Analyzing Sharpness along GD Trajectory: Progressive Sharpening and Edge of Stability

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent findings demonstrate that modern neural networks trained by full-batch gradient descent typically enter a regime called Edge of Stability (EOS). In this regime, the sharpness, i.e., the maximum Hessian eigenvalue, first increases to the value 2/(step size) (the progressive sharpening phase) and then oscillates around this value (the EOS phase). This paper aims to analyze the GD dynamics and the sharpness along the optimization trajectory. Our analysis naturally divides the GD trajectory into four phases depending on the change of the sharpness. We empirically identify the norm of output layer weight as an interesting indicator of the sharpness dynamics. Based on this empirical observation, we attempt to theoretically and empirically explain the dynamics of various key quantities that lead to the change of the sharpness in each phase of EOS. Moreover, based on certain assumptions, we provide a theoretical proof of the sharpness behavior in EOS regime in two-layer fully-connected linear neural networks. We also discuss some other empirical findings and the limitation of our theoretical results.

1 Introduction

Deep learning has achieved great success in a variety of machine learning applications, and gradient based algorithms are the prevailing optimization methods for training deep neural networks. However, mathematically understanding the behavior of the optimization methods for deep learning is highly challenging, due to non-convexity, over-parameterization and complicated architectures. In particular, some recent empirical findings in deep networks contradicts our traditional understandings of the gradient methods. For example Wu et al. [2018] observed that the solution found by gradient descent has sharpness approximