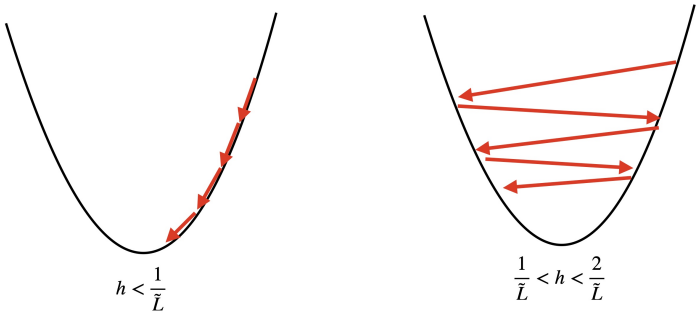


Figure 7: GD trajectory under different learning rate h. The black curve represents the objective function; the red arrows represent the evolution of GD iterations.

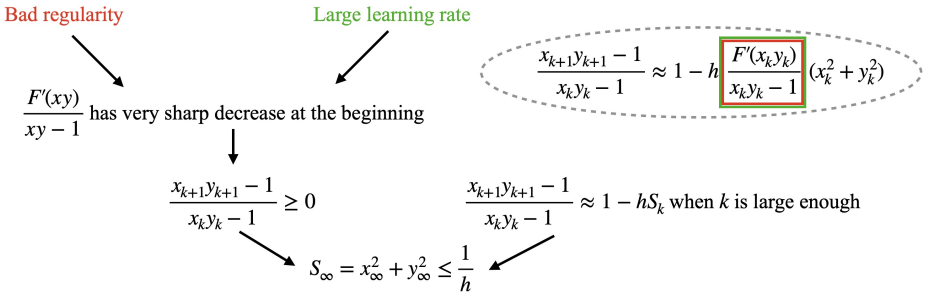


Figure 8: Proof idea of $S_\infty \leq \frac{1}{h}$ for bad regularity functions. The formula inside the dashed ellipse is identical to that in Figure 6. Different from good regularity cases (Figure 6), here bad regularity (red color) and large learning rate (green color) mainly affect the same term in the dash-circled expression.

the term $\frac{F'(xy)}{xy-1}$, which can be seen as an approximation of $F''(xy)$ near $xy = 1$, exhibits a sharp decrease at the beginning of the GD iterations; after that, $\frac{F'(xy)}{xy-1}$ will be small enough throughout the rest of the trajectory such that $xy - 1$ has monotone decrease, which leads to the $\frac{1}{h}$ bound of limiting sharpness.

We now demonstrate that our choice of the large learning rate is not 'cherry-picking', meaning the disappearance of the EoS phenomenon is not a consequence of specific choices of the learning rate, whether it is larger or smaller. Figure 9 exhibits the changes of sharpness (and also $(x_k - y_k)^2$) along GD trajectories under different learning rates chosen equidistantly until divergence. None of the plots shows EoS (and balancing) behavior since the sharpness stays far below $2/h$ at the limit.

Note that Theorem 8 does not contain the $b = 1$ case due to technical reasons. The convergence pattern of $b = 1$ is different from the one described in the above theorem, which can be shown in, for example, the initial condition set of Theorem 8: if $b = 1$, the lower bound $2^{\frac{1}{b-1}}$ of xy is $+\infty$. Nevertheless, our main focus is to show the effect of regularity on large learning rate phenomena. The results under $b \geq 3$ already serve this purpose on

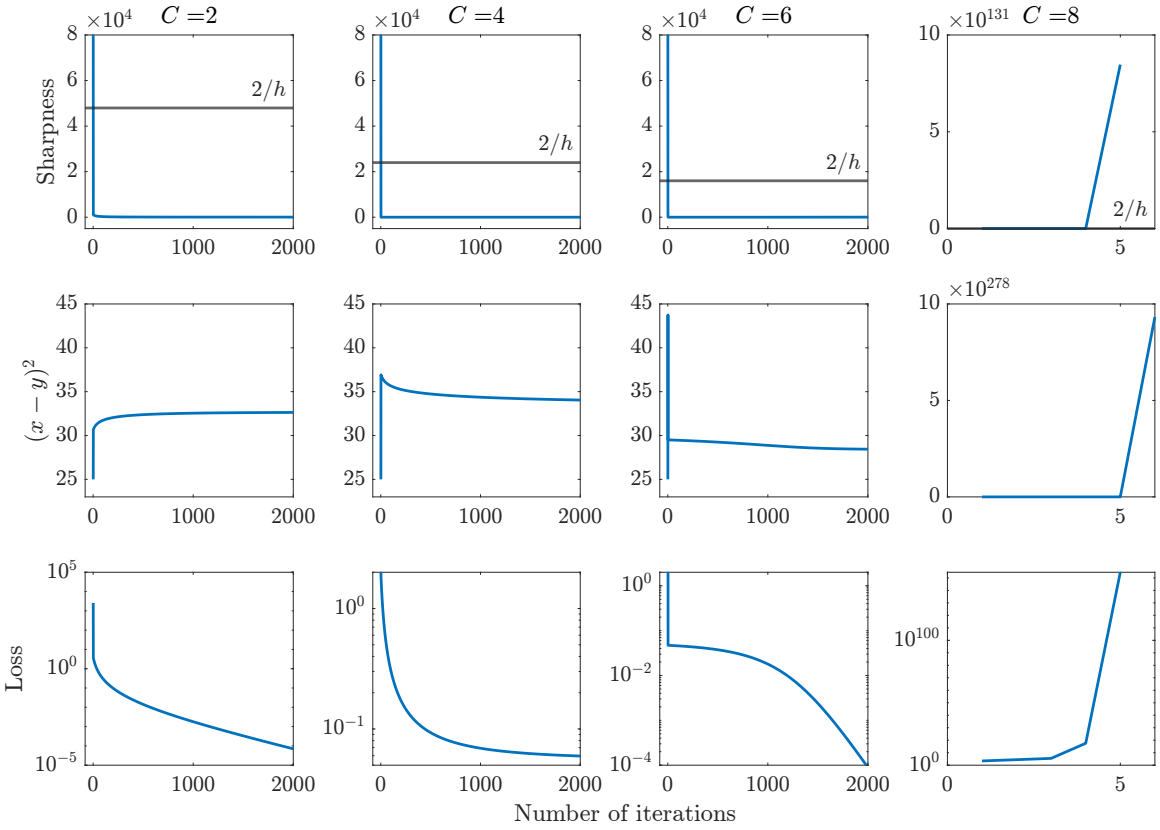


Figure 9: No EoS or balancing for bad regularity function (2) with $b = 3$ under different learning rates. All the figures share the same initial condition $x_0 = 6, y_0 = 1$; the learning rate is chosen to be $h = \frac{C}{(x_0^2+y_0^2)(x_0y_0)^{2b-2}}$ with $b = 3$, where $C = 2, 4, 6, 8$, until divergence.

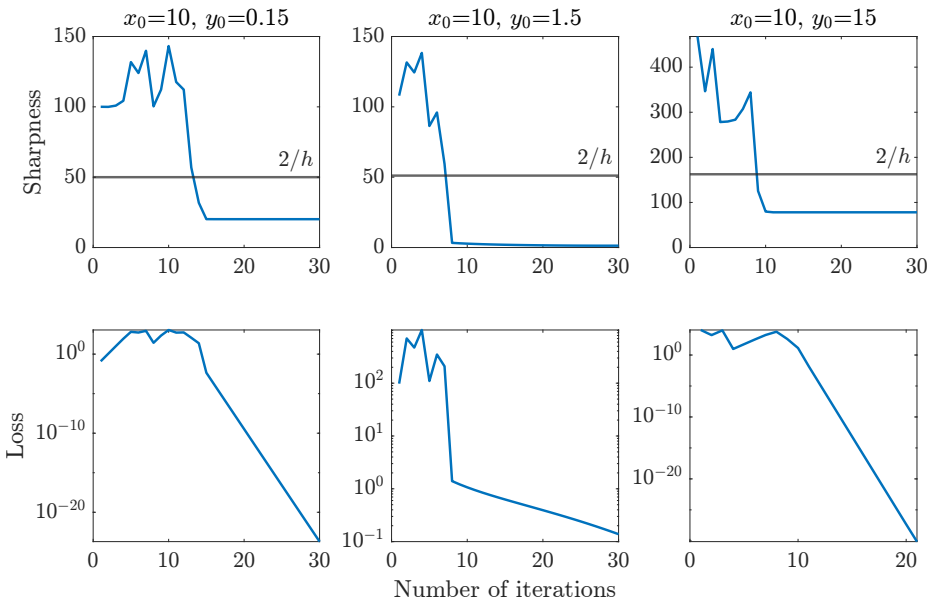


Figure 10: No EoS when $b = 1$. For the initial condition, we fix $x_0 = 10$, and choose y_0 to be 0.15, 1.5, and 15; the learning rates are all taken to be $h = 4/(x_0^2 + y_0^2)$. It turns out that none of the three cases have the EoS phenomenon. The convergence pattern of both the sharpness and the loss is very irregular, especially in the early stage of the convergence.

the bad regularity side. For completeness, we illustrate the $b = 1$ case by experiments (see Figure 10) and will leave the theoretical result for future exploration.

★ **Example implications of the general theory.** Consider the family of functions in (2), i.e., with both good and bad regularities.

- *Bad regularity eliminates the EoS phenomenon.* If we fix the initial condition set for all functions, i.e., take the intersection of all initial condition sets

$$(x_0, y_0) \in \{\sqrt{2} < x_0 y_0 \le 3, x_0^2 + y_0^2 \gtrsim \max_a 4c_1(a)^{-4/3}\} \backslash \mathcal{B}, \text{ where } \mathcal{B} = \bigcup_a \mathcal{B}_a \cup \bigcup_b \mathcal{B}_b, \quad (7)$$

$$\text{we have: when dor} \le 1,\ S_\infty \approx \tfrac{2}{h}; \text{when dor} > 2,\ S_\infty \le \tfrac{1}{h}.$$

  This means as the regularity increases, the same initial condition will lose the ability to lead to EoS, even if a large learning rate is used. Moreover, worse regularity (larger degree) results in a larger set of initial conditions that do not lead to EoS. When dor > 2, the initial condition set that does not show EoS expands as dor increases, i.e.,

$$(x_0, y_0) \in \{(x, y) : xy > 4^{\frac{1}{\mathrm{dor}(F)-2}}, x^2 + y^2 \ge 4\} \backslash \mathcal{B}, \text{ where } \mathcal{B} = \bigcup_b \mathcal{B}_b. \quad (8)$$

- *Larger learning rate leads to smaller limiting sharpness, i.e., flatter minima.* Previous results, as are based on the stability theory, state that *if* GD converges, then the

limiting sharpness is upper bounded by $\mathcal{O}(1/h)$; this is a more precise version of the common belief that larger h leads to a flatter minimum. The results above (Theorem 6 and 8), however, no longer require the assumption of convergence (i.e., the '*if*' clause), have it proved and yield the same $\mathcal{O}(1/h)$ bound, for both good and bad regularity cases.

Although we do not characterize all the initial conditions in the space, our results are still quantitative and suggestive of the effect of increasing regularity as reducing large learning rate phenomena.

4.1 Critical effect of large learning rate on EoS in both the good and bad regularity cases

In this section, we will first demonstrate that the learning rates used in Theorem 6 and 8 are indeed in the large learning rate regime, i.e., $h \gtrsim \frac{2}{L}$, where L is the local Lipschitz constant of the gradient. Then, we will theoretically elucidate how large learning rates change the local Lipschitz constant as GD evolves, in both the good and bad regularity cases.

For most of the initial conditions, our choices of learning rate h depend on L_0, the local Lipschitz constant near the initial condition (see Section 2.3). To see this, for good regularity functions (Theorem 6), if GD is initialized near the minima, then according to Proposition 5, the sharpness $S_0 \approx x_0^2 + y_0^2$, which is also $\approx L_0$. Therefore, $h = \frac{C}{x_0^2+y_0^2} \approx \frac{C}{L_0}$. Such L_0 controls the sharpness along the whole trajectory (see Figure 1(c) for example), namely, $L_0 \approx L$. Consequently, $h \approx \frac{C}{L}$ and additionally $\frac{2}{L} \lesssim h \lesssim \frac{4}{L}$. For bad regularity functions (Theorem 8), by Lemma 32, the lower bound of $h \gtrsim \frac{2}{L_0}$ and the upper bound of $h \gtrsim \frac{4}{L_0}$. Similarly, we have $h \gtrsim \frac{2}{L}$ (see for example Figure 2(c)).

Apart from the above situations, there exist exceptions of initial conditions, where their corresponding h does not depend on L_0. For good regularity functions (Theorem 6), if the initialization is away from the minima, L_0 is very small and in this case, and h depends on L_∞, which is the local Lipschitz constant near the limiting point (see Section 2.3). Then L is controlled by L_∞ (see for example Figure 1(d)), i.e., $L \approx L_\infty$. By Theorem 6, $L \approx L_\infty \approx S_\infty \approx \frac{2}{h}$ and thus $h \approx \frac{2}{L}$.

From a theoretical perspective, the mechanism of GD under large learning rates can be described in the following: when $h > \frac{2}{L}$, according to stability analysis (see Theorem 16 in Appendix A), GD cannot converge in the region where the minimum local Lipschitz constant $\approx L$. Instead, GD tends to search for a region with smaller local Lipschitz constant $\tilde{L} < L$ (which corresponds to *Phase 1* discussed above), s.t. $\frac{2}{L} < h < \frac{2}{\tilde{L}}$, i.e., h is inside the regular learning rate regime for the new flatter region (which is in *Phase 2*). $h \approx \frac{2}{L}$ is on the boundary of the two regions, and therefore there is no 'searching' for the flatter region.

5. Balancing

The existence of the balancing phenomenon is originally proved in matrix factorization problem with objective function $\frac{1}{2}\|A - XY^\top\|_{\mathrm{F}}^2$ (Wang et al., 2022a), for which GD optimization

can still converge when learning rate h exceeds $2/L$, and its limiting point (X_∞, Y_∞) satisfies

$$|\|X_\infty\|_{\mathrm{F}} - \|Y_\infty\|_{\mathrm{F}}|^2 < \frac{2}{c_1 h} - c_2, \text{ for some } c_1, c_2 > 0.$$

The larger the learning rate is, the smaller the gap between the magnitudes of X and Y will be, i.e., X and Y will become more balanced when compared to their initials.

One feature of balancing is that it can be used as an explicit characterization of sharpness. To see this, let us first recall that at any minimizer (satisfying $xy = 1$), the sharpness of any function in our function class (2) is x^2+y^2, by Proposition 5. Then at that minimizer, we have

$$(|x| - |y|)^2 = (x - y)^2 = x^2 + y^2 - 2xy = x^2 + y^2 - 2,$$

i.e., $(x - y)^2$ is equivalent to the sharpness $x^2 + y^2$ up to a constant shift. Therefore, if GD starts near a minimum, we can compare the limiting difference $(x_\infty - y_\infty)^2$ with its initial $(x_0 - y_0)^2$ to gain an understanding of whether GD converges to a sharper or a flatter minimum.

In this section, we generalize the balancing phenomenon scrutinized in (Wang et al., 2022a) to the larger class of functions given by (2) (when dor ≤ 2), and prove that bad regularity can eliminate balancing.

The following theorem shows the balancing phenomenon for functions with good regularities.

Theorem 9 (Balancing) *Consider $0 < a \le 1$ in (2). Assume the initial condition satisfies $(x_0, y_0) \in \{(x, y) : 1 < xy < M_1(a), x^2 + y^2 \gtrsim 4c_1^{-4/3}\} \backslash \mathcal{B}_a$, for some $c_1 = c_1(a) > 0$ and $M_1(a) > 3$, where $\mathcal{B}_a$ is a Lebesgue measure-0 set. Let the learning rate be $h = \frac{C}{x_0^2+y_0^2}$ for $2 < M_2(a) \le C \le 4$, where $M_2(a) \lesssim 2.6$, where the precise definitions of M_1, M_2, c_1 are in Theorem 13. Then GD converges to a global minimum (x_∞, y_∞), and we have*

$$(x_\infty - y_\infty)^2 \le \frac{2}{C}(x_0 - y_0)^2 + 2(x_0 y_0 - 1).$$

The above theorem shows that given a large h corresponding to $C > 2$, if $x_0 y_0$ is close to 1, $(x_\infty - y_\infty)^2 < (x_0 - y_0)^2$ and the balancing phenomenon occurs. Additionally, a larger h leads to a smaller upper bound of $(x_\infty - y_\infty)^2$, and therefore is more balanced at the limit. Especially, if $C = 4$, we have

$$(x_\infty - y_\infty)^2 \le \frac{1}{2}(x_0 - y_0)^2 + 2(x_0 y_0 - 1) \lesssim \frac{1}{2}(x_0 - y_0)^2. \tag{9}$$

The case with $b = 1$ was already studied in the literature, and it also exhibits the balancing phenomenon as reviewed below:

Theorem 10 (Balancing; Corollary of Theorem 3.2 (Wang et al., 2022a)) *Let $b = 1$ in (2). Assume the initial condition satisfies $(x_0, y_0) \in \left\{(x, y) : x^2 + y^2 > 8\right\} \backslash \mathcal{B}$, where $\mathcal{B}$ is some Lebesgue measure-0 set. Let the learning rate be $h = \frac{C}{x_0^2+y_0^2+4}$, where $2 \le C \le 4$. Then GD converges to a global minimum (x_∞, y_∞), and we have*

$$(x_\infty - y_\infty)^2 \le \frac{2}{C}(x_0 - y_0)^2 + \frac{4}{C}(x_0 y_0 + 2) - 2.$$

In the above theorem, the range of h is approximately the same as Theorem 9 when $x_0^2+y_0^2$ is large, and consistent with Theorem 11. Especially, if $C=4$, we have

$$(x_\infty-y_\infty)^2\le\frac{1}{2}(x_0-y_0)^2+x_0y_0,$$

which has a similar upper bound as (9) when x_0y_0 is small.

More important than proving balancing for more functions, the following theorem demonstrates that the balancing is lost when the regularity of the objective becomes bad.

Theorem 11 (no Balancing) *Consider $b=2n+1$ for $n\in\mathbb{Z}$ in (2). Assume the initial condition satisfies $(x_0,y_0)\in\{(x,y):xy>2^{\frac{1}{b-1}},x^2+y^2\ge 4\}\backslash\mathcal{B}_b$, where $\mathcal{B}_b$ is a Lebesgue measure-0 set. Let the learning rate be $h=\frac{C}{(x_0^2+y_0^2+4)(x_0y_0)^{2b-2}}$ for $2\le C\le M_3(b,x_0,y_0)$, where $3<M_3(b,x_0,y_0)\le 4$, and the precise definition of M_3 is given in Theorem 15. Then GD converges to a global minimum (x_∞,y_∞), and we have*

$$(x_\infty-y_\infty)^2\ge(x_0-y_0)^2+\min\left\{2(x_0y_0-1)-\frac{2C}{b}x_0y_0,\left(2-\frac{8b}{4b-C}\right)(x_0y_0-1)\right\}$$

The above theorem states that although a larger learning rate (i.e., larger C) can still reduce the lower bound of $(x_\infty-y_\infty)^2$, this limiting difference may not be smaller than the initial difference $(x_0-y_0)^2$, especially when b is large, i.e., there is no balancing in this case. Note this does not contradict the stability theory (see Appendix A), despite the fact that GD does not converge to a flatter minimum (see more discussions below). In addition, similar to the discussion in Section 4 on EoS, a larger learning rate does not help recover the balancing in bad regularity cases either (see Figure 9). A more detailed version of Theorem 11 can be found in Theorem 30.

★ **Example implications of the general theory.** Consider the family of functions in (2), i.e., with both good and bad regularities.

- *Bad regularities eliminate the balancing phenomenon.* Consider the same set of initial conditions (7) as in Section 4.

  When $\mathrm{dor}(F)\le 2$, we have $(x_\infty-y_\infty)^2\le\frac{2}{C}(x_0-y_0)^2+\tilde{c}$,
  where $2.6\lesssim C=C(h)\le 4$, $\tilde{c}$ is a constant independent of h,x_0,y_0,a.

  When $\mathrm{dor}(F)>2$, we have $(x_\infty-y_\infty)^2\ge(x_0-y_0)^2-\hat{c}\frac{C}{b}\to(x_0-y_0)^2$ as $b\to\infty$,
  where $2\le C=C(h)\le M_3(b,x_0,y_0)\le 4$, $\hat{c}$ is independent of h,x_0,y_0,b.

  This shows that if the initial condition is near a sharp minimum, i.e., $(x_0-y_0)^2$ is large, then for functions with good (small) regularities, $(x_\infty-y_\infty)^2\lesssim(x_0-y_0)^2$, while for functions with bad (large) regularities, $(x_\infty-y_\infty)^2\gtrsim(x_0-y_0)^2$.

  Similarly to Section 4, a worse regularity (larger degree) also results in a larger set of initial conditions that do not lead to balancing.

- *When balancing appears, larger learning rate leads to more balancedness.* Note the above $C \approx h(x_0^2 + y_0^2)$[6]. Then, for dor ≤ 2, we equivalently have

$$(x_\infty - y_\infty)^2 \lesssim \frac{2(x_0 - y_0)^2/(x_0^2 + y_0^2)}{h} + \bar{c}, \qquad \text{where } \bar{c} \text{ is independent of } h, x_0, y_0, a.$$

  Namely, for any fixed initial condition, large h reduces the gap between x_∞ and y_∞, i.e., brings more balancedness.

- *Large learning rate does not necessarily help GD converge to a flatter minimum.* For bad regularity cases, $(x_\infty - y_\infty)^2$ does not decrease compared to its initial $(x_0 - y_0)^2$. This implies that if starting near a sharp minimum (but not too close; see Section 8), GD will just converge to it instead of searching for flatter ones. Note this does not contradict our claim that large learning rate drives GD towards flatter regions. For functions with bad regularity, the minimizer large-learning-rate GD converges to is, roughly speaking, already flatter than its neighborhood (see Figure 5). As the degree of regularity increases, the difference in sharpness between this sharp minimum and its neighborhood also increases, i.e., the minimum becomes much flatter than its neighborhood. Thus converging to this sharp minimum is already a reflection of GD escaping to a flatter region.

Remark 12 (EoS vs Balancing) *Note that EoS and balancing are different phenomena despite both being large learning rate behaviors of GD. More precisely, balancing mainly stems from the* ***de-sharpening (pre-EoS)*** *stage in EoS but is almost unrelated to the* ***progressive sharpening*** *and* ***limiting stabilization of sharpness near*** $2/h$*. In contrast, the* ***pre-EoS*** *is not a prerequisite for EoS to manifest; instead, EoS relies on the latter two stages.*

6. Convergence

The conditions required by theorems in both Section 4 and 5 are to ensure the convergence of GD even with a large learning rate. The proofs of those theorems rely on a detailed understanding of this convergence. It is essential to emphasize that achieving convergence under large learning rates is a non-trivial task. Our results differ from existing milestones that demonstrate certain implicit biases based on unverified assumptions of convergence. Next, we will provide precise definitions of the settings used for the convergence results.

For the convenience of stating our convergence results, let us first begin with some additional **notations** and **definitions**:

Consider GD update for the objective function $f(x, y) \coloneqq F(xy)$

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ y_k \end{pmatrix} - h\ell_k \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_k \\ y_k \end{pmatrix} = \begin{pmatrix} 1 & -h\ell_k \\ -h\ell_k & 1 \end{pmatrix} \begin{pmatrix} x_k \\ y_k \end{pmatrix}, \tag{10}$$

where $\ell_k = F'(x_k y_k)$.

6. Indeed, $C = h(x_0^2 + y_0^2)$ for dor ≤ 1; $C = h(x_0^2 + y_0^2 + 4)$ for dor $= 2$

By GD update (10), we have

$$x_{k+1}y_{k+1} - 1 = (x_k y_k - 1)\left(1 - h\frac{\ell_k}{x_k y_k - 1}(x_k^2 + y_k^2 - h\ell_k x_k y_k)\right) = r_k(x_k y_k - 1).$$

where $r_k = 1 - h\frac{\ell_k}{x_k y_k - 1}(x_k^2 + y_k^2 - h\ell_k x_k y_k)$. Let $\delta := xy - 1$. We define

$$q(\delta) := \frac{F'(xy)}{xy - 1} = \frac{F'(\delta + 1)}{\delta}, \text{ and then } q_k := q(x_k y_k - 1) = \frac{\ell_k}{x_k y_k - 1}. \tag{11}$$

Moreover, for $0 < a \le 1$,

$$1 - c_1\delta^2 \le q(\delta) \le 1, \text{ where } c_1 = -\frac{-3 + 3a - 2\log(2)}{24\log(2)}. \tag{12}$$

In the next three theorems, we provide the detailed convergence results of GD for our family of functions under large learning rates, along with precise definitions of M_1, M_2, M_3, and c_1. These theorems are the foundation of all the earlier results on EoS and balancing. Note GD still converges when h is smaller, but that is not our focus here. See more discussions in the last paragraph of this section.

Theorem 13 *Consider $0 < a \le 1$ in (2). Assume the initial condition*

$$(x_0, y_0) \in \{(x, y) : 1 < xy \lesssim M_1, x^2 + y^2 \gtrsim 4c_1^{-4/3}\}\backslash\mathcal{B}_a$$

for some Lebesgue measure-0 set $\mathcal{B}_a$, where c_1 is defined in (12), $M_1 = q^{-1}\left(\frac{q(1)}{2}\right) + 1$. Let the learning rate be

$$\frac{2}{x_0^2 + y_0^2} < \frac{M_2}{x_0^2 + y_0^2} \le h \le \frac{4}{x_0^2 + y_0^2}, \text{ where } M_2 = \frac{2}{q(1)} + \mathcal{O}\left(\frac{1}{x_0^2 + y_0^2}\right).$$

Then GD converges to a global minimum.

Theorem 14 (Theorem 3.1 in Wang et al. (2022a)) *Consider $b = 1$ in (2). Assume the initial condition*

$$(x_0, y_0) \in \{(x, y) : x^2 + y^2 \ge 8\}\backslash\mathcal{B}$$

for some Lebesgue measure-0 set $\mathcal{B}$.Let the learning rate be

$$\frac{2}{x_0^2 + y_0^2 + 4} \le h \le \frac{4}{x_0^2 + y_0^2 + 4}.$$

Then GD converges to a global minimum.

Theorem 15 *Consider $b = 2n + 1$ for $n \in \mathbb{Z}$ in (2). Assume the initial condition*

$$(x_0, y_0) \in \{(x, y) : xy > 2^{\frac{1}{b-1}}, x^2 + y^2 \ge 4\}\backslash\mathcal{B}_b, \text{ for some Lebesgue measure-0 set } \mathcal{B}_b.$$

Let the learning rate be

$$\frac{2}{(x_0^2 + y_0^2 + 4)(x_0 y_0)^{2b-2}} \le h \le \frac{M_3}{(x_0^2 + y_0^2 + 4)(x_0 y_0)^{2b-2}},$$

	tanh	ReLU	ReLU^3
huber	0	1	3
ℓ^2	0	2	6

Table 1: Degrees of regularities (dor) for different compositions of loss and activation functions.

where $M_3 = \min\left\{4, \arg\max_C \left\{\left(1 + \left(\frac{C}{(x_0^2+y_0^2+4)(x_0y_0)^{2b-2}}\right)^2 \ell_0^2\right)x_0y_0 - \frac{C}{(x_0^2+y_0^2+4)(x_0y_0)^{2b-2}}\ell_0(x_0^2 + y_0^2) \geq \epsilon\right\}\right\} > 3$, *for some small* $\epsilon > 0$. *Then GD converges to a global minimum. Moreover, we have the rate of convergence*

$$|x_k y_k - 1| \leq |x_1 y_1 - 1| S^{k-1}, \quad \textit{for } k \geq 1,$$

$$\textit{where } S = \begin{cases} 1 - h(2 - h\ell_0), & \textit{if } r_0 > 0 \\ 1 - h(2-h)q_1 x_1 y_1, & \textit{otherwise} \end{cases}, \textit{ and } 0 < S < 1.$$

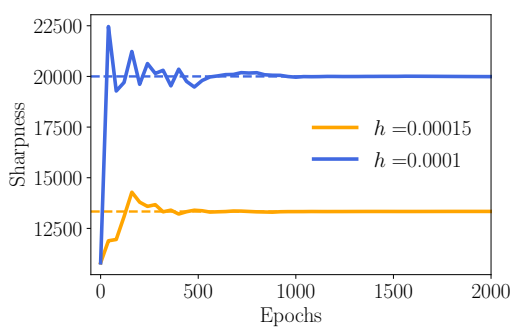
Note the convergence rate quantified by Theorem 15 is a global one. This is the first characterization of the convergence speed of GD under large learning rates for non-convex functions (2) ($b \geq 3$) that do not even have globally Lipschitz gradients (note the quantification of convergence speed of GD for nonconvex functions without Lipschitz gradient is already highly nontrivial, even for regular learning rates; examples of great work in that direction include Zhang et al. (2019); Ward and Kolda (2023); Li et al. (2023)).

The main idea of the proof is to construct a quantitative version of the two-phase convergence pattern of large learning rate GD as discussed in Section 4. That is, GD first tries to escape from the sharp region to a flat region (*Phase 1*) and then converges inside the flat one (*Phase 2*). For more detailed discussions of this two-phase convergence and its interpretation under large learning rates, please see Section 4.

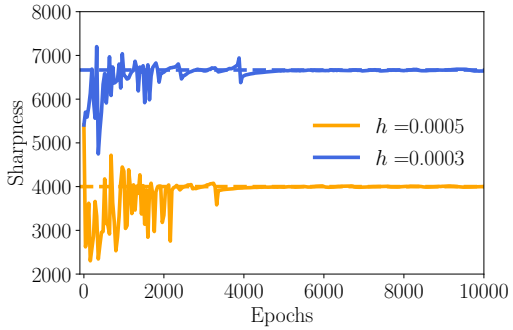
The convergence of GD also holds in the regular learning rate regime (we omit it here since it is not our main focus). The analysis is more standard although still non-trivial for non-convex functions. The convergence pattern for regular learning rate contains only *Phase 2* and GD will converge to a nearby minimum.

7. Experiments

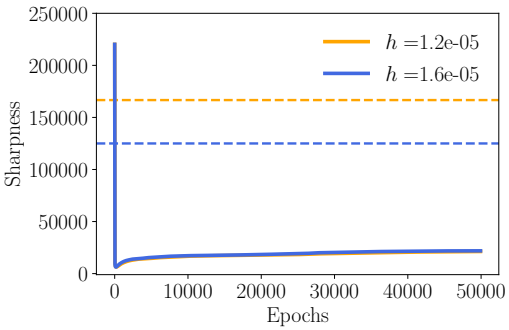
This section is dedicated to demonstrating that our results still hold in more complicated setups corresponding to deep learning problems, and regularity is an essential factor determining large learning rate implicit biases. We also verify that these biases only occur when learning rates are sufficiently *large*. Moreover, batch normalization, a popular technique for improving training and generalization, indeed works and can turn the neural network training objective that has bad regularity into one with good regularity, creating pleasant large learning rate implicit biases. Detailed setup and more experiments can be found in Appendix G.



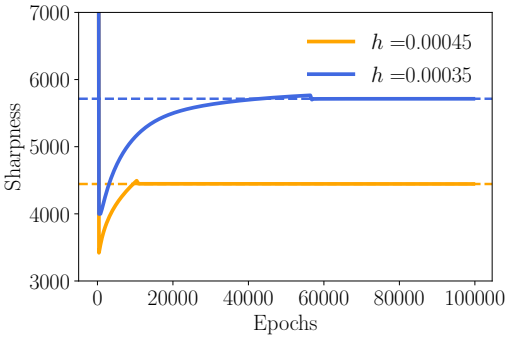
(a) huber+tanh → EoS



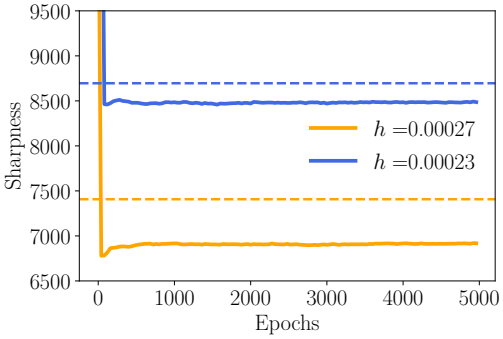
(b) huber+ReLU → EoS



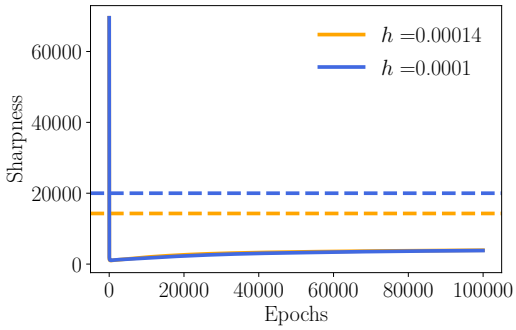
(c) huber+ReLU^3 → no EoS



(d) ℓ^2+tanh → EoS

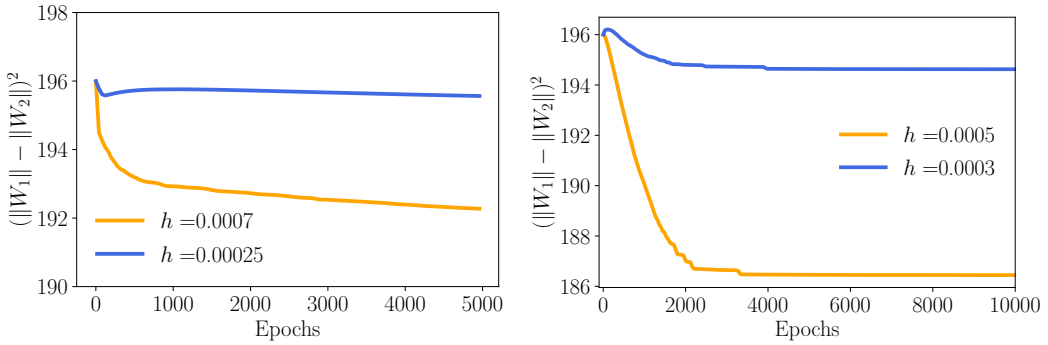


(e) ℓ^2+ReLU → no EoS



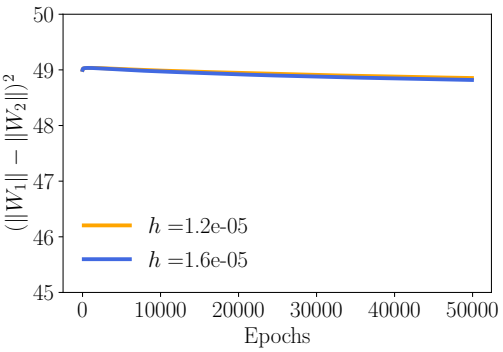
(f) ℓ^2+ReLU^3 → no EoS

Figure 11: EoS and non-EoS phenomena. The solid lines are the evolution of sharpness in different situations; the dashed lines are $2/h$ for different learning rates h.

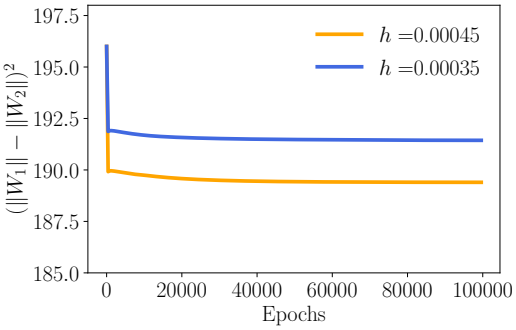


(a) huber+tanh → balancing

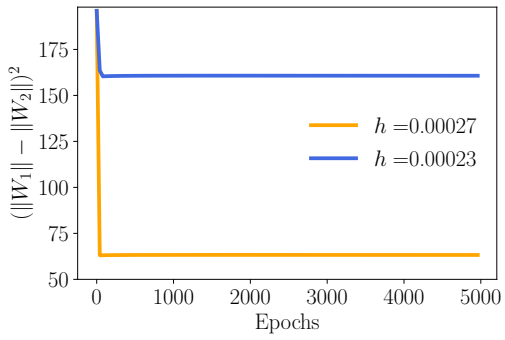
(b) huber+ReLU → balancing



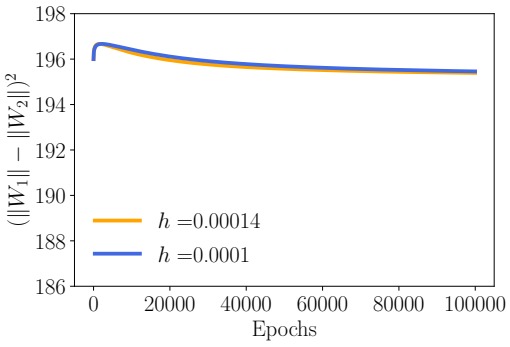
(c) huber+ReLU^3 → no balancing



(d) ℓ^2+tanh → balancing



(e) ℓ^2+ReLU → balancing



(f) ℓ^2+ReLU^3 → no balancing

Figure 12: Balancing and non-balancing. For (a) huber+tanh, we use larger learning rates than that in Figure 11 to further highlight the balancing phenomenon; all the other cases use the same learning rates as in Figure 11.

★ **Good regularity vs bad regularity.** In this section we focus on 6 different objective functions, whose degrees of regularities are computed under (5) for objective function (4) with various loss (huber and ℓ^2)and activation (tanh, ReLU, and ReLU^3) functions (see more details in Section 3.2) in Tab. 1. Although the neural network models used in the experiments are mostly 2-layer which is different from (4) (except for huber+ReLU^3, where we use the same structure as Equation 4 to highlight the effect of regularity; see details in Appendix G), Tab. 1 still helps illustrate the insight of regularity affecting on large learning rate phenomena in different models. Figure 11 and 12 shows the evolution of sharpness and $(\|W_1\|_{\mathrm{F}} - \|W_2\|_{\mathrm{F}})^2$ along GD iterations respectively under the six cases of Tab. 1. When dor > 1, there is no EoS (see Figure 11(e) 11(c) 11(f)); when dor > 2, there is no balancing (see Figure 12(c) 12(f)). This validates our conjecture that in neural network models, bad regularity can also eliminate large learning rate phenomena.

★ **Restoration of EoS: batch normalization.** In Figure 13, we add batch normalization to the three bad regularity cases that do not lead to EoS (see Figure 11(e) 11(c) 11(f)). It is shown in the figures that EoS reappears for each case, which shows that batch normalization turns bad regularity into good one and consequently lead to large learning rate phenomena.

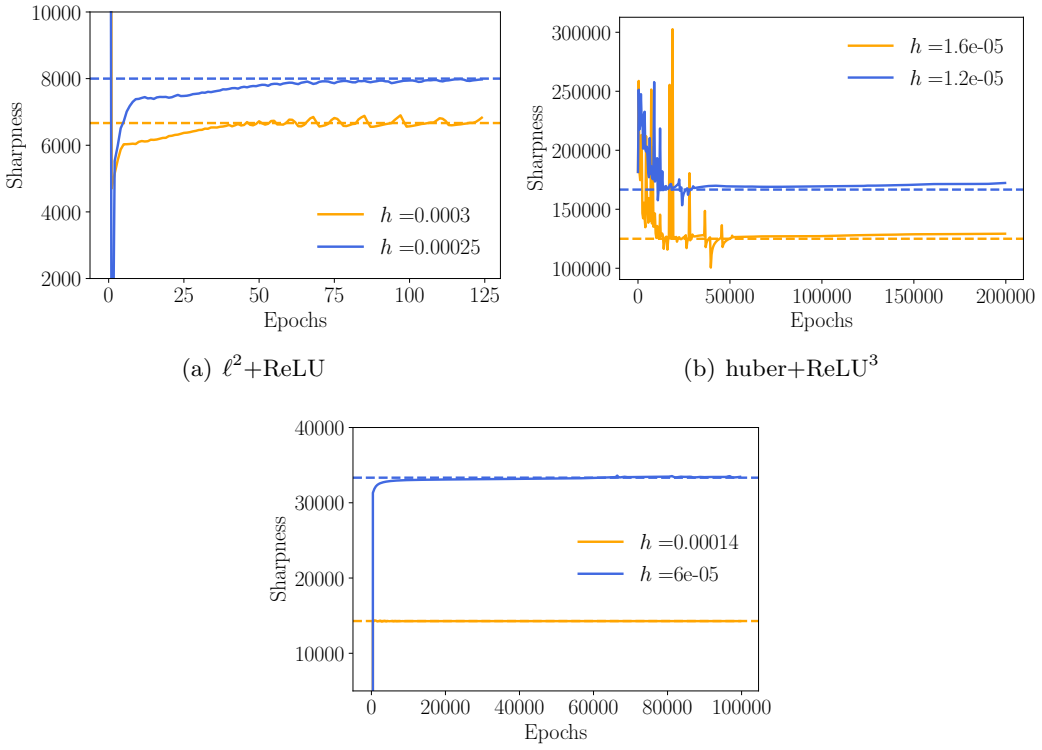
★ **Large learning rate vs small learning rate.** Figure 14 shows two examples where we use both small and large learning rates for optimization. It turns out that a small learning rate can indeed vanish the large learning rate phenomena.

★ **Independence of data sets: CIFAR-10 and MNIST.** All the results above are performed on CIFAR-10. In Figure 15, we also show results on MNIST, which are consistent with those on CIFAR-10. See more results in Appendix G.

8. Discussions

★ **Large learning rate phenomena under bad regularity.** Our findings indicate that, for a substantial set of initial conditions and within our family of functions, the occurrence of large learning rate phenomena becomes less likely as the regularity of objective function worsens (i.e., a larger degree of regularity). In the meantime, a small portion of 'good' regions of initial conditions may still exist, for which the large learning rate phenomena (partially) persist even for functions with bad regularities. For example, when the initial condition is very close to the minima, Figure 16 shows that balancing and the limiting behavior (sharpness near $2/h$) of EoS still occur, although there is no progressive sharpening stage. A similar limiting behavior of EoS is theoretically described in Zhu et al. (2023) with $b = 2$. Figure 17 shows that when we increase the degree of regularity, large learning rate phenomena will still disappear even if we use the same initialization as Figure 16.

★ **Beyond our function class.** Within our function class (2), regularity is a sufficient and necessary condition for the implicit biases of large learning rate discussed here; however, beyond these functions, more elements can contribute to these phenomena. One of them is symmetry. For example, in (2), if we consider the asymmetrical version $\log(\exp(s-1)+1)$, the minimizer is at $-\infty$ with 0 Hessian. This means there is no 'climb up the hill' for sharpness towards $2/h$ and thus no EoS. Additional factors may also affect these phenomena, such as the involvement of bias in the neural network, the usage of multiple data points, and more complicated network architectures. We will leave these for future explorations.



(a) ℓ^2+ReLU

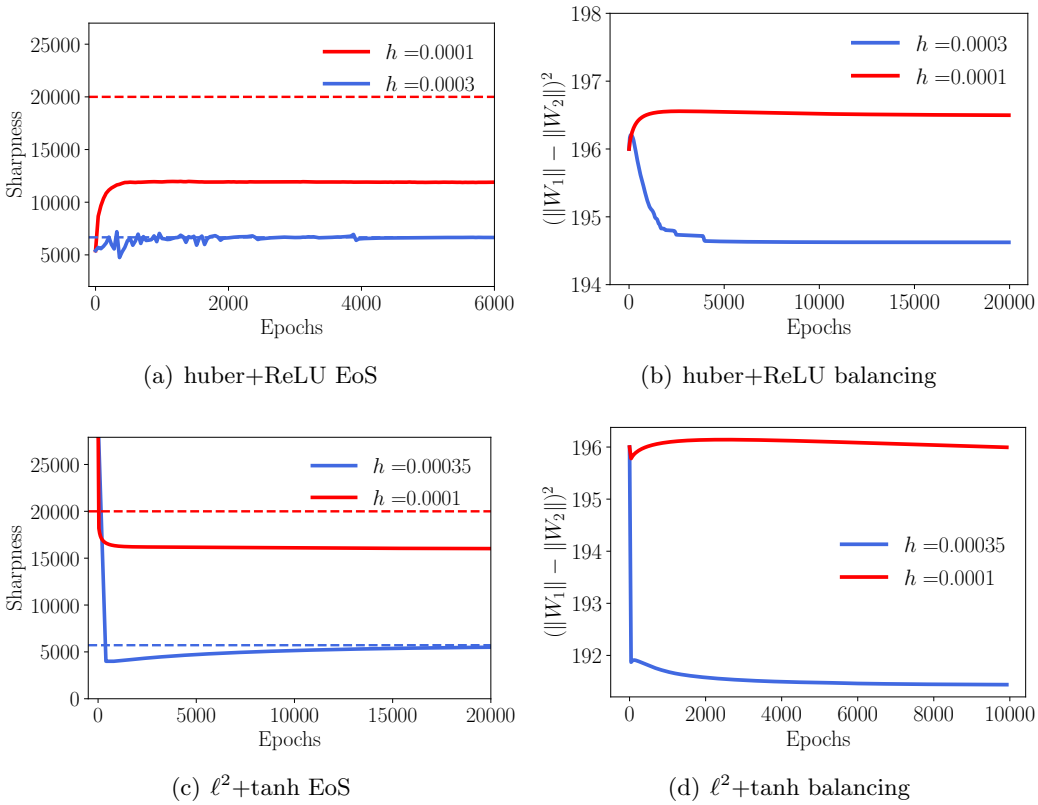
(b) huber+ReLU^3

(c) ℓ^2+ReLU^3

Figure 13: Restoration of EoS for bad regularity cases under batch normalization. The three cases above correspond to the same setting as Figure 11(e) 11(c) 11(f) with batch normalization. Without batch normalization, EoS does not occur.

★ **The effect of depth.** Our framework currently lacks the analysis of the effect of depth, due to the fact that it is not a multivariate theory. Extending the current techniques and/or results to multivariate cases is highly nontrivial. Besides, there might be additional consequences or phenomena of increasing depth, for example, more complicated balancing between subsets of layers compared to the two-layer case, where balancing can only occur between the two layers. We will also need to leave this for future exploration.

★ **Comparison between our analysis techniques and existing ones.** Damian et al. (2022) and Wang et al. (2022b) analyzed the generic objective functions and therefore required extra assumptions involving information along the GD trajectory. These assumptions cannot be easily verified theoretically. In contrast, our work follows Wang et al. (2022a), Zhu et al. (2023), Ahn et al. (2023), and Song and Yun (2023) by constructing explicit example functions (see detailed discussions in Section 3.1). Specifically, inspired by Wang et al. (2022a), we deal with more complicated objective functions that are structurally close to how neural network composes activation and loss functions. Our theory is enabled by



(a) huber+ReLU EoS

(b) huber+ReLU balancing

(c) ℓ^2+tanh EoS

(d) ℓ^2+tanh balancing

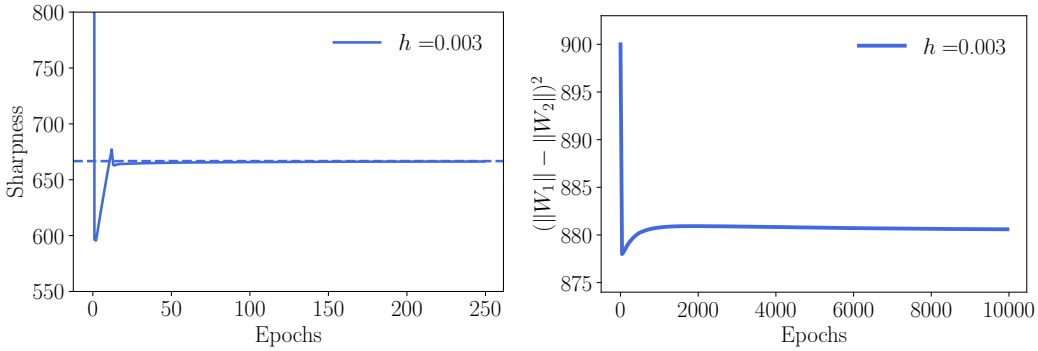
Figure 14: Small learning rate (red) vs large learning rate (blue). All the small learning rate cases (red) fail to exhibit EoS and balancing even under good regularity functions.

a new analysis technique, namely scrutinizing the co-evolution of two auxiliary quantities[7] x^2+y^2 and xy through the entire GD process, which helps us provide a first characterization of the transition between the first two stages of EoS.

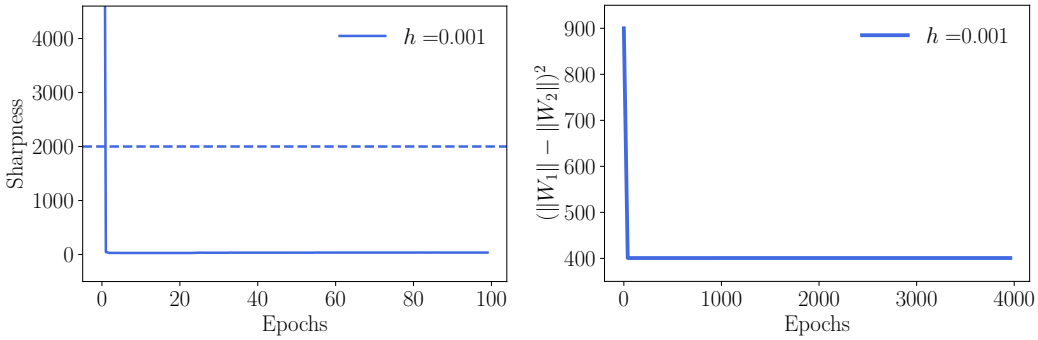
Acknowledgment

YW and MT were partially supported by NSF DMS-1847802, NSF ECCS-1942523, Cullen-Peck Scholarship, and GT-Emory AI Humanity Award, and MT also recently supported by NSF DMS-2513699, DOE NA0004261, SC0026274, and Richard Duke Fellowship. The authors thank Sinho Chewi, Yuejie Chi, Jeremy Cohen, Rong Ge, Francesco Orabona, Suvrit Sra for inspiring discussions, and John Zhang for proofreading the paper.

7. Note here both x^2+y^2 and xy are generic expressions that work for various functions, as the former is the sharpness at minimizer, and the later is the minimizer; all our objective functions are designed such that these two expressions remain the same so that a comparison across different functions is fair.



(a) huber+tanh → EoS (b) huber+tanh → balancing



(c) ℓ^2+ReLU → no EoS (d) ℓ^2+ReLU → no balancing

Figure 15: Large learning rate phenomena on MNIST. The above plots show that the effect of regularity of functions is not unique to CIFAR-10: under MNIST data set, good regularity function leads to EoS (a) and balancing (b), while bad regularity function fail to ensure EoS (c) or balancing (d).

Appendix

Appendix A. Necessary condition of convergence from stability theory

For an objective function $f(u)$, consider the GD update in terms of a iterative map ψ

$$u_{k+1} = \psi(u_k) := u_k - h\nabla f(u_k).$$

We further consider a stationary point u^* of the objective function f, i.e., $\nabla f(u^*) = 0$. Then this point u^* is a fixed point of the map ψ since

$$u^* = u^* - h\nabla f(u^*) = \psi(u^*).$$

If all the magnitudes of the eigenvalues of Jacobian matrix $\nabla\psi(u^*)$ are less than 1, u^* is a stable fixed point (see more explanation in Section 5 of Wang et al. (2022a)). Consequently, we have the following theorem

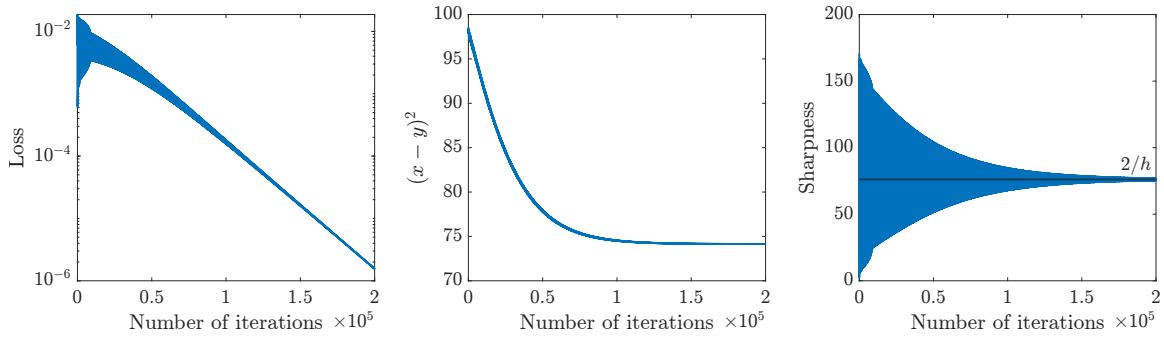


Figure 16: Appearance of large learning rate phenomena in 'good' region of functions with bad regularities. The objective function is (2) with $b = 3$; the initial condition is $x = 10, y = 0.11$; the learning rate is $h = \frac{4}{(x_0^2+y_0^2+4)(x_0y_0)^{2b-2}}$.

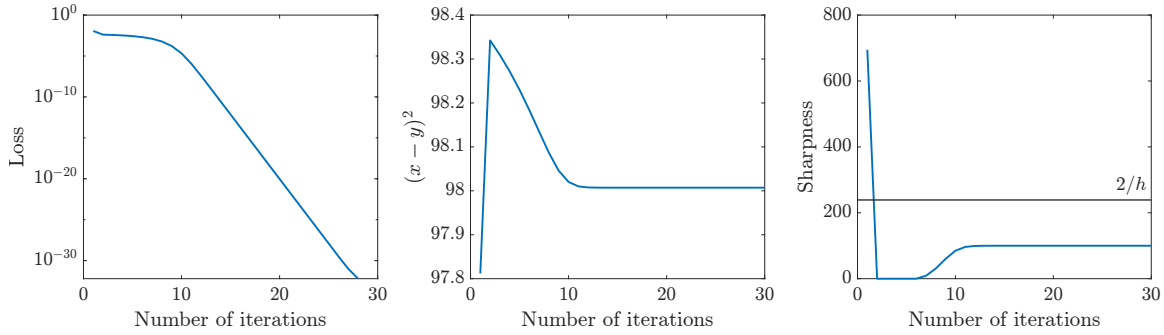


Figure 17: Disappearance of large learning rate phenomena in 'good' region of functions with bad regularities. This Figure uses the same initialization and rule of choosing learning rate with $b = 9$.

Theorem 16 (Necessary condition, see for example Wu et al.) *Let u^* be a local minimum point of $f(u)$ and consider GD updates. If $-I \prec I - h\nabla^2 f(u^*) \prec I$, i.e., $h < \frac{2}{L^*}$, where $\nabla^2 f(u^*) \preceq L^* I$, we have that u^* is a stable fixed point of GD map.*

Note the above theorem is a necessary condition of the convergence of GD to a minimizer. This is due to the fact that if $h > \frac{2}{L^*}$, there exists at least one eigendirection s.t. the magnitude of its eigenvalue is greater than 1. Namely, there will be an unstable direction of the map that prevents GD from converging towards the point u^*.

Appendix B. Degree of regularity

In this section, we would like to prove (5).

Define

$$\mathrm{dor}_+(F) = \inf\left\{n : |F(s)| \le C_1|s|^n, \text{ for } s \ge C_0 \text{ with some constant } C_0, C_1 > 0\right\}$$
$$\mathrm{dor}_-(F) = \inf\left\{n : |F(s)| \le C_1|s|^n, \text{ for } s \le -C_0 \text{ with some constant } C_0, C_1 > 0\right\}.$$

Theorem 17 *Let $F(s), G(s)$ be continuous functions in $\mathbb{R}$. Assume $\mathrm{dor}_+(F) = \mathrm{dor}_-(F)$.*

1. *Assume $|F(s)| \to \infty$ as $s \to \infty$. Assume there exists at least one subsequence $\{s_i\}$, s.t., $|G(s_i)| \to \infty$ as $s_i \to \infty$. Then*

$$\mathrm{dor}(F \circ G) = \mathrm{dor}(F)\mathrm{dor}(G)$$

2. *Assume $|G(s)| \le M$ for some constant $0 < M < +\infty$. Then*

$$\mathrm{dor}(F \circ G) = 0 = \mathrm{dor}(F)\mathrm{dor}(G)$$

Proof For all the three cases, there exists some $C_1, C_2, n_F, n_G \ge 0$ for $|s| \ge C_0$, s.t.,

$$|F(G(s))| \le C_1|G(s)|^{n_F} \le C_1(C_2|s|^{n_G})^{n_F} = C_3|s|^{n_G n_F},$$

where $C_3 = C_1 C_2^{n_F} \ge 0$.

For 1, from the conditions, we know there exists at least one subsequence $\{s_i\}$, s.t., $F(G(s_i)) \to +\infty$ as $s_i \to +\infty$. Then

$$\begin{aligned}|F(G(s))| &\le \inf_{n_F} C_1|G(s)|^{n_F} = C_1|G(s)|^{\mathrm{dor}(F)} \\ &\le \inf_{n_G} C_1(C_2|s|^{n_G})^{\mathrm{dor}(F)} = C_3|s|^{\mathrm{dor}(G)\mathrm{dor}(F)},\end{aligned}$$

where the first inequality follows from $F(G(s))$ is unbounded and the definition of dor for F; the second inequality follows from the definition of dor for G. Thus by definition,

$$\mathrm{dor}(F \circ G) = \mathrm{dor}(F)\mathrm{dor}(G).$$

For 2, since $|G(s)| \le M$, we have $\mathrm{dor}(G) = 0$ by definition. Then, since $F(s)$ is continuous, we have $|F(G(s))|$ is bounded, i.e., $\mathrm{dor}(F \circ G) = 0$. Therefore, we also have

$$\mathrm{dor}(F \circ G) = \mathrm{dor}(F)\mathrm{dor}(G).$$

■

Appendix C. Preparation for proofs

Before the proofs, we first take a closer look at the GD iteration for the function $f(x, y) = F(xy)$ (2)

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ y_k \end{pmatrix} - h\ell_k \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_k \\ y_k \end{pmatrix} = \begin{pmatrix} 1 & -h\ell_k \\ -h\ell_k & 1 \end{pmatrix} \begin{pmatrix} x_k \\ y_k \end{pmatrix},$$

where $\ell_k = F'(x_k y_k)$. Let $u_k = \begin{pmatrix} x_k \\ y_k \end{pmatrix}$. Then

$$u_{k+1}^\top u_{k+1} = (1 + h^2 \ell_k^2) u_k^\top u_k - 4h\ell_k x_k y_k, \tag{13}$$

$$\begin{aligned} u_{k+2}^\top u_{k+2} &= (1 + h^2 \ell_{k+1}^2) u_{k+1}^\top u_{k+1} - 4h\ell_{k+1} x_{k+1} y_{k+1} \\ &= (1 + h^2 \ell_{k+1}^2)((1 + h^2 \ell_k^2) u_k^\top u_k - 4h\ell_k x_k y_k) - 4h\ell_{k+1} x_{k+1} y_{k+1}. \end{aligned} \tag{14}$$

Lemma 18 *Under the same assumption as Theorem 13, we have*

$$u_{k+1}^\top u_{k+1} \lesssim u_k^\top u_k - 4h\ell_k (x_k y_k - \ell_k)$$

Proof By Lemma 23, we have

$$u_k^\top u_k \le \frac{4}{h} + \mathcal{O}(h).$$

Then by (13), we have

$$u_{k+1}^\top u_{k+1} \le u_k^\top u_k - 4h\ell_k (x_k y_k - \ell_k) + \mathcal{O}(h^3 \ell_k^2)$$

■

For the update of $x_k y_k$, we have

$$x_{k+1} y_{k+1} = (1 + h^2 \ell_k^2) x_k y_k - h\ell_k u_k^\top u_k, \tag{15}$$

$$x_{k+1} y_{k+1} - 1 = (x_k y_k - 1) \left(1 - h \frac{\ell_k}{x_k y_k - 1} (u_k^\top u_k - h\ell_k x_k y_k)\right). \tag{16}$$

Let $\delta := xy - 1$ and $\delta_k := x_k y_k - 1$. Then we define the following functions:

$$\begin{gathered} \ell(\delta) = F'(\delta + 1), \quad \text{and then } \ell_k = \ell(\delta_k) = F'(x_k y_k); \\ q(\delta) = \frac{\ell(\delta)}{\delta}, \quad \text{and then } q(\delta_k) = \frac{\ell_k}{x_k y_k - 1}; \\ r(u_k, \delta) = 1 - hq(\delta)(u_k^\top u_k - h\ell(\delta)(\delta + 1)), \\ \text{and then } r_k = r(u_k, \delta_k) = 1 - h \frac{\ell_k}{x_k y_k - 1} (u_k^\top u_k - h\ell_k x_k y_k). \end{gathered}$$

Let

$$C_k = \frac{1 - r_k}{q(\delta_k)}, \quad \text{and then } r_k = 1 - C_k q(\delta_k).$$

From Lemma 18, we also define

$$L(\delta) = \ell(\delta)(\delta + 1 - \ell(\delta)), \text{ and then } L(\delta_k) = \ell_k (x_k y_k - \ell_k).$$

All the above functions also depend on a or b. If not specified, all the properties for these functions used in the proofs are valid for all $0 < a \le 1$ or $b = 2n + 1$ with $n \in \mathbb{N}$.

Throughout the proof, we use characters with subscripts (e.g. x_k, y_k, δ_k) to represent the point on the GD trajectory, and use characters without subscripts (e.g. x, y, δ) when we would like to study the expressions as functions depending on these variables.

Appendix D. Proofs of results for functions (2) with $0 < a \leq 1$

Consider function

$$f(x,y) = \frac{(\log(\exp(xy-1)+1) + \log(\exp(1-xy)+1))^a}{a2^{a-2}\log^{a-1}(2)} \text{ with } 0 < a \leq 1.$$

The following propositions show the properties of the functions defined above under this objective. The proof is based on simple analysis and Taylor expansion and is thus omitted. If not specified, these properties are independent of a.

Proposition 19 *The function* $\ell(\delta) = \frac{2^{2-a}\left(e^{\delta}-1\right)\log^{1-a}(2)\left(\log\left(e^{-\delta}+1\right)+\log\left(e^{\delta}+1\right)\right)^{a-1}}{\left(e^{\delta}+1\right)}$ *has the following properties:*

- $\ell(\delta) = -\ell(-\delta)$
- $\ell(\delta) > 0$ *for* $\delta > 0$.
- $|\ell(\delta)| \leq 3$ *for all* $\delta > 0$.

Proposition 20 *The function* $L(\delta) = \ell(\delta)(\delta + 1 - \ell(\delta))$ *has the following properties:*

- $L(\delta)$ *monotonically increases for* $\delta \geq 0$, *and* $L(\delta) \geq L(0)$ *for* $\delta \geq 0$.
- $L(\delta) + L(-\delta) \geq 0$ *for all* δ.
- $L(\delta_1) + L(\delta) \geq 0$ *for* $\delta_1 \geq 1$ *and all* δ.
- $L(\delta) + L(r\delta) \geq 0.8(1+r)\delta$ *for* $\delta > 0$ *and* $-1 < r < 0$.

Proposition 21 *The function* $q(\delta) = \frac{\ell(\delta)}{\delta}$ *has the following properties:*

- $q(\delta)$ *is symmetric with respect to* $\delta = 0$, *i.e.,* $q(\delta) = q(-\delta)$.
- *For* $\delta \geq 0$, $q(\delta)$ *monotonically decreases as* δ *increases, and then* $0 \leq q(\delta) \leq q(0) = 1$.
- *From Taylor expansion, we have*

$$q(\delta) \geq 1 - c_1\delta^2, \text{ where } c_1 = -\frac{-3+3a-2\log(2)}{24\log(2)} > 0.$$

  Also, when $\delta \leq 1$,

$$q(\delta) \leq 1 - c_2\delta^2, \text{ where } c_2 = -\frac{-3+3a-2\log(2)}{30\log(2)}.$$

- *For* $2 \leq C \leq 4$,

$$\delta = q^{-1}\left(\frac{1}{C}\right) \leq (1+a)C.$$

From the above propositions, we will use $\delta_k = |x_k y_k - 1|$ instead of $x_k y_k - 1$ for the rest of the proofs.

D.1 Proof of convergence

Proof [Proof of Theorem 13]

By Proposition 21 and Proposition 19, we have $q(\delta) \geq 0$ and $|\ell(\delta)| \leq 3$. Then

$$r_k = 1 - hq(\delta_k)(u_k^\top u_k - h\ell_k x_k y_k) \leq 1 - hq(\delta_k)(2 - h\ell_k)|x_k y_k| < 1,$$

where the first inequality follows from $u_k^\top u_k = x_k^2 + y_k^2 \geq 2|x_k y_k|$, and the second inequality follows from $|\ell_k| \leq 3$ and $2 - 3h > 0$.

As is discussed in Lemma 24, the initial condition set removes a null set of converging initial conditions within finite steps, i.e., $r_k \neq 0$. By our choice of initial condition and learning rate h,

$$\begin{aligned} r_0 &\leq 1 - \Big(\frac{2/q(1)}{u_0^\top u_0} + \mathcal{O}(h^2)\Big) q\left(q^{-1}\Big(\frac{q(1)}{2}\Big)(u_0^\top u_0 - \mathcal{O}(h))\right) \\ &= 1 - (1 - \mathcal{O}(h^2) + \mathcal{O}(h) - \mathcal{O}(h^3)) = -\mathcal{O}(h) < 0. \end{aligned}$$

By Lemma 25, for all $n \geq k$, if $x_n y_n > 1$, then $x_{n+1} y_{n+1} < 1$ and vice versa. Then we can consider the kth iteration when $x_k y_k > 1$ and we have $x_{k+2n} > 1$ for $n \geq 1$.

By Lemma 29, $|x_k y_k - 1|$ is guaranteed to decrease monotonically in $|x_k y_k - 1| \leq R_2(a)$, with $|r_k| < 1$. Therefore, GD will converge to $xy = 1$ (otherwise, if $x_k y_k$ converges to $c \neq 1$, with $|r_k| < 1$, $|x_k y_k - 1|$ will keep decreasing, contradiction). ■

D.2 Proofs of EoS

Proof [Proof of Theorem 6]

By Theorem 13, GD converges to a global minimum. Throughout this proof, we will use the big $\mathcal{O}$ notation for complexity and distance. The proof can be made more rigorous by considering some specific constant scaling of these orders.

(1) Analyze the change of $u_k^\top u_k$ and r_k throughout the trajectory; in the end, bound the limiting sharpness

By the lower bound of h, $u_0^\top u_0 \gtrsim \frac{2/q(1)+\mathcal{O}(h)}{h}$, which implies $|r_0(1)| \geq 1 + \mathcal{O}(h)$. Then we consider the decrease of $u_k^\top u_k$ until GD enters the region $|r_k| < 1$. More precisely, we consider three regions: 1) when $|\delta_k|$ is small enough s.t. $u_k^\top u_k$ does not decrease every two steps. By the following bound,

$$\begin{aligned} u_{k+2}^\top u_{k+2} &= (1 + h^2\ell_{k+1}^2) u_{k+1}^\top u_{k+1} - 4h\ell_{k+1} x_{k+1} y_{k+1} \\ &= (1 + h^2\ell_{k+1}^2)((1 + h^2\ell_k^2) u_k^\top u_k - 4h\ell_k x_k y_k) - 4h\ell_{k+1} x_{k+1} y_{k+1} \\ &\geq u_k^\top u_k - 4h[\ell_k x_k y_k - \frac{1}{2}\ell_k^2 + \ell_{k+1} x_{k+1} y_{k+1} - \frac{1}{2}\ell_{k+1}^2] - 4h^3\ell_{k+1}^2 \ell_k x_k y_k + h^4 \ell_{k+1}^2 \ell_k^2 u_k^\top u_k \\ &\geq u_k^\top u_k - 4h[(1 + r_k)(x_k y_k - 1) + \frac{1}{2}(1 + r_k^2)(x_k y_k - 1)^2] \\ &\quad - 4h^3\ell_{k+1}^2 \ell_k x_k y_k + h^4 \ell_{k+1}^2 \ell_k^2 u_k^\top u_k, \end{aligned}$$

we have that such $|\delta_k| \leq \mathcal{O}(h)$. 2) Starting from $|\delta_k| = \mathcal{O}(h)$, consider $|\delta_k|$ increases to some region that is $\mathcal{O}(h)$ away from 1, i.e., the rate of increase is at least $1 + \mathcal{O}(h)$. In this region,

it takes GD at most $\mathcal{O}(-\log h/h)$ steps and $u_k^\top u_k$ may decrease every two steps, where the order of decrease is $\mathcal{O}(h)$. 3) Starting from the end of 2), if the rate of increase is at least $1+\mathcal{O}(h)$, then the complexity of entering the region $|r_k|<1$ is at most $\mathcal{O}(1/h)$ which follows the same derivation as 2). Otherwise, i.e., the rate of increase is less than $1+\mathcal{O}(h)$. Since the decrease of $u_k^\top u_k$ is $\mathcal{O}(h)$ and $q(\delta_k)$ keep decreasing before $|r_k|<1$, it takes GD at most $\mathcal{O}(1/h)$ steps such that $|r_k|$ is $\mathcal{O}(h)$ less than $1+\mathcal{O}(h)$, i.e., $|r_k|<1$. Therefore, the overall decrease of $u_k^\top u_k$ is at most $\tilde{\mathcal{O}}(1)$ and we still have $u_k^\top u_k \geq \frac{2/q(1)-\mathcal{O}(h)}{h}$ at the end of all the above processes. Also, according to Lemma 26, $u_k^\top u_k$ will decrease to at least $\frac{2/q(1)+\mathcal{O}(h)}{h}$. Otherwise, $|r_k|>1$ in $|x_k y_k-1|\geq R_1(a)$ and therefore $u_k^\top u_k$ will keep decreasing.

Before further discussing the complexity of GD, we still need an upper bound of $x_k y_k$. Let $u_k^\top u_k = \frac{C_k}{h}$. Then since $r_k<0$ for all k, we can consider the maximum δ s.t. $r_0\approx 0$, i.e., $\delta = q^{-1}(\frac{1}{C_k}) \leq (1+a)C_k$ by Proposition 21. Therefore, $\delta_k \lesssim (1+a)C_k \leq 5$. Here we just relax C_k to be upper bounded by 2.5 based on $2/q(1)+\mathcal{O}(h)$.

If $|x_k y_k-1|>R_2(a)$, suppose $|x_k y_k-1|>1.5$ then $|r_k|<r_k(1.5)<0.8+\mathcal{O}(h^2)<0.85$. Then the complexity of entering $\delta_k\leq 1.5$ is $\frac{\log(1.5/5)}{\log(0.85)}=\mathcal{O}(1)$. Therefore, the step of GD entering this region is less than the order $\mathcal{O}(1/\log(r_k(1.5)))=\tilde{\mathcal{O}}(1)$.

Then consider the complexity of GD entering $\{r<1, |xy-1|<1-\mathcal{O}(h^2)\}$ (By Lemma 27, $R_2(a)\geq 1-\mathcal{O}(h^2)$). We can consider $|r_k|$ via the map $g(\delta)=\delta(C_k q(\delta)-1)$, where we ignore the $\mathcal{O}(h^2)$ terms and fix $C_k=C_N$ for some N. We would like to analyze the complexity of the map converging to $\mathcal{O}(h)$ error of the fixed point δ^* near $\delta=1$, i.e., $(C_k q(\delta)-1)=1$. Since $(C_k q(1-\mathcal{O}(h^2))-1)=1-\mathcal{O}(h)$,

$$\begin{aligned}1=C_k q(\delta^*)-1 &\geq C_k(1-c_1(\delta^*)^2)-1\\ &= C_k(1-c_1)-1+(1-(\delta^*)^2)C_k c_1\\ &= 1-\mathcal{O}(h)+(1-(\delta^*)^2)C_k c_1\end{aligned}$$

Therefore $\delta^*\geq 1-\mathcal{O}(h/c_1)$.

By checking the derivative of the map, we have $g'(\delta)=C_k q(\delta)-1+C_k\delta q'(\delta)$ and $q'(\delta)<-c_1$ for $0.8<\delta<1.5$. Since $C_k>2$, when $\delta>\delta^*$, $C_k q(\delta)-1<1$ and thus $0<g'(\delta)\leq 1-C_k\delta^* c_1$. Then the map will decrease from above to enter the region within $\mathcal{O}(h)$ distance of δ^* with the complexity of $\mathcal{O}(-\log(h)/c_1)$. For the $\mathcal{O}(h^2)$ terms, the error between this map and the true δ_k update is of the order $\mathcal{O}(\exp(h^2\cdot\log(h)/c_1)-1)<\mathcal{O}(h)$ and thus can be omitted. Then GD starts to decrease in the monotone decreasing region of $|x_k y_k-1|$ with $u_k^\top u_k\geq \frac{2}{h}+\mathcal{O}(1)$.

Next we consider the Nth iteration with $u_N^\top u_N=\frac{2}{h}+\bar{C}_N$, for some $\bar{C}_N=\mathcal{O}(1)$. For $k\geq N$, $u_k^\top u_k>\frac{2}{h}$, and $x_k y_k>1$, we have that $\bar{C}_k-\bar{C}_{k+2}=\mathcal{O}(h)$. We would like to prove that eventually $|r_k|\geq 1-\mathcal{O}(h^2)$ when $u_k^\top u_k\geq\frac{2}{h}$, i.e., we would like to analyze the complexity of $|r_k|$ increasing to $|r_k|=1-\mathcal{O}(h^2)$.

First note

$$|r_k|\geq q(\delta_k)h(u_k^\top u_k-h\ell_k x_k y_k)-1=1-2c_1\delta_k^2+2\bar{C}_N h\pm\mathcal{O}(h^2).$$

Then the complexity is upper bounded by the complexity of δ_k decreasing to the value s.t. $-2c_1\delta_k^2+2\bar{C}_N h=0$, i.e., $\delta_k=\mathcal{O}(\sqrt{h/c_1})$. Let $c_1=h^p$. By the assumption of h, we have $p\leq\frac{3}{4}$.

Therefore $\sqrt{h/c_1} = h^{\frac{1-p}{2}}$. Then we analyze

$$\delta_{n+N} = \delta_N - 2h^p\delta_N^3 + 2\bar{C}_N h\delta_N - 2h^p\delta_N^3(1 - 2h^p\delta_N^2 + 2\bar{C}_N h)^2 + \cdots = \mathcal{O}(h^{\frac{1-p}{2}})$$

We further remove all the $\bar{C}_N$ terms since the sum of them is $\mathcal{O}(h) < \mathcal{O}(h^{\frac{1-p}{2}})$. We first consider

$$\delta_N - 2h^p\delta_N^3 - 2h^p\delta_N^3(1 - 2h^p\delta_N^2)^2 + \cdots = h^{p_1}$$

Ignoring the $\mathcal{O}(h^2)$ part, we have that $|r_k|$ is increasing as $|\delta_k|$ decreases. Then

$$h^{p_1} = LHS \lesssim \delta_N - h^p h^{3p_1} - h^p h^{3p_1} + \cdots = \delta_N - n_1 h^{3p_1+p} \le 1 - n_1 h^{3p_1+p}$$

Then

$$n_1 \le h^{-(3p_1+p)}$$

Iteratively, we consider $\delta_N = h^{p_1}$ and $p_2 = \frac{4}{3}p_1$. Then

$$h^{p_2} = LHS \lesssim h^{p_1} - h^p h^{3p_2} - h^p h^{3p_2} + \cdots = \delta_N - n_2 h^{3p_2+p} \le h^{p_1} - n_2 h^{3p_2+p}$$

$$n_2 \le h^{-(3p_2+p-p_1)} = h^{-(3p_1+p)}$$

We can next solve i s.t.

$$\left(\frac{4}{3}\right)^{i-1} p_1 = \frac{1-p}{2} \Rightarrow i = 1 + \frac{\log \frac{1-p}{2p_1}}{\log \frac{4}{3}}$$

From the above, we also need to consider the $\bar{C}_N$ part and thus require

$$3p_1 + p < \min\{\frac{1+3p}{2}, 1\}$$

Then

$$\frac{1-p}{2p_1} > \frac{3}{2} \text{ and } i \text{ is at most } \mathcal{O}(\log h)$$

Then the total complexity of achieving $h^{\frac{1-p}{2}}$ is $\mathcal{O}(h^{-(3p_1+p)}) < \mathcal{O}(h^{-1})$. Thus we will eventually have $|r_k| = 1 - \mathcal{O}(h^2)$ when $u_k^\top u_k \ge \frac{2}{h}$. For the $k+2$th iteration, the increase of $1 - c\delta_k^2$ is $\mathcal{O}((1-|r_k|)\delta_k^2 + (1-|r_{k+1}|)\delta_k^2) = \mathcal{O}((1-|r_k|)\delta_k^2)$, where one step change of $|r_k|$ is at most $\mathcal{O}(h^2)$ and thus can be omitted, and the decrease of $h(u_k^\top u_k - h\ell_k x_k y_k)$ is $\mathcal{O}(h^2\delta_k(1-|r_k|))$ by Lemma 26. Thus when $\delta_k \ge \mathcal{O}(h^2)$, $\mathcal{O}((1-|r_k|)\delta_k^2) \ge \mathcal{O}(h^2\delta_k(1-|r_k|))$, i.e., $|r_k| \ge 1 - \mathcal{O}(h^2)$; when $\delta_k < \mathcal{O}(h^2)$, from the expression of $|r_k|$, we have $|r_k| \ge 1 - \mathcal{O}(h^2)$ for the kth iteration s.t. $u_k^\top u_k \ge \frac{2}{h} - \mathcal{O}(h)$. Therefore, $|r_k| \ge 1 - \mathcal{O}(h^2)$ for all the kth iteration s.t. $u_k^\top u_k \ge \frac{2}{h} - \mathcal{O}(h)$.

Next, consider the first step when $\frac{2}{h} \ge u_N^\top u_N \ge \frac{2}{h} - \mathcal{O}(h)$ and $x_N y_N > 1$ since the decrease of $u_k^\top u_k$ at each step is at most $\mathcal{O}(h)$. Then based on the expression of r_N, we have $|r_N| \le 1 - \mathcal{O}(h^2)$ and additionally by Lemma 26, $u_k^\top u_k$ decreases every two steps and thus we

have $|r_k| = 1 - \mathcal{O}(h^2)$ for all $k \geq N$. By series expansion, $\ell_{N+k} x_{N+k} y_{N+k} = \mathcal{O}(|x_{N+k} y_{N+k} - 1|)$. Then

$$\begin{aligned}
{u_{N+k+2}}^\top u_{N+k+2} &\geq {u_{N+k}}^\top u_{N+k} - 4h[\ell_{N+k} x_{N+k} y_{N+k} + \ell_{N+k+1} x_{N+k+1} y_{N+k+1}] - \mathcal{O}(h^3) \\
&\geq {u_{N+k}}^\top u_{N+k} - \mathcal{O}\left(h[\ell_{N+k} x_{N+k} y_{N+k} + \ell_{N+k+1} x_{N+k+1} y_{N+k+1}]\right) \\
&\geq {u_{N+k}}^\top u_{N+k} - \mathcal{O}\left(h[(1 - |r_{N+k}|)\delta_{N+k}]\right) \\
&\geq {u_N}^\top u_N - \mathcal{O}\left(h \sum_{i=0}^{k} [(1 - |r_{N+i}|)\delta_{N+i}]\right) \\
&\geq {u_N}^\top u_N - \mathcal{O}\left(h \sum_{i=0}^{k} [(1 - |r_{N+i}|) \prod_{j=0}^{i} r_{N+j} \delta_N]\right) \\
&\geq {u_N}^\top u_N - \mathcal{O}\left(h^3 \delta_N \sum_{i=0}^{k} (1 - \mathcal{O}(h^2))^i\right)
\end{aligned}$$

Take $k \to \infty$ for both side and we have $u_\infty^\top u_\infty \geq u_N^\top u_N - \mathcal{O}(h)$. Therefore, the limiting sharpness is at most $\mathcal{O}(h)$ away from $2/h$.

(2) Prove that $\exists$ one step with small sharpness

In the rest of the proof, we would like to show that GD will enter a flat region. Consider the Hessian $\nabla^2 f$. The trace is

$$\operatorname{tr}(\nabla^2 f) = (x^2 + y^2) G_1(xy - 1),$$

where

$$\begin{aligned}
G_1(\delta) &= \frac{2^{2-a} \log^{1-a}(2) \left(\log\left(e^{-\delta} + 1\right) + \log\left(e^{\delta} + 1\right)\right)^{a-2}}{\left(e^{\delta} + 1\right)^2} \\
&\qquad \times \left((a-1)\left(e^{\delta} - 1\right)^2 + 2e^{\delta} \log\left(e^{-\delta} + 1\right) + 2e^{\delta} \log\left(e^{\delta} + 1\right)\right) \\
&= \frac{2^{3-a} e^{\delta} \log^{1-a}(2) \left(\log\left(e^{-\delta} + 1\right) + \log\left(e^{\delta} + 1\right)\right)^{a-1}}{\left(e^{\delta} + 1\right)^2} \\
&+ \frac{2^{2-a} (a-1) \left(e^{\delta} - 1\right)^2 \log^{1-a}(2) \left(\log\left(e^{-\delta} + 1\right) + \log\left(e^{\delta} + 1\right)\right)^{a-2}}{\left(e^{\delta} + 1\right)^2} \\
&= q(\delta) \frac{2\delta}{e^{\delta} - e^{-\delta}} + q(\delta) \frac{(a-1)\left(e^{\delta} - 1\right)\delta}{\left(e^{\delta} + 1\right)\left(\log\left(e^{-\delta} + 1\right) + \log\left(e^{\delta} + 1\right)\right)} \\
&\leq q(\delta) \frac{2\delta}{e^{\delta} - e^{-\delta}} + q(\delta)(a-1)\left(1 - \frac{2\delta}{e^{\delta} - e^{-\delta}}\right) \\
&= \left((2-a)\frac{2\delta}{e^{\delta} - e^{-\delta}} - (1-a)\right) q(\delta).
\end{aligned}$$

Let

$$q_1(\delta, a) = (2-a)\frac{2\delta}{e^{\delta} - e^{-\delta}} - (1-a).$$

Then both $q_1(\delta, a)$ and $q(\delta)$ decreases as δ increases for $\delta \geq 0$. Also, fix δ, $q_1(\delta, a)$ and $q(\delta, a)$ increases as a increases.

The determinant is

$$\det(\nabla^2 f) = G_2(xy-1)$$

where

$$\begin{aligned}G_2(\delta) = &-\frac{4^{2-a}\left(e^{\delta}-1\right)\log^{2-2a}(2)\left(\log\left(e^{-\delta}+1\right)+\log\left(e^{\delta}+1\right)\right)^{2a-3}}{\left(e^{\delta}+1\right)^3}\\ &\times\left(2(a-1)(\delta+1)\left(e^{\delta}-1\right)^2+\left(4e^{\delta}(\delta+1)+e^{2\delta}-1\right)\left(\log\left(e^{-\delta}+1\right)+\log\left(e^{\delta}+1\right)\right)\right).\end{aligned}$$

Also,

$$\begin{aligned}|G_2(\delta)| &\le \frac{4^{2-a}\left(e^{\delta}-1\right)\log^{2-2a}(2)\left(\log\left(e^{-\delta}+1\right)+\log\left(e^{\delta}+1\right)\right)^{2a-3}}{\left(e^{\delta}+1\right)^3}\\ &\quad\times\left(4e^{\delta}(\delta+1)+e^{2\delta}-1\right)\left(\log\left(e^{-\delta}+1\right)+\log\left(e^{\delta}+1\right)\right)\\ &= \frac{4^{2-a}\left(e^{\delta}-1\right)\log^{2-2a}(2)\left(4e^{\delta}(\delta+1)+e^{2\delta}-1\right)}{\left(e^{\delta}+1\right)^3\left(\log\left(e^{-\delta}+1\right)+\log\left(e^{\delta}+1\right)\right)^{2-2a}}\\ &= 4^{2-a}\log^{2-2a}(2)\frac{\left(e^{\delta}-1\right)\left(4e^{\delta}(\delta+1)+e^{2\delta}-1\right)}{\left(e^{\delta}+1\right)^3}\frac{1}{\left(\log\left(e^{-\delta}+1\right)+\log\left(e^{\delta}+1\right)\right)^{2-2a}}\\ &\le 4^{2-a}\log^{2-2a}(2)\frac{\left(e^{\delta}-1\right)\left(4e^{\delta}(\delta+1)+e^{2\delta}-1\right)}{\left(e^{\delta}+1\right)^3}\\ &\le 32\log^2(2).\end{aligned}$$

where the first inequality follows from $2(a-1)(\delta+1)\left(e^{\delta}-1\right)^2 \le 0$ and $|2(a-1)(\delta+1)\left(e^{\delta}-1\right)^2| \le \left(4e^{\delta}(\delta+1)+e^{2\delta}-1\right)\left(\log\left(e^{-\delta}+1\right)+\log\left(e^{\delta}+1\right)\right)$; the second inequality follows from $1/\left(\log\left(e^{-\delta}+1\right)+\log\left(e^{\delta}+1\right)\right)^{2-2a} \le 1$, with equality achieved at $a=0,\delta=0$; the last inequality follows from $\frac{\left(e^{\delta}-1\right)\left(4e^{\delta}(\delta+1)+e^{2\delta}-1\right)}{\left(e^{\delta}+1\right)^3} \le 2$ and $4^{2-a}\log^{2-2a}(2) \le 16\log^2(2)$. Then we have

$$\sqrt{|G_2(\delta)|} \le \sqrt{32\log^2(2)} \le \frac{1}{7}4c_1^{-4/3} \lesssim \frac{1}{7}u_0^{\top}u_0.$$

By Taylor expansion, the magnitude of the largest eigenvalue of Hessian (sharpness) is upper bounded by

$$\max\left\{|G_1(\delta)|(x^2+y^2)+\mathcal{O}\left(\frac{|G_2(\delta)/G_1(\delta)|}{x^2+y^2}\right), \sqrt{|G_2(\delta)|}+\mathcal{O}\left(\frac{G_1(\delta)^2}{\sqrt{|G_2(\delta)|}}(x^2+y^2)^2\right)\right\}.$$

Let $h=\frac{C}{x_0^2+y_0^2}$. If $|r_0|<1$, then

$$q(\delta_0) \lesssim \frac{2}{C}.$$

Otherwise, similar to the proof of limiting sharpness, $u_k^\top u_k$ decreases at most $\tilde{\mathcal{O}}(1)$ before GD enters the region where δ_k starts to decrease for the first time, i.e., $|r_k| = 1 - \mathcal{O}(h)$. Therefore, for such $\delta = \delta_k$, we also have

$$1 - c_1\delta^2 \le q(\delta) \lesssim \frac{2}{C},$$

where $\tilde{\mathcal{O}}(h)$ term is omitted. Then

$$\delta \gtrsim \frac{\sqrt{C-2}}{\sqrt{Cc_1}},$$

and

$$G_1(\delta) \lesssim \frac{2}{C} q_1(\frac{\sqrt{C-2}}{\sqrt{Cc_1}}, a) \le \frac{2}{C} q_1(\frac{\sqrt{C-2}}{\sqrt{C/12}}, 1),$$

where the last inequality follows easily from checking the derivative of $q_1(\frac{\sqrt{C-2}}{\sqrt{Cc_1(a)}}, a)$ with respect to a. Also, when $G_1(\delta) < 0$, by checking the derivative and values of $G_1(\delta)$, we can bound the minimum

$$|G_1(\delta)| \le |\min_\delta G_1(\delta)| \le \frac{1}{2} q_1(\frac{\sqrt{4-2}}{\sqrt{4/12}}, 1) \le \frac{2}{C} q_1(\frac{\sqrt{C-2}}{\sqrt{C/12}}, 1).$$

Also, $S_\infty \approx \frac{2}{h} = \frac{2}{C} u_0^\top u_0$. Based on the above discussion, $u_k^\top u_k \le u_0^\top u_0 - \mathcal{O}(h^{1-m})$. Therefore,

$$S_k \lesssim q_1(\frac{\sqrt{C-2}}{\sqrt{C/12}}, 1) S_\infty \le \frac{1}{4}(6-C) S_\infty,$$

where the last inequality follows from linearization and shift. ■

D.3 Proof of balancing

Proof [Proof of Theorem 9] By Theorem 6 and the global minima $xy = 1$,

$$(x_\infty - y_\infty)^2 = u_\infty^\top u_\infty - 2 \le \frac{2}{h} - 2 = \frac{2}{C}(x_0^2 + y_0^2 - 2x_0y_0) + \frac{4}{C}x_0y_0 - 2 \le \frac{2}{C}(x_0 - y_0)^2 + 2(x_0y_0 - 1).$$

■

D.4 Supplementary lemmas

We first show a summary of the behavior of $u_k^\top u_k$ in the following lemma, which is the main idea of the proof of convergence.

Lemma 22 *Under the assumption of Theorem 13, there exist an increasing sequence $\{2N_i\}_{i\in\mathbb{Z}}$ with $N_i \in \mathbb{Z}$ and $N_{i+1} > N_i$, s.t., ${u_{2N_{i+1}}}^\top u_{2N_{i+1}} \le {u_{2N_i}}^\top u_{2N_i}$. More precisely, consider the even number of iterations, i.e., $k = 2i$ for $i \in \mathbb{Z}$,*

- *Stage I (not necessary):* $u_k^\top u_k$ *increases but is bounded by* $\frac{4}{h} + \mathcal{O}(h)$. *This happens when* $|x_k y_k - 1|$ *is too small at the early stage of iterations (see Lemma 23);*
- *Stage II:* $u_k^\top u_k$ *decreases every two steps when* x_k, y_k *is outside a 'monotone decreasing region' (see Lemma 26);*
- *Stage III: when* x_k, y_k *is near the 'monotone decreasing region', there exists* N, *s.t.* $u_{k+2N}^\top u_{k+2N}$ *is smaller than* $u_k^\top u_k$ *(see Lemma 29);*
- *Stage IV: when* $|x_k y_k - 1|$ *monotonically decreases,* $u_k^\top u_k$ *decreases every two steps (see Lemma 26 and Lemma 29)*

Lemma 23 *Under the assumption of Theorem 13, we have*

$$u_k^\top u_k \le \frac{4}{h} + \mathcal{O}(h), \forall k \in \mathbb{N}.$$

Proof

By Lemma 25 and Lemma 26, we know that the increase of $u_k^\top u_k$ after two step can only happen when $r_k \le -1$ and $0 < x_k y_k - 1 < 1$.

WOLG assume N is the first step s.t. $u_N^\top u_N = 4/h + \mathcal{O}(h)$. By Lemma 26, we would like to show that there exists k s.t. for $x_k y_k > 1$, we have either $x_k y_k - 1 > 1$ or $-1 < r_k < 0$.

Let $\delta_k = |x_k y_k - 1|$. Then

$$|r_k| \ge 4(1 - c_1\delta_k^2) - 1 + \mathcal{O}(h^2) \ge 4(1 - \delta_k^2/4) - 1 + \mathcal{O}(h^2)$$

and therefore

$$\delta_{k+1} \ge \delta_k(3 - 4c_1\delta_k^2 + \mathcal{O}(h^2)) \ge \delta_k(3 - \delta_k^2 + \mathcal{O}(h^2)).$$

If $\delta_k < 1$, it takes GD $\mathcal{O}(-\log(\delta_0))$ steps to go to $\delta_k \ge 1$. When $\delta_k \ge 1$ $u_k^\top u_k$ already starts to decrease every two steps. Moreover,

$$\delta_{k+1} \ge \delta_k(3 - \delta_k^2 + \mathcal{O}(h^2)) = \delta_k + 2\delta_k - \delta_k^3 + \mathcal{O}(h^2).$$

Therefore, δ_k keeps increasing for a while, staying in $\delta_k \ge 1$. Therefore, $u_k^\top u_k$ stays in the decreasing region with increase of at most $\mathcal{O}(h)$ for each step (for more detailed version, see Lemma 22), i.e., $u_k^\top u_k \le \frac{4}{h} + \mathcal{O}(h)$ for all k. ■

Lemma 24 *Under the assumption of Theorem 13, GD does not converge outside* $\{(x, y)|x^2 + y^2 \le 2/h\}$.

Proof We first would like to remove the initial condition that can converge in finite steps to the minima outside of this region. It turns out that such initial conditions form a null set. The proof is almost the same as Wang et al. (2022a) except for some easy calculations and thus omitted. We further remove all the initial conditions that converges to the periodic orbits, i.e., $\prod_{i=1}^n r_k \cdots r_{k+2i-1} = 1$ for all k and n. By similar argument in Wang et al. (2022a) (i.e., for each k, n, this is a null set, and therefore the union of countably many null sets is still a null set), such initial conditions also form a null set.

Assume $u_k^\top u_k \geq 2/h + \epsilon_0$ for all k, where $\epsilon_0 > 0$. Consider the case when $|x_k y_k - 1| < \epsilon_1 = \frac{\sqrt{h\epsilon_0}}{2}$. Then

$$\begin{aligned}|r_k| &\geq |h(1-\mathcal{O}(\epsilon_1^2))(2/h+\epsilon_0) + \mathcal{O}(h^2\epsilon_1) - 1| \\ &= |1 - 2\epsilon_1^2 + h\epsilon_0 + \mathcal{O}(h^2\epsilon_1)| \\ &= |1 + \frac{h}{2}\epsilon_0| > 0\end{aligned}$$

Namely, GD cannot converge with $u_k^\top u_k > 2/h$. ■

Lemma 25 *Under the assumption of Theorem 13, if $r_k < 0$, then $r_n < 0$ for all $n \geq k$.*

Proof From Lemma 18, we have $|C_{k-1} - C_k| = \mathcal{O}(h^2)$. Also $C_k = \frac{1-r_k}{q(\delta_k)}$. If $r_k < 0$,

$$\begin{aligned}r_{k+1} = 1 - C_{k+1}q(\delta_{k+1}) &\leq 1 - \frac{1-r_k}{q(\delta_k)}q(\delta_{k+1}) + \mathcal{O}(h^2) \\ &= 1 - \frac{1-r_k}{q(\delta_k)}q(r_k\delta_k) + \mathcal{O}(h^2)\end{aligned}$$

By Proposition 21, it suffices to prove that

$$R_0(\delta_k, -r_k) := (1-r_k)q(-r_k\delta_k) - q(\delta_k) > 0, \ \forall \delta_k > 0$$

We abuse the notation and let $r_k(\delta) = 1 - C_k q(\delta)$ for fixed $C_k > 0$. By Proposition 21, $r_k(\delta)$ monotonically increases for $\delta > 0$ and there is only one root denoted as δ_0 s.t. $r_k(\delta) = 0$ (Since $r_k < 0$, there exists δ s.t. $r_k(\delta) < 0$ and when δ is large, $r_k(\delta) > 0$; therefore, it has a root). When $r_k(\delta_0) = r_k = 0$, by Proposition 21, $R_0(\delta_0, -r_k(\delta_0)) = R_0(\delta_0, 0) = 1 - q(\delta_0) > 0$. Also

$$\begin{aligned}R_0(\delta, -r_k(\delta)) &= C_k q(\delta) q\left((C_k q(\delta) - 1)\delta\right) - q(\delta) \\ &= q(\delta)\left[C_k q\left((C_k q(\delta) - 1)\delta\right) - 1\right]\end{aligned}$$

Therefore, we only need to show $K(\delta) := (C_k q(\delta) - 1)\delta < \delta_0$ for $\delta \in (0, \delta_0)$. The derivative of K is

$$K'(\delta) = C_k q(\delta) - 1 + C_k \delta q'(\delta) < C_k q(\delta) - 1.$$

Then the maximum point of $K(\delta)$ is achieved at $\delta_1 < \delta_0$ by Proposition 21. Thus,

$$R_0(\delta, -r_k(\delta)) > 0,$$

and consequently

$$R_0(\delta_k, -r_k) \geq q(\delta_k)(C_k q(K(\delta_1)) - 1) > 0,$$

$$r_{k+1} \leq -\frac{R_0(\delta_k, -r_k)}{q(\delta_k)} + \mathcal{O}(h^2) \leq -(C_k q(R_1(\delta_1)) - 1) + \mathcal{O}(h^2) < 0,$$

since this $\mathcal{O}(h^2)$ is indeed bounded by $c\,\delta_k h^2$ for some universal constant $c > 0$ and can be controlled by the first term. ■

Lemma 26 *Under the assumption of Theorem 13, the following properties are the two step decrease of $u_k^\top u_k$ in different cases:*

1. *If $r_k < 0$ and $u_k^\top u_k \le \frac{4}{h} + \mathcal{O}(h)$, there exists $R_1(a, r_k) \le 1$, s.t., when $x_k y_k - 1 > R_1(a, r_k)$, we have*
$$u_{k+2}^\top u_{k+2} \le u_k^\top u_k - \frac{1}{2}h + \mathcal{O}(h^3 \ell_{k+1}^2).$$
2. *If $-1 < r_k < 0$ and $0 < x_k y_k - 1 \le 1$, then*
$$u_{k+2}^\top u_{k+2} \le u_k^\top u_k - 3.2h(1 + r_k)(x_k y_k - 1).$$
3. *If $r_k \le -1$, $u_k^\top u_k \le \frac{3}{h}$, and $x_k y_k < 1$, then*
$$u_{k+2}^\top u_{k+2} \le u_k^\top u_k - \min\{\mathcal{O}(h), \mathcal{O}(h(x_k y_k - 1))\}.$$

Proof First, we consider the two-step update of $u_k^\top u_k$ in the following

$$\begin{aligned}
u_{k+2}^\top u_{k+2} &= (1 + h^2 \ell_{k+1}^2) u_{k+1}^\top u_{k+1} - 4h\ell_{k+1} x_{k+1} y_{k+1} \\
&= (1 + h^2 \ell_{k+1}^2)((1 + h^2 \ell_k^2) u_k^\top u_k - 4h\ell_k x_k y_k) - 4h\ell_{k+1} x_{k+1} y_{k+1} \\
&\le u_k^\top u_k - 4h[\ell_k x_k y_k - \ell_k^2 + \ell_{k+1} x_{k+1} y_{k+1} - \ell_{k+1}^2] \\
&\quad - 4h^3 \ell_{k+1}^2 \ell_k x_k y_k + h^4 \ell_{k+1}^2 \ell_k^2 u_k^\top u_k + \mathcal{O}(h^3 \ell_{k+1}^2) \\
&= u_k^\top u_k - 4h[\ell_k x_k y_k - \ell_k^2 + \ell_{k+1} x_{k+1} y_{k+1} - \ell_{k+1}^2] + \mathcal{O}(h^3 \ell_{k+1}^2).
\end{aligned}$$

It suffices to analyze

$$\ell_k x_k y_k - \ell_k^2 + \ell_{k+1} x_{k+1} y_{k+1} - \ell_{k+1}^2 = L(\delta_k) + L(r_k \delta_k), \text{ where } L(\delta) = \ell(\delta)(\delta + 1 - \ell(\delta)).$$

By checking the slope and values of the above function, we have

$$L(\delta_k) + L(r_k \delta_k) \ge L(1) + L(r_k) > 0.14$$

for all $r_k < 0$, $0 < a \le 1$, and $\delta_k \ge 1$.

When $0 < x_k y_k - 1 \le 1$, in order to make $-1 < r_k < 0$, we should at least have $u_k^\top u_k \le \frac{4}{h}$. Also we have $\ell_k \le x_k y_k - 1 \le 1$ in this region by Proposition 21. Therefore, by Proposition 20

$$\begin{aligned}
u_{k+2}^\top u_{k+2} &= (1 + h^2 \ell_{k+1}^2) u_{k+1}^\top u_{k+1} - 4h\ell_{k+1} x_{k+1} y_{k+1} \\
&= (1 + h^2 \ell_{k+1}^2)((1 + h^2 \ell_k^2) u_k^\top u_k - 4h\ell_k x_k y_k) - 4h\ell_{k+1} x_{k+1} y_{k+1} \\
&\le u_k^\top u_k - 4h[\ell_k x_k y_k - \ell_k^2 + \ell_{k+1} x_{k+1} y_{k+1} - \ell_{k+1}^2] - 4h^3 \ell_{k+1}^2 \ell_k x_k y_k + h^4 \ell_{k+1}^2 \ell_k^2 u_k^\top u_k \\
&\le u_k^\top u_k - 4h[\ell_k x_k y_k - \ell_k^2 + \ell_{k+1} x_{k+1} y_{k+1} - \ell_{k+1}^2] \\
&\le u_k^\top u_k - 3.2h(1 + r_k)(x_k y_k - 1).
\end{aligned}$$

When $r_k \le -1$, $u_k^\top u_k \le \frac{3}{h}$, and $x_k y_k < 1$, by Taylor series and simple calculation, we have $L(\delta) + L(r\delta) \ge \min\{0.5(1 + r)\delta, L(-1) + L(1)\}$, where $L(-1) + L(1) \ge 0.14$. Then

$$\begin{aligned}
u_{k+2}^\top u_{k+2} &\le u_k^\top u_k - 4h[\ell_k x_k y_k - \ell_k^2 + \ell_{k+1} x_{k+1} y_{k+1} - \ell_{k+1}^2] - h(\ell_k^2 + \ell_{k+1}^2) + \mathcal{O}(h^3 \ell_{k+1}^2) \\
&\le u_k^\top u_k - 4h \min\{0.5(1 + r_k)(x_k y_k - 1), L(-1) + L(1)\}.
\end{aligned}$$

■

Lemma 27 *Under the assumption of Theorem 13, if $-1 \le r_k < 0$, there exists $R_2(a, r_k) \ge 1 - \mathcal{O}(h^2)$, s.t., for $|x_k y_k - 1| \le R_2(a, r_k)$, we have $r_{k+1} > -1$.*

Proof Since $r_k = 1 - C_k q(\delta_k)$, we have

$$r_{k+1} = 1 - C_{k+1} q(\delta_{k+1}) = 1 - \frac{1 - r_k}{q(\delta_k)} q(r_k \delta_k) \pm \mathcal{O}(h^2).$$

We denote

$$\delta r(r_k, \delta_k) \coloneqq 2 - \frac{1 - r_k}{q(\delta_k)} q(r_k \delta_k).$$

When $\delta_k > 0$, $\delta r(r_k, \delta_k)$ monotonically decreases as δ_k increases. Consider

$$\delta r(r_k, 1) = 2 - \frac{1 - r_k}{q(1)} q(r_k).$$

This function $\delta r(r_k, 1)$ monotonically decreases when r_k increases between -1 and 0 and $\delta r(-1, 0) > 0$. Therefore, $r_{k+1} > -1 \pm \mathcal{O}(h^2)$ and the conclusion follows from series expansion. ■

Lemma 28 *$R_2(a) - R_1(a) > c > 0$ for some constant c.*

Proof Instead of analyzing the expressions of R_1 and R_2, we will just check this property by dividing the interval of a, i.e., $(0, 1]$, into two subintervals and analyze a relaxation of R_1 and R_2. Consider $a \in [0.75, 1]$ and $a \in (0, 0.75)$. For the former one, $a \in [0.75, 1]$, $R_1 \le 1$ and $R_2 \ge 1.5$; for the latter one, $a \in (0, 0.75)$, $R_1 \le 0.85$ and $R_2 \ge 1$. The detailed analysis is by checking the trend of the following two quantities in the intervals bounded by the above-mentioned values, the decrease of $u_k^\top u_k$ every two steps and the value of r_{k+1} once $-1 < r_k < 0$ in previous lemmas. ■

Lemma 29 *Under the assumption of Theorem 13, if $r_k < 0$ for all $k \ge n$ given some $n \ge 0$, $|x_k y_k - 1|$ will eventually start to decrease in the region $|x_k y_k - 1| < R_2(a)$.*

Proof

By Lemma 26, when $|x_k y_k - 1| \ge R_1(a, r_k)$, $u_k^\top u_k$ keeps decreasing every two step with certain amount away from 0. Eventually, by the expression of $|r_k|$, $u_k^\top u_k$ will decrease until GD enters $|x_k y_k - 1| \le R_2(a, r_k)$. Note by Lemma 28, $R_2(a, r_k) > R_1(a, r_k)$. However, $|x_k y_k - 1|$ may not keep decreasing inside this region. WOLG, consider a lower bound of $R_2(a, r_k)$ to be $R_2(a) \ge 1 - \mathcal{O}(h^2)$ and we will use $R_2(a)$ independent of iterations. From the expression of $|r_k|$, we have at least $u_k^\top u_k \le \frac{3}{h}$ in this case.

First consider $|x_k y_k - 1| > R_2(a)$. According to Lemma 26 and Proposition 21, $u_k^\top u_k$ decreases every two steps and the function q decreases when $|x_k y_k - 1|$ increases. Therefore there exists k s.t. $|r_k| < 1$. Moreover, for $|r_k| < 1$, there exists n, s.t. $x_{k+2n} y_{k+2n} < x_k y_k$ and $|r_{k+2n}| < 1$ if $x_{k+2n} y_{k+2n} - 1 \ge R_1(a)$.

Next, assume k is such that $|x_k y_k - 1| > R_2(a)$ but $|x_{k+2} y_{k+2} - 1| \le R_2(a)$. According to the initial condition, there is no periodic orbit in the trajectory (see details in the proof of Lemma 24). We claim that there exists n, s.t., $|r_{k+2n}| < 1$ and $|x_{k+2n} y_{k+2n} - 1| \le R_2(a)$. Otherwise, if $u_{k+2}^\top u_{k+2}$ will still decrease every two steps, then it returns to the above cases when $|x_k y_k - 1| > R_2(a)$. Then we analyze the case where $u_{k+2}^\top u_{k+2}$ will increase after two steps. For the first n s.t. $x_{k+2n} y_{k+2n} > R_2(a)$, if we have $|r_{k+2n}(R_2(a))| < |r_k(R_2(a))|$, this implies $u_{k+2n}^\top u_{k+2n} < u_k^\top u_k$ which can be either absorbed in the above cases, or eventually fall into the following case. For the first n s.t. $x_{k+2n} y_{k+2n} > R_2(a)$, consider $|r_{k+2n}(R_2(a))| > |r_k(R_2(a))|$, with $|r_{k+2i}| \ge 1$ for all $i < n$ and $u_{k+2n}^\top u_{k+2n} > u_k^\top u_k$. Then if such process repeats, $|r_{k+2n}(R_2(a))|$ will be larger and larger until GD either starts to decrease in $|xy-1| \le R_2(a)$ or enters the two-step decreasing region of $u_k^\top u_k$ in $|xy-1| \le R_2(a)$ (we can use $|xy-1| \ge R_1(a)$ to represent this region; however, $R_1(a)$ is just a bound, meaning the actual region is larger), which will lead to $|r_{k+2n}| < 1$ inside this region due to the same reasoning as the previous paragraph. Detailed characterization of this stage can be seen in the proof of limiting sharpness.

■

Appendix E. Proofs of results for functions (2) with $b = 1$

Proof [Proof of Theorem 10] By Theorem 3.1 in Wang et al. (2022a),

$$(x-y)^2 \le \frac{2}{h} - 2.$$

Then since $h = \frac{C}{x_0^2+y_0^2+4}$, we have

$$(x-y)^2 \le \frac{2}{h} - 2 = \frac{2}{C}(x_0^2 + y_0^2 - 2x_0y_0 + 2x_0y_0 + 4) - 2 = \frac{2}{C}(x_0 - y_0)^2 + \frac{4}{C}(x_0y_0 + 2) - 2.$$

■

Appendix F. Proofs of results for functions (2) with $b \ge 3$

Consider function

$$f(x,y) = (1 - (xy)^b)^2/(2b^2), \text{ with } b = 2n+1 \text{ for } n \in \mathbb{Z}.$$

Let $p(s) = \sum_{i=0}^{b-1} s^i$. Then

$$1 - (xy)^b = (1-xy)p(xy).$$

Apart from the equations in Appendix C, we will also use the following two equations in our proofs

$$x_{k+1}y_{k+1} - 1 = (x_k y_k - 1)\left(1 - \frac{h}{b}(x_k y_k)^{b-1} p(x_k y_k) u_k^\top u_k + \frac{h^2}{b^2}(x_k y_k - 1)(x_k y_k)^{2b-2} p(x_k y_k)^2 x_k y_k\right),$$

$$u_{k+1}^\top u_{k+1} = u_k^\top u_k - \frac{h}{b}(x_k y_k - 1) \times \left((x_k y_k)^{b-1} p(x_k y_k) x_k y_k \left(4 - \frac{h}{b}(x_k y_k)^{b-1} p(x_k y_k) u_k^\top u_k \right) + \frac{h}{b}(x_k y_k)^{2b-2} p(x_k y_k)^2 u_k^\top u_k \right).$$

F.1 Proof of convergence

Proof [Proof of Theorem 15]

By Lemma 24 (with a different choice of null set based on the functions), GD can only converge to the point s.t. $x^2 + y^2 \le 2/h$. Also we remove all the points (which is in a null set) s.t. they converge in finite step. Therefore, $r_k \neq 0$ for all $k \ge 0$.

If $0 < r_0 < 1$, we have $x_1 y_1 > 1$ and for any $x_k y_k > 1$,

$$\begin{aligned} u_{k+1}^\top u_{k+1} &= u_k^\top u_k - \frac{h}{b}(x_k y_k - 1)\left((x_k y_k)^{b-1} p(x_k y_k) x_k y_k \left(4 - \frac{h}{b}(x_k y_k)^{b-1} p(x_k y_k) u_k^\top u_k \right) \right. \\ &\quad \left. + \frac{h}{b}(x_k y_k)^{2b-2} p(x_k y_k)^2 u_k^\top u_k \right) \\ &\le u_k^\top u_k. \end{aligned}$$

Assume $r_i > 0$ for $i = 0, \cdots, k$. Consider

$$r_{k+1} = \left(1 - h \frac{\ell_{k+1}}{x_{k+1} y_{k+1} - 1} (u_{k+1}^\top u_{k+1} - h \ell_{k+1} x_{k+1} y_{k+1}) \right)$$

where $\ell_k = \frac{(x_k y_k)^{b-1}((x_k y_k)^b - 1)}{b}$ and therefore $q_k = \frac{\ell_k}{x_k y_k - 1} = \frac{1}{b}(x_k y_k)^{b-1} p(x_k y_k)$. Moreover, $W(s) = \frac{1}{b} s^{b-1} p(s)$ monotonically increases when $s > 1$, which implies $q_{k+1} \le q_k$. Also,

$$\begin{aligned} 0 < u_{k+1}^\top u_{k+1} - h \ell_{k+1} x_{k+1} y_{k+1} &= (1 + h^2 \ell_k^2) u_k^\top u_k - 4 h \ell_k x_k y_k - h \ell_{k+1} x_{k+1} y_{k+1} \\ &\le u_k^\top u_k - h \ell_k x_k y_k + h \frac{C}{b} \ell_k x_k y_k - 3 h \ell_k x_k y_k - h \ell_{k+1} x_{k+1} y_{k+1} \\ &\le u_k^\top u_k - h \ell_k x_k y_k. \end{aligned}$$

Therefore, $r_{k+1} \ge r_k > 0$. Moreover, we have

$$\begin{aligned} r_{k+1} &= 1 - h \frac{\ell_{k+1}}{x_{k+1} y_{k+1} - 1} (u_{k+1}^\top u_{k+1} - h \ell_{k+1} x_{k+1} y_{k+1}) \\ &\le 1 - h \frac{\ell_{k+1}}{x_{k+1} y_{k+1} - 1} (2 x_{k+1} y_{k+1} - h \ell_{k+1} x_{k+1} y_{k+1}) \\ &\le 1 - h(2 - h \ell_0) < 1 \end{aligned}$$

where the second inequality follows from the value at $x_{k+1} y_{k+1} = 1$. Then $x_k y_k - 1$ exponentially decreases until it converges.

If $r_0 < 0$, from the upper bound of h, we have $0 < r_1 < 1$, $0 < x_1 y_1 < 1$, and $u_1^\top u_1 \le u_0^\top u_0$ (according to the above discussion for $x_0 y_0 > 1$). Then if $0 < x_k y_k < 1$,

$$\begin{aligned} u_{k+1}^\top u_{k+1} &= (1 + h^2 \ell_k^2) u_k^\top u_k - 4 h \ell_k x_k y_k \\ &= u_k^\top u_k + h |\ell_k| (4 x_k y_k + h |\ell_k| u_k^\top u_k) \\ &\le u_k^\top u_k + (4 + C u_k^\top u_k / (b u_0^\top u_0)) h |\ell_k| x_k y_k \end{aligned}$$

where the inequality follows from $x_k y_k < 1$ and $x_0 y_0 > 1$.

Also, when $0 < x_k y_k < 1$,

$$\begin{aligned} x_{k+1}y_{k+1} &= x_k y_k + (1 - r_k)(1 - x_k y_k) \\ &= x_k y_k - h\ell_k(u_k^\top u_k - h\ell_k x_k y_k) \\ &\geq x_k y_k - h\ell_k x_k y_k(2 - h\ell_k) \\ &= x_k y_k + h|\ell_k| x_k y_k(2 + h|\ell_k|) \\ &\geq x_k y_k + 2h|\ell_k| x_k y_k. \end{aligned}$$

We claim that when $u_k^\top u_k \leq u_0^\top u_0$ and $0 < x_k y_k < 1$, we have $u_n^\top u_n \leq u_0^\top u_0 + C_1$ for all $n \geq k$ and for constant $\frac{10}{3} < C_1 \leq \frac{7}{2}$. Otherwise, consider N s.t. $u_n^\top u_n \leq u_0^\top u_0 + C_1$ for $k \leq n \leq N$ and $u_{N+1}^\top u_{N+1} \geq u_0^\top u_0 + C_1$. If $0 < x_n y_n < 1$,

$$\begin{aligned} r_n &= 1 - h\frac{\ell_n}{x_n y_n - 1}(u_n^\top u_n - h\ell_n x_n y_n) \\ &\geq 1 - h(u_n^\top u_n - h\ell_n x_n y_n) \\ &\geq 1 - \frac{C}{4(u_0^\top u_0 + 4)}(u_n^\top u_n - h\ell_n x_n y_n) \\ &\geq 1 - \frac{u_0^\top u_0 + C_1 + h|\ell_n| x_n y_n}{u_0^\top u_0 + 4} \\ &\geq 1 - \frac{u_0^\top u_0 + 15/4}{u_0^\top u_0 + 4} > 0 \end{aligned}$$

where the first inequality follows from $q_n \leq 1$ for $0 < x_n y_n < 1$; the second inequality follows from the initial condition and the requirement of h; the last inequality follows from $h|\ell_n| x_n y_n < \frac{1}{4}$ for $0 < x_n y_n < 1$ (can be easily checked from the initial condition that $h \leq \frac{1}{8}$, and the rest follows from analyzing the expression of the function). Therefore, $0 < x_{n+1} y_{n+1} < 1$ and consequently, $0 < r_n < 1$ for $n = k, \cdots, N$. Especially, consider $n = N$. Then iteratively from the lower bound of $x_{N+1} y_{N+1}$ above, we have

$$\sum_{i=k}^{N} 2h|\ell_i| x_i y_i < 1$$

and thus

$$\begin{aligned} \sum_{i=k}^{N}(4 + Cu_i^\top u_i/(bu_0^\top u_0))h|\ell_i| x_i y_i &\leq (4 + C(u_0^\top u_0 + C_1)/(bu_0^\top u_0)) \sum_{i=k}^{N} h|\ell_i| x_i y_i \\ &< \frac{20}{3}\sum_{i=k}^{N} h|\ell_i| x_i y_i < \frac{10}{3} < C_1. \end{aligned}$$

Contradiction.

Then, we have $0 < r_n < 1$ for all $n \geq k$. Take $k = 1$, and then we have the monotone decreasing of $1 - x_n y_n$. Also,

$$\begin{aligned} r_n &= 1 - h\frac{\ell_n}{x_n y_n - 1}(u_n^\top u_n - h\ell_n x_n y_n) \\ &\leq 1 - hq_1(2x_n y_n - h\ell_n x_n y_n) \\ &\leq 1 - hq_1 x_1 y_1(2 - h\ell_n) \\ &\leq 1 - h(2-h)q_1 x_1 y_1 < 1. \end{aligned}$$

Thus, GD will converge to $xy = 1$.

■

F.2 Proof of non-EoS

Proof [Proof of Theorem 8]

From the proof of convergence, we know $r_k > 0$ for $k \geq 1$. Thus when $x_k y_k$ is very close to 1, $r_k > 0$. Take the limit and we have

$$\lim_{k\to\infty} r_k = 1 - hu_\infty^\top u_\infty \geq 0.$$

Therefore,

$$S_\infty = u_\infty^\top u_\infty \leq \frac{1}{h}$$

■

F.3 Proof of non-balancing

Below is a more detailed version of Theorem 11.

Theorem 30 (no Balancing, formal version) *Consider $b = 2n + 1$ for $n \in \mathbb{Z}$. Assume the initial condition satisfies*

$$(x_0, y_0) \in \{(x, y) : xy > 4^{\frac{1}{2b-2}}, x^2 + y^2 \geq 4\} \backslash \mathcal{B}_b, \text{ where } \mathcal{B}_b \text{ is a Lebesgue measure-0 set.}$$

Let the learning rate be

$$h = \frac{C}{(x_0^2 + y_0^2 + 4)(x_0 y_0)^{2b-2}} \text{ for } 2 \leq C \leq M_3(b, x_0, y_0), \text{ where } 3 < M_3(b, x_0, y_0) \leq 4,$$

and the precise definition of M_3 is given in Theorem 15. Then GD converges to a global minimum (x_∞, y_∞), and we have

$$(x_\infty - y_\infty)^2 \geq \begin{cases} (x_0 - y_0)^2 - \frac{2C}{4b-C}(x_0 y_0 - 1), & \text{if } r_0 > 0, \\ (x_0 - y_0)^2 + 2(x_0 y_0 - 1) - \frac{2C}{b} x_0 y_0, & \text{otherwise} \end{cases}$$

Proof [Proof of Theorem 30 (and Theorem 11)]
Let $h = \frac{C}{(x_0^2+y_0^2+4)(x_0y_0)^{2b-2}}$. Before the proof, let first consider

$$h\ell_0 = \frac{C}{(x_0^2+y_0^2+4)(x_0y_0)^{2b-2}} \frac{(x_0y_0)^{b-1}\left((x_0y_0)^b-1\right)}{b} \leq \frac{Cx_0y_0}{b(x_0^2+y_0^2+4)} \leq \frac{C}{2b}.$$

If $r_0 > 0$, from the proof of convergence, we know: $r_k > 0$ for all k, and $\ell_k \geq 0$ for all k. Moreover, we have $x_k y_k \geq x_{k+1}y_{k+1}$, and consequently $\ell_k \leq \ell_0$ for all k (by the monotone decreasing of this function; see details in the proof of convergence). Then

$$\begin{aligned} x_{k+1}y_{k+1} &= x_ky_k - h\ell_k(u_k^\top u_k - h\ell_k x_k y_k) \leq x_ky_k - h\ell_k x_k y_k(2-h\ell_k) \\ &\leq x_ky_k - h\ell_k x_k y_k(2-\frac{C}{2b}) \end{aligned}$$

Therefore

$$x_0y_0 - 1 = \sum_{k=0}^{\infty} x_ky_k - x_{k+1}y_{k+1} \geq \sum_{k=0}^{\infty} h\ell_k x_k y_k(2-\frac{C}{2b}).$$

Also, we have

$$\begin{aligned} u_{k+1}^\top u_{k+1} &= (1+h^2\ell_k^2)u_k^\top u_k - 4h\ell_k x_k y_k \\ &\geq u_k^\top u_k - 4h\ell_k x_k y_k \end{aligned}$$

Then

$$u_0^\top u_0 - u_\infty^\top u_\infty = \sum_{k=0}^{\infty} u_k^\top u_k - u_{k+1}^\top u_{k+1} \leq \sum_{k=0}^{\infty} 4h\ell_k x_k y_k \leq \frac{8b}{4b-C}(x_0y_0-1) \leq \frac{2b}{b-1}(x_0y_0-1)$$

Then

$$\begin{aligned} (x_\infty - y_\infty)^2 = u_\infty^\top u_\infty - 2x_\infty y_\infty &\geq u_0^\top u_0 - 2x_\infty y_\infty - \frac{8b}{4b-C}(x_0y_0-1) \\ &= u_0^\top u_0 - 2x_0y_0 + 2x_0y_0 - 2 - \frac{8b}{4b-C}(x_0y_0-1) \\ &= (x_0-y_0)^2 - \frac{2C}{4b-C}(x_0y_0-1). \end{aligned}$$

If $r_0 < 0$, from the proof of convergence, we have: $r_k > 0$ for $k \geq 1$, and $u_{k+1}^\top u_{k+1} \geq u_k^\top u_k$ for $k \geq 1$. Thus

$$u_0^\top u_0 - u_\infty^\top u_\infty \leq u_0^\top u_0 - u_1^\top u_1 \leq 4h\ell_0 x_0 y_0 \leq \frac{2C}{b} x_0 y_0$$

Then

$$\begin{aligned} (x_\infty - y_\infty)^2 = u_\infty^\top u_\infty - 2x_\infty y_\infty &\geq u_0^\top u_0 - 2x_\infty y_\infty - \frac{2C}{b} x_0 y_0 \\ &= u_0^\top u_0 - 2x_0y_0 + 2x_0y_0 - 2 - \frac{2C}{b} x_0 y_0 \\ &= (x_0-y_0)^2 + 2(x_0y_0-1) - \frac{2C}{b} x_0 y_0. \end{aligned}$$

■

F.4 Supplementary lemmas

Lemma 31 *Under the assumption of Theorem 15, $M_3 > 3$.*

Proof By the assumption, we have

$$\begin{aligned}
&\left(1+\left(\frac{C}{(x_0^2+y_0^2+4)(x_0y_0)^{2b-2}}\right)^2\ell_0^2\right)x_0y_0 - \frac{C}{(x_0^2+y_0^2+4)(x_0y_0)^{2b-2}}\ell_0(x_0^2+y_0^2)\\
&= \left(1+\left(\frac{C\left((x_0y_0)^b-1\right)}{b(x_0^2+y_0^2+4)(x_0y_0)^{b-1}}\right)^2\right)x_0y_0 - \frac{C\left((x_0y_0)^b-1\right)}{b(x_0^2+y_0^2+4)(x_0y_0)^{b-1}}(x_0^2+y_0^2)\\
&\geq x_0y_0 - \frac{C(x_0^2+y_0^2)\left((x_0y_0)^b-1\right)}{b(x_0^2+y_0^2+4)(x_0y_0)^{b-1}}
\end{aligned}$$

Since $b \geq 3$, when $C = 3$, we have

$$x_0y_0 - \frac{C(x_0^2+y_0^2)\left((x_0y_0)^b-1\right)}{b(x_0^2+y_0^2+4)(x_0y_0)^{b-1}} > 0.$$

■

Lemma 32 (stepsize) *Let C_1=4. When $x_0^2+y_0^2 \geq 4\left(\sqrt{2}+2\right)C_1$, the learning rate bound in Theorem 15*

$$\frac{2}{(u_0^\top u_0+C_1)(x_0y_0)^{2b-2}} \leq h \leq \frac{M_3(x_0,y_0)}{(u_0^\top u_0+C_1)(x_0y_0)^{2b-2}}$$

satisfies

$$\frac{2}{S_0} < \frac{2}{(u_0^\top u_0+C_1)(x_0y_0)^{2b-2}}, \text{ and } \frac{4}{S_0} \leq \frac{M_3(x_0,y_0)}{(u_0^\top u_0+C_1)(x_0y_0)^{2b-2}}.$$

Proof Consider the Hessian $\nabla^2 f(x,y)$. The trace is

$$\begin{aligned}
\operatorname{tr}\nabla^2 f(x,y) &= (1+(1-1/b)(1-1/(xy)^b))(x^2+y^2)(xy)^{2b-2}\\
&> (1+(1-1/b)(1-\frac{1}{4^{\frac{b}{2b-2}}}))(x^2+y^2)(xy)^{2b-2}\\
&\geq \left(1+\frac{2}{3}\left(1-\frac{1}{2\sqrt{2}}\right)\right)(x^2+y^2)(xy)^{2b-2}\\
&> \frac{4}{3}(x^2+y^2)(xy)^{2b-2},
\end{aligned}$$

where the second inequality is via lower bounding of the expression at $b = 3$.

The determinant is

$$\det\nabla^2 f(x,y) = -\frac{(xy)^{2b}\left((xy)^b-1\right)\left(-(xy)^b+b\left(4(xy)^b-2\right)+1\right)}{b^2x^2y^2}.$$

When $xy > 1$, $\det \nabla^2 f(x,y) < 0$. Therefore, initial sharpness

$$S_0 > \operatorname{tr} \nabla^2 f(x_0, y_0).$$

Then when $x_0^2 + y_0^2 \geq 4\left(\sqrt{2}+2\right)C_1 = 16\left(\sqrt{2}+2\right)$, we have

$$S_0 > \left(1 + \frac{2}{3}\left(1 - \frac{1}{2\sqrt{2}}\right)\right)(x_0^2 + y_0^2)(x_0 y_0)^{2b-2} \geq \frac{4}{3}(x_0^2 + y_0^2 + C_1)(x_0 y_0)^{2b-2},$$

and thus the lower bound of h is greater than $\frac{2}{S_0}$, actually $\frac{8}{3S_0}$, and the upper bound of h is greater than $\frac{4M_3}{3S_0} > \frac{4}{S_0}$ since $M_3 > 3$. ■

Appendix G. Additional experiments

In this section, we provide the experiment details. We perform experiments on the CIFAR-10 and MNIST data sets with models of different regularities. Throughout the training process, we apply the full batch gradient descent with a group of constant learning rates to illustrate the presence/absence of large learning rate phenomena in different regularity settings. The algorithm does not use any weight decay or momentum.

G.1 CIFAR-10

To shorten the training time, we train our model on CIFAR-10-1k, which consists of 1000 randomly chosen training samples from CIFAR-10. The data dimension is $N_0 = 3 \times 32 \times 32 = 3072$. The labels are transformed into their one-hot encoding, and the output dimension is $N_2 = 10$.

We consider 2-layer neural networks in (3), i.e., $l = 2$, with $N_1 = 200$ hidden neurons. Correspondingly, the weight matrices for the two layers are $W_1 \in \mathbb{R}^{N_0 \times N_1}$ and $W_2 \in \mathbb{R}^{N_1 \times N_2}$. We exclude any bias terms in our models. The loss function $\mathcal{L}$ is chosen from ℓ^2 (MSE) loss and the huber loss. The activation function σ_1 is chosen from ReLU, tanh and ReLU^3.

For the weight initialization, we first use the default uniform initialization in PyTorch. Each weight in the i-th layer ($i = 1, 2$) follows the uniform distribution on $[-\frac{1}{\sqrt{N_{i-1}}}, \frac{1}{\sqrt{N_{i-1}}}]$. We then scale the weight matrices so that $\|W_1\|_{\mathrm{F}} = 6$ and $\|W_2\|_{\mathrm{F}} = 20$.

We adopt the same Lanczos algorithm as in Cohen et al. (2021) to compute the leading eigenvalue of Hessian, i.e., the sharpness. For the settings with total epochs no greater than $20,000$, we compute the sharpness after every 40 epochs. For those settings with total epochs greater than $20,000$, we compute the sharpness after every 400 epochs.

The results of the above mentioned six cases are shown in Figure 181920212223. All the cases are consistent with our theory that bad regularity vanishes large learning rate phenomena, except for huber loss with ReLU^3. The reason for this inconsistency potentially lies in the fact that the ReLU^3 is only added to W_1, while W_2 is still linear under the corresponding objective function (and linear function, from our theory, has good regularity). The disagreement of the 'order' of weights is beyond the scope of our theory and we will leave it for future exploration.

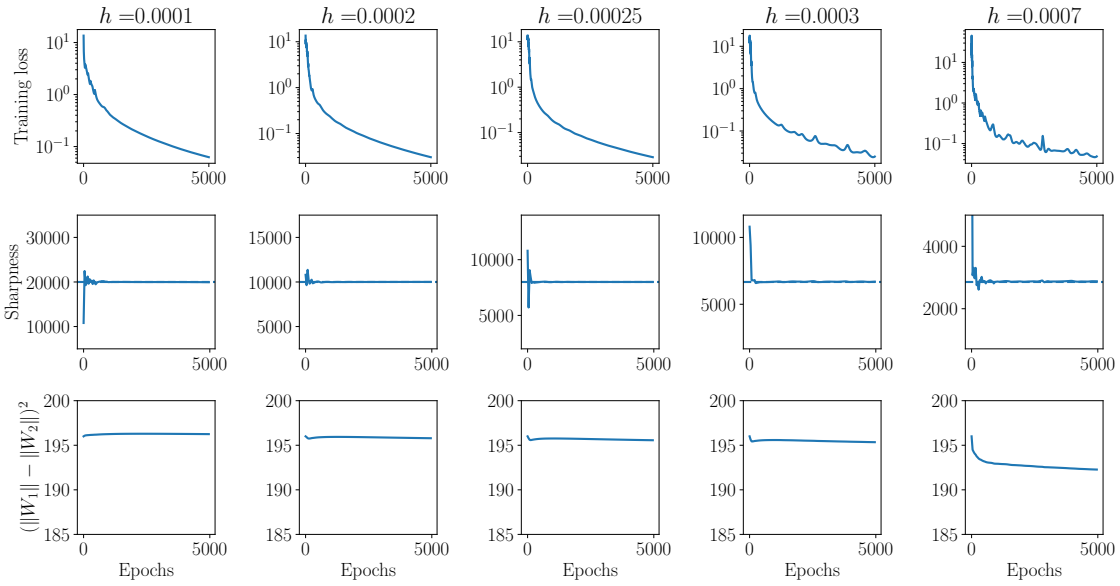


Figure 18: CIFAR-10: huber+tanh, 2-layer. Each column shows training with h increasing from small to near-divergent values. The initial condition is $\|W_1\|_{\mathrm{F}} = 6$ and $\|W_2\|_{\mathrm{F}} = 20$.

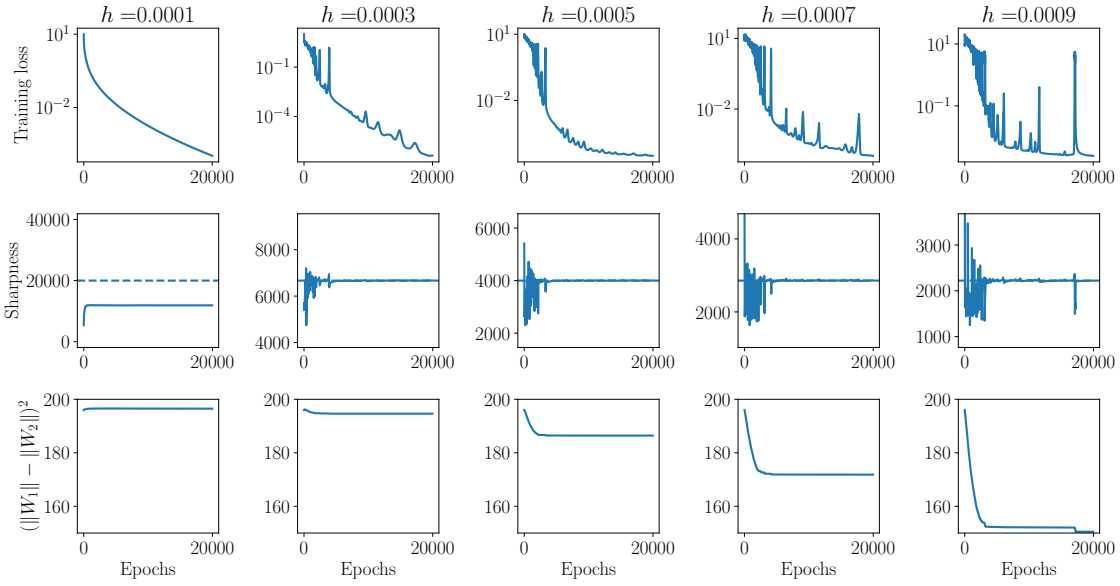


Figure 19: CIFAR-10: huber+ReLU, 2-layer. Each column shows training with h increasing equidistantly from small to near-divergent values. The initial condition is $\|W_1\|_{\mathrm{F}} = 6$ and $\|W_2\|_{\mathrm{F}} = 20$.

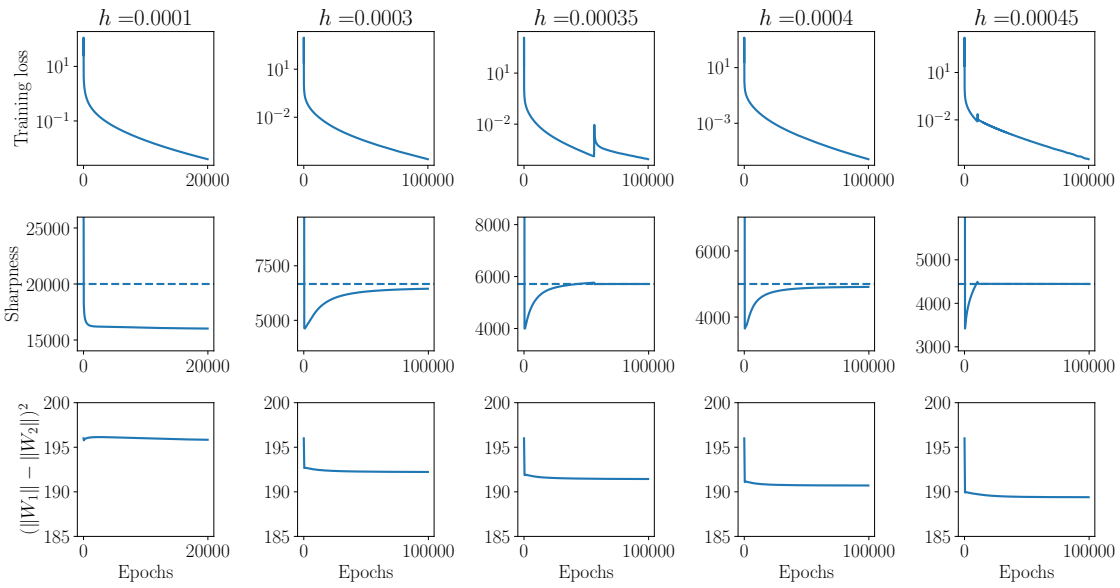


Figure 20: CIFAR-10: ℓ^2+tanh, 2-layer. Each column shows training with h increasing from small to near-divergent values. The initial condition is $\|W_1\|_{\mathrm{F}} = 6$ and $\|W_2\|_{\mathrm{F}} = 20$.

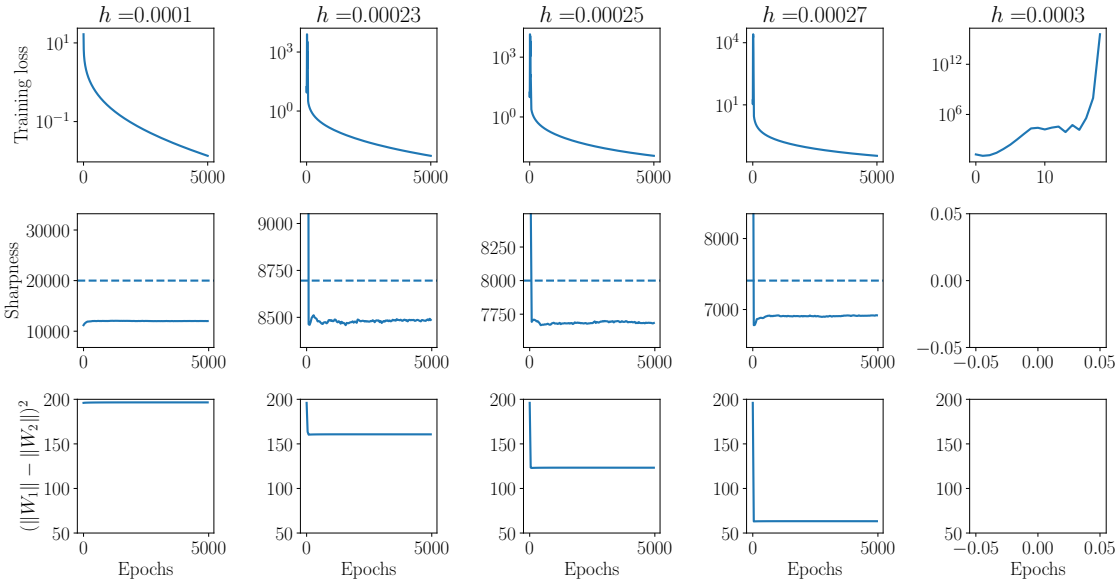


Figure 21: CIFAR-10: ℓ^2+ReLU, 2-layer. Each column shows training with h increasing equidistantly from small to divergent values. The initial condition is $\|W_1\|_{\mathrm{F}} = 6$ and $\|W_2\|_{\mathrm{F}} = 20$.

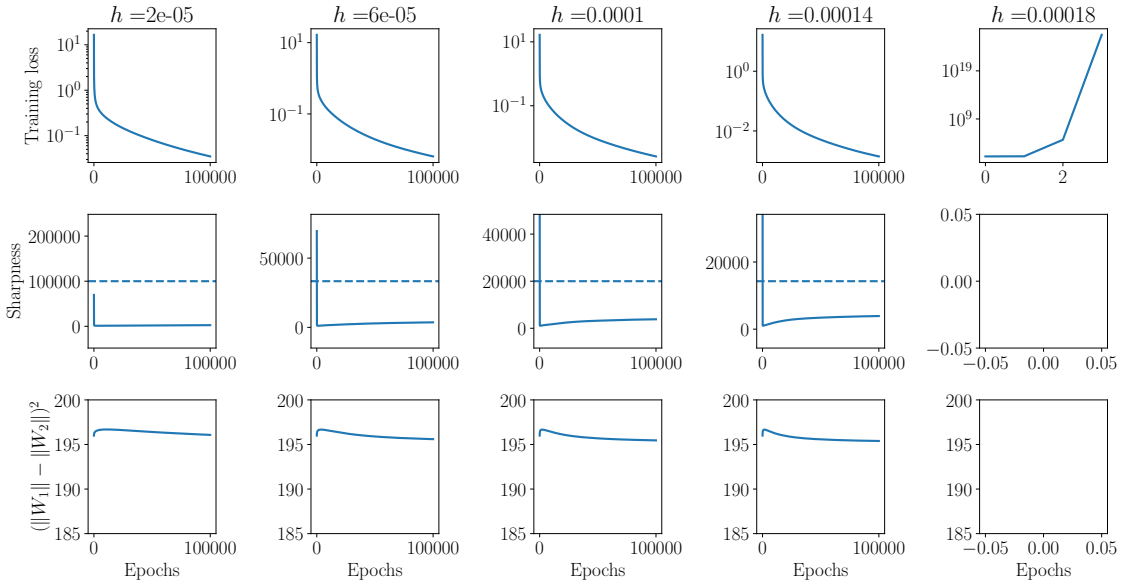


Figure 22: CIFAR-10: ℓ^2+ReLU^3, 2-layer. Each column shows training with h increasing from small to divergent values. The initial condition is $\|W_1\|_{\mathrm{F}} = 6$ and $\|W_2\|_{\mathrm{F}} = 20$.

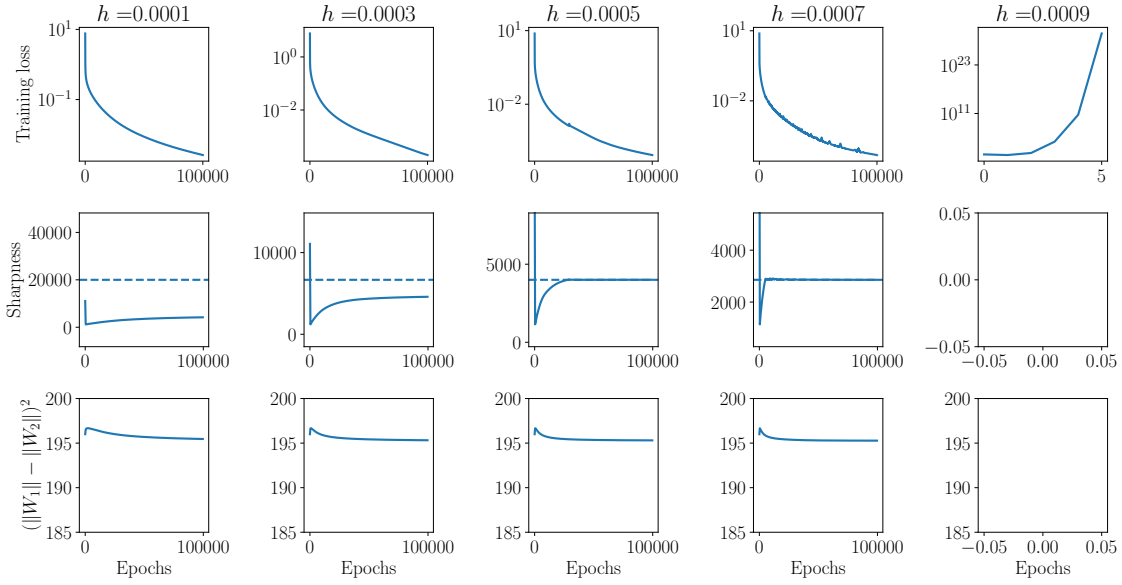


Figure 23: CIFAR-10: ℓ^2+ReLU^3, 2-layer. Each column shows training with h increasing from small to divergent values. The initial condition is $\|W_1\|_{\mathrm{F}} = 6$ and $\|W_2\|_{\mathrm{F}} = 20$.

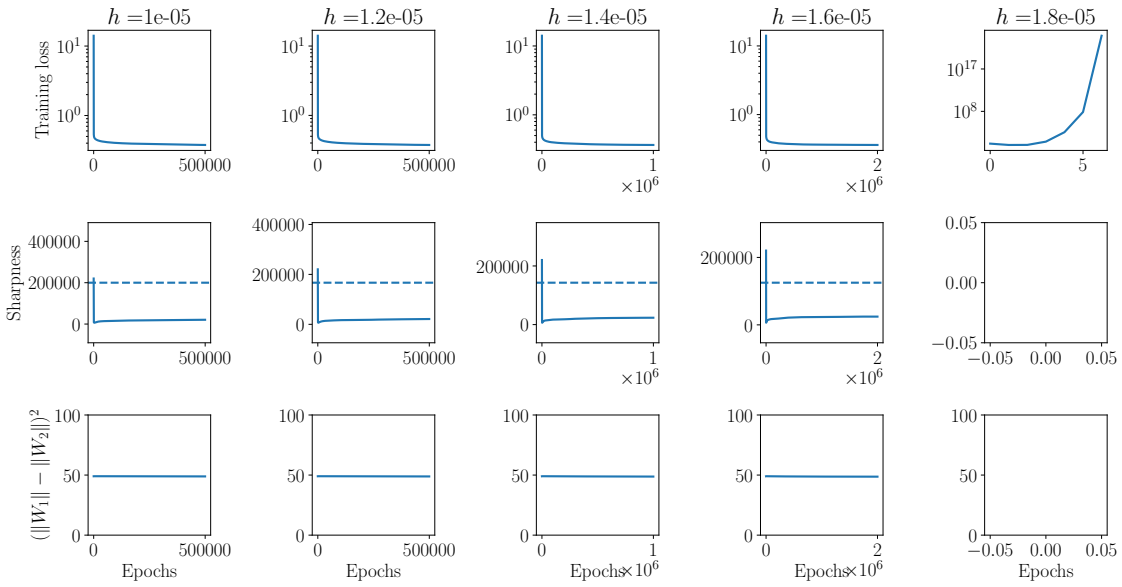


Figure 24: CIFAR-10: ℓ^2+ReLU^3, 3-layer network. Each column shows training with h increasing from small to divergent values. The initial condition is $\|W_1\|_{\mathrm{F}} = 3$ and $\|W_2\|_{\mathrm{F}} = 10$

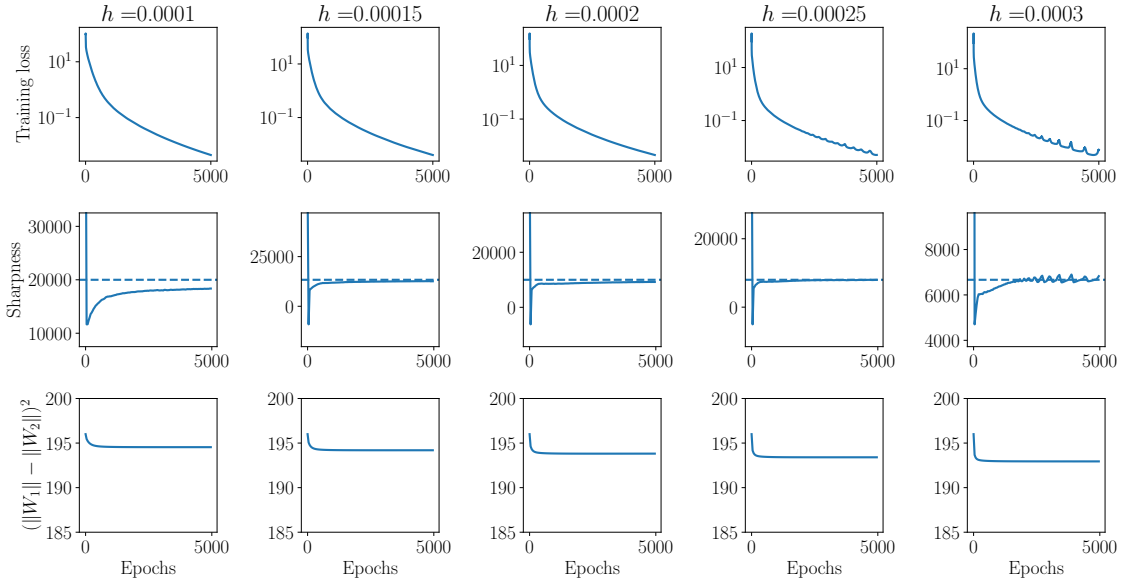


Figure 25: CIFAR-10: ℓ^2+ReLU with batch normalization, 2-layer network. Each column shows training with h increasing from small to near-divergent values.

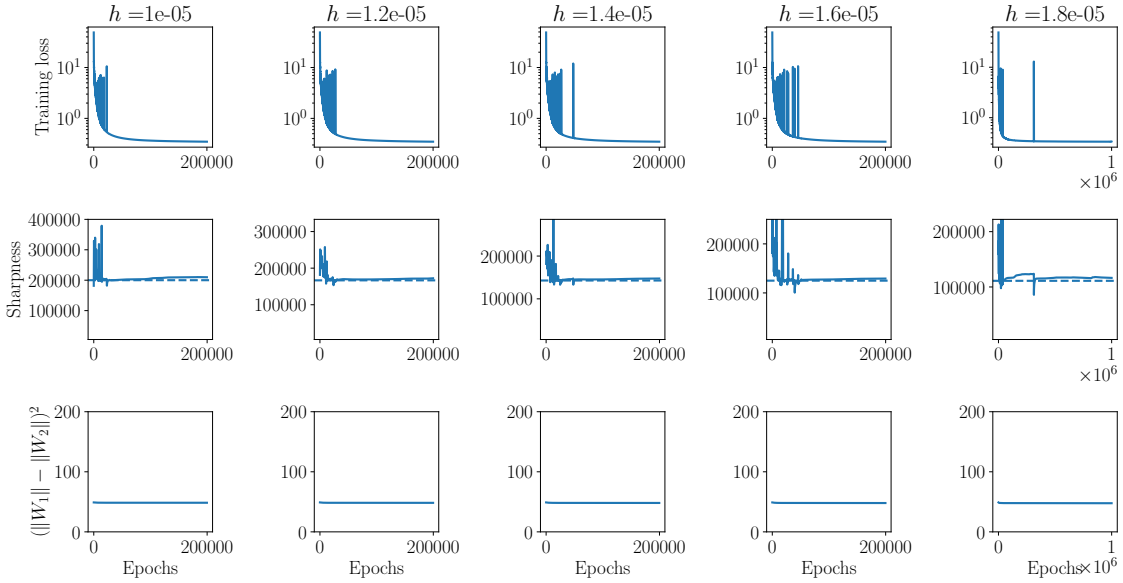


Figure 26: CIFAR-10: huber+ReLU^3 with batch normalization, 3-layer network. Each column shows training with h increasing from small to near-divergent values.

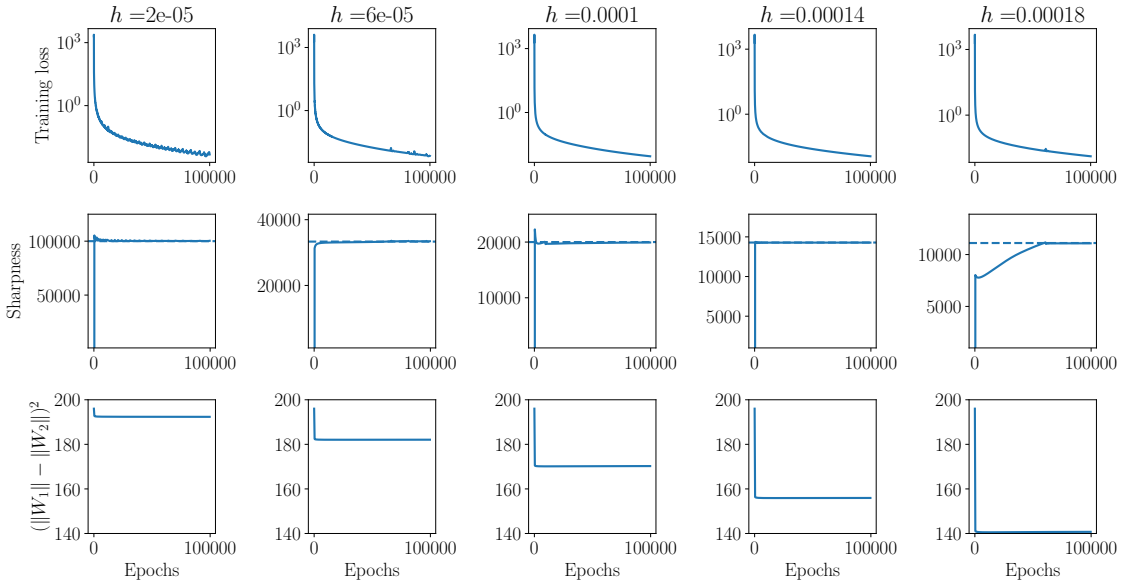


Figure 27: CIFAR-10: ℓ^2+ReLU^3 with batch normalization, 2-layer network. Each column shows training with h increasing from small to near-divergent values.

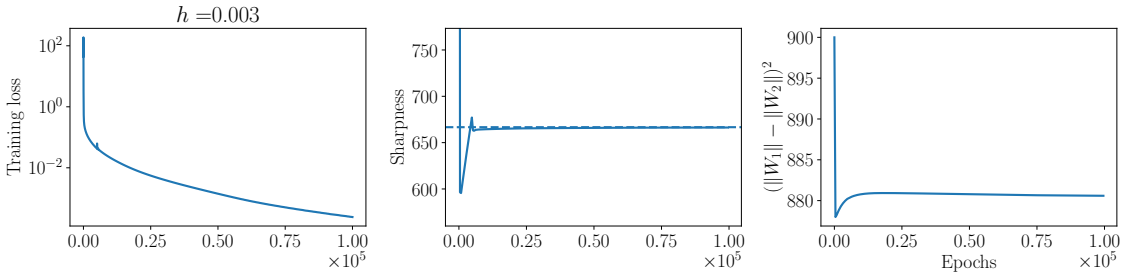


Figure 28: MNIST: huber+tanh, 2-layer. The above shows the training process under one large h. The initialization is $\|W_1\|_F = 10$ and $\|W_2\|_F = 40$.

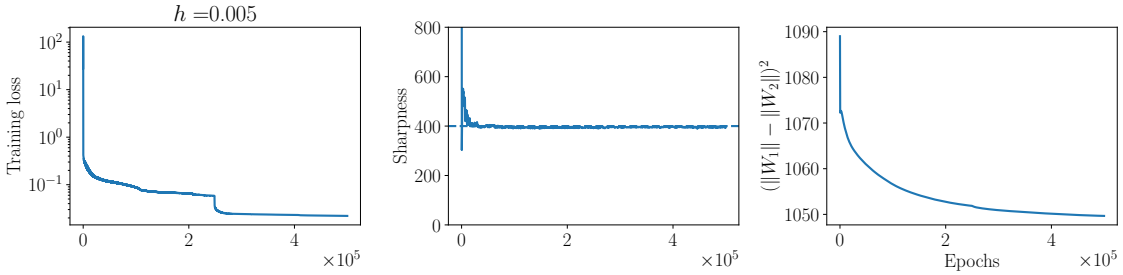


Figure 29: MNIST: huber+ReLU, 2-layer. The above shows the training process under one large h. The initialization is $\|W_1\|_F = 7$ and $\|W_2\|_F = 40$

Nevertheless, to provide further support for our regularity theory, we consider the 3-layer model (5) with a fixed last layer for the combination of ReLU^3 activation and huber loss. In this case, we add a fixed last layer $W_3 \in \mathbb{R}^{N_2 \times N_2}$ whose entries are uniformly sampled from $\{+1, -1\}$. We set σ_1 be the linear activation, $\sigma_2 = \mathrm{ReLU}^3$, and $\mathcal{L} = \text{Huber}$. To prevent the initial sharpness from being too large, we use a slightly smaller initialization where $\|W_1\|_F = 3$ and $\|W_2\|_F = 10$. The result is shown in Figure 24, which is consistent with our claim.

For bad regularity models (degree greater than or equal to 2 in Tab. 1), we also implement experiments with batch normalization to see if the large learning rate phenomena can be recovered. For the cases using 2-layer network (see Figure 2527), we add a standard `BatchNorm1d` layer in PyTorch between the first layer W_1 and the activation σ_1. We set `affine=False` so that the normalization layer would not contain any learnable parameters. For 3-layer network (see Figure 26), it is added between W_2 and σ_2.

G.2 MNIST

The setting in MNIST is similar to CIFAR-10. We again perform experiments on a 1000 sample subset of MNIST. The input dimension now is $N_0 = 1 \times 28 \times 28 = 784$, and the

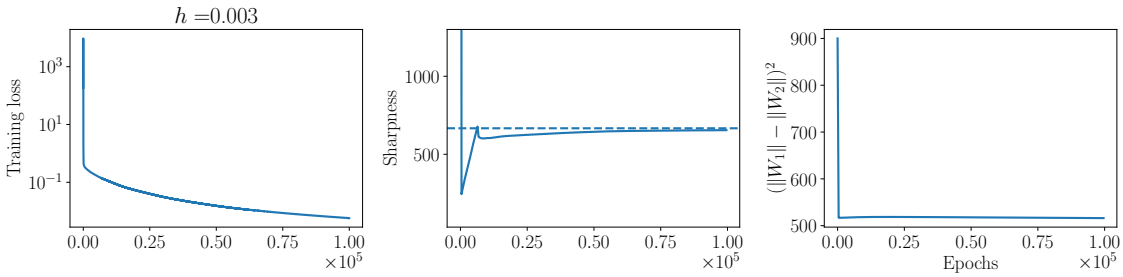


Figure 30: MNIST: ℓ^2+tanh, 2-layer. The above shows the training process under one large h. The initialization is $\|W_1\|_{\mathrm{F}} = 10$ and $\|W_2\|_{\mathrm{F}} = 40$.

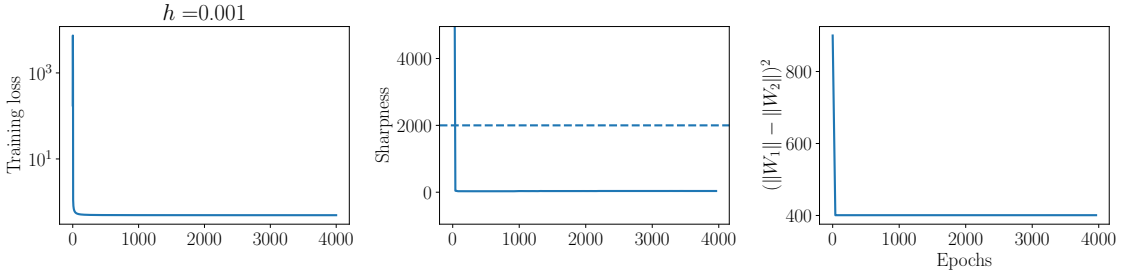


Figure 31: MNIST: ℓ^2+ReLU, 2-layer. The above shows the training process under one large h. The initialization is $\|W_1\|_{\mathrm{F}} = 10$ and $\|W_2\|_{\mathrm{F}} = 40$.

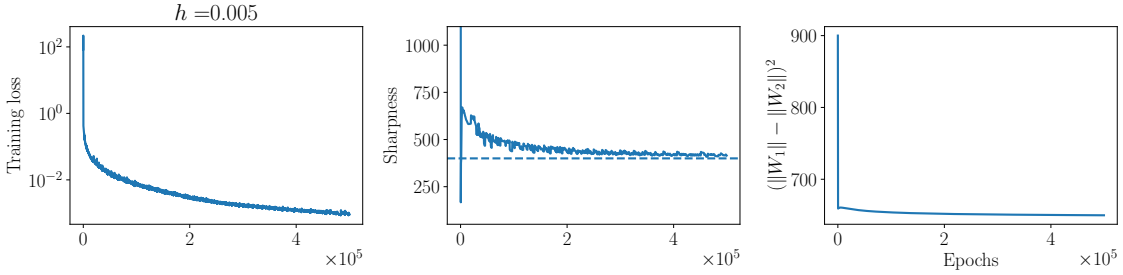


Figure 32: MNIST: huber+ReLU^3, 2-layer. The above shows the training process under one large h. The initialization is $\|W_1\|_{\mathrm{F}} = 10$ and $\|W_2\|_{\mathrm{F}} = 40$.

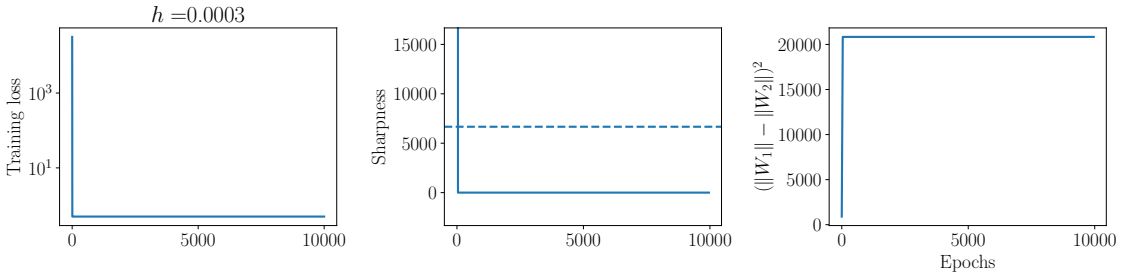


Figure 33: MNIST: huber+ReLU^3, 3-layer. The above shows the training process under one large h. The initialization is $\|W_1\|_{\mathrm{F}} = 6$ and $\|W_2\|_{\mathrm{F}} = 15$

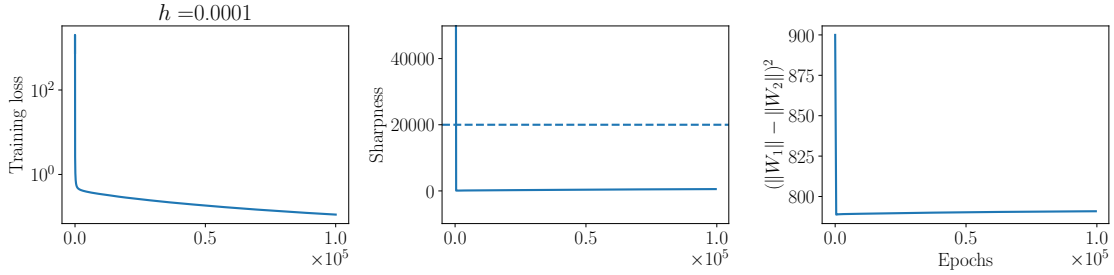


Figure 34: MNIST: ℓ^2+ReLU^3, 2-layer. The above shows the training process under one large h. The initialization is $\|W_1\|_{\mathrm{F}} = 10$ and $\|W_2\|_{\mathrm{F}} = 40$.

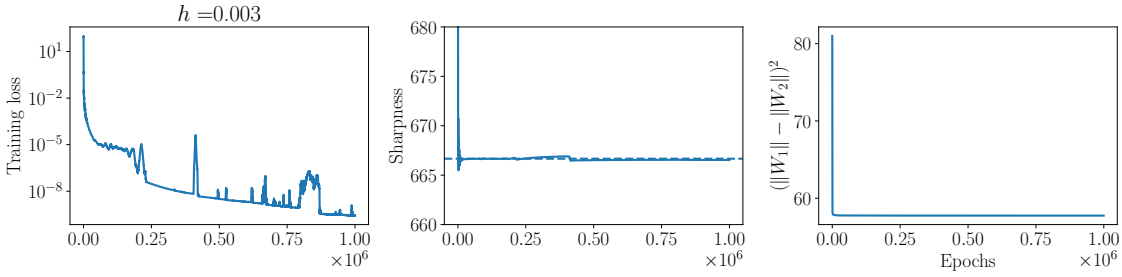


Figure 35: MNIST: huber+ReLU^3 with batch normalization, 3-layer. The above shows the training process under one large h.

output dimension is still $N_2 = 10$. For the weight initialization, we use $\|W_1\|_F = 7$ and $\|W_2\|_F = 40$ for huber+ReLU, $\|W_1\|_F = 10$ and $\|W_2\|_F = 40$ for all the other 2-layer models, and $\|W_1\|_F = 6$ and $\|W_2\|_F = 15$ for the 3-layer model with last layer fixed. All the cases (6 cases under 2-layer network, and huber+ReLU^3 under 3-layer network) exhibit the same results on MNIST (see Figure 28293031323334). We also test batch normalization on an example of bad regularity, huber+ReLU^3 (3-layer), and observe the restoration of large learning rate phenomena (see Figure 35).

References

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. *International Conference on Machine Learning*, pages 247–257, 2022.

Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability. *Advances in Neural Information Processing Systems*, 36:19540–19569, 2023.

Jason Altschuler. *Greed, hedging, and acceleration in convex optimization.* PhD thesis, Massachusetts Institute of Technology, 2018.

Jason M Altschuler and Pablo A Parrilo. Acceleration by stepsize hedging: Multi-step descent and the silver stepsize schedule. *Journal of the ACM*, 72(2):1–38, 2025.

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. *International Conference on Machine Learning*, 2023a.

Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Sgd with large step sizes learns sparse features. *International Conference on Machine Learning*, pages 903–925, 2023b.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. *International Conference on Machine Learning*, pages 244–253, 2018.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. *International Conference on Machine Learning*, pages 948–1024, 2022.

Nils Bjorck, Carla P Gomes, Bart Selman, and Kilian Q Weinberger. Understanding batch normalization. *Advances in neural information processing systems*, 31, 2018.

Lei Chen and Joan Bruna. Beyond the edge of stability via two-step gradient updates. *International Conference on Machine Learning*, pages 4330–4391, 2023.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *International Conference on Learning Representations*, 2021.

Jeremy M Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E Dahl, et al. Adaptive gradient methods at the edge of stability. *arXiv preprint arXiv:2207.14484*, 2022.

Alex Damian, Eshaan Nichani, and Jason D Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. *International Conference on Learning Representations*, 2022.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. *International Conference on Machine Learning*, pages 1019–1028, 2017.

Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in Neural Information Processing Systems*.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *International Conference on Learning Representations*, 2021.

Yihang Gao, Yiqi Gu, and Michael Ng. Gradient descent finds the global optima of two-layer physics-informed neural networks. *International Conference on Machine Learning*, pages 10676–10707, 2023.

Khashayar Gatmiry, Zhiyuan Li, Ching-Yao Chuang, Sashank Reddi, Tengyu Ma, and Stefanie Jegelka. The inductive bias of flatness regularization for deep matrix factorization. *Advances in Neural Information Processing Systems*, 2023.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. *International Conference on Machine Learning*, pages 2232–2241, 2019.

Mariano Giaquinta. Growth conditions and regularity, a counterexample. *manuscripta mathematica*, 59(2), 1987.

Mariano Giaquinta. Quasiconvexity, growth conditions and partial regularity. *Partial differential equations and calculus of variations*, pages 211–237, 1988.

Benjamin Grimmer. Provably faster gradient descent via long steps. *SIAM Journal on Optimization*, 34(3):2588–2608, 2024.

Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, 1993.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning*, pages 448–456, 2015.

Rares Iordan, Marc Peter Deisenroth, and Mihaela Rosca. Investigating the edge of stability phenomenon in reinforcement learning. *arXiv preprint arXiv:2307.04210*, 2023.

Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J Geras. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. *International Conference on Machine Learning*, 139:4772–4784, 18–24 Jul 2021.

Dayal Singh Kalra and Maissam Barkeshli. Phase diagram of early training dynamics in deep neural networks: effect of the learning rate, depth, and width. *Advances in Neural Information Processing Systems*, 2023.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. The normalization method for alleviating pathological sharpness in wide neural networks. *Advances in neural information processing systems*, 32, 2019.

Lingkai Kong and Molei Tao. Stochasticity of deterministic gradient descent: Large learning rate for multiscale objective function. *Advances in Neural Information Processing Systems*, 33:2625–2638, 2020.

Itai Kreisler, Mor Shpigel Nacson, Daniel Soudry, and Yair Carmon. Gradient descent monotonically decreases the sharpness of gradient flow solutions in scalar networks and beyond. *International Conference on Machine Learning*, pages 17684–17744, 2023.

John Langford and Rich Caruana. (not) bounding the true error. *Advances in Neural Information Processing Systems*, 14, 2001.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. *Advances in Neural Information Processing Systems*, 36:40238–40271, 2023.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, 20(1):1474–1520, 2019.

Tao Luo and Haizhao Yang. Chapter 11 - two-layer neural networks for partial differential equations: optimization and generalization theory. volume 25 of *Handbook of Numerical Analysis*, pages 515–554. Elsevier, 2024.

Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. *Advances in Neural Information Processing Systems*, 35:34689–34708, 2022.

Cong Ma, Yuanxin Li, and Yuejie Chi. Beyond procrustes: Balancing-free gradient descent for asymmetric low-rank matrix sensing. *IEEE Transactions on Signal Processing*, 69: 867–877, 2021.

Lachlan Ewen MacDonald, Jack Valmadre, and Simon Lucey. On progressive sharpening, flat minima and generalisation. *arXiv preprint arXiv:2305.14683*, 2023.

Paolo Marcellini. Growth conditions and regularity for weak solutions to nonlinear elliptic pdes. *Journal of Mathematical Analysis and Applications*, 501(1):124408, 2021.

Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit bias of the step size in linear diagonal neural networks. *International Conference on Machine Learning*, 162:16270–16295, 17–23 Jul 2022.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. *Advances in Neural Information Processing Systems*, 32, 2019.

Samet Oymak. Provable super-convergence with a large cyclical learning rate. *IEEE Signal Processing Letters*, 28:1645–1649, 2021.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in neural information processing systems*, 31, 2018.

Sihyeon Seong, Yegang Lee, Youngwook Kee, Dongyoon Han, and Junmo Kim. Towards flatter loss surface via nonmonotonic learning rate scheduling. *Conference on Uncertainty in Artificial Intelligence*, pages 1020–1030, 2018.

Leslie N Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

Minhak Song and Chulhee Yun. Trajectory alignment: Understanding the edge of stability phenomenon via bifurcation theory. *Advances in Neural Information Processing Systems*, 2023.

JG Stampfli and James P Williams. Growth conditions and the numerical range in a banach algebra. *Tohoku Mathematical Journal, Second Series*, 20(4):417–424, 1968.

Yuqing Wang, Minshuo Chen, Tuo Zhao, and Molei Tao. Large learning rate tames homogeneity: Convergence and balancing effect. *International Conference on Learning Representations*, 2022a.

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

Rachel Ward and Tamara Kolda. Convergence of alternating gradient descent for matrix factorization. *Advances in Neural Information Processing Systems*, 36:22369–22382, 2023.

Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. *Advances in Neural Information Processing Systems*, 36:1024–1035, 2023.

Jingfeng Wu, Vladimir Braverman, and Jason D Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. *Advances in Neural Information Processing Systems*, 36:74229–74256, 2023.

Lei Wu, Chao Ma, and Weinan E. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*.

Yunfei Yang and Ding-Xuan Zhou. Optimal rates of approximation by shallow ReLU^k neural networks and applications to nonparametric regression. *Constructive Approximation*, pages 1–32, 2024.

Tian Ye and Simon S Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. *Advances in Neural Information Processing Systems*, 34:1429–1439, 2021.

David Young. On richardson's method for solving linear systems with positive definite matrices. *Journal of Mathematics and Physics*, 32(1-4):243–255, 1953.

Xubo Yue, Maher Nouiehed, and Raed Al Kontar. Salr: Sharpness-aware learning rate scheduler for improved generalization. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *International Conference on Learning Representations*, 2019.

Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. *International Conference on Learning Representations*, 2023.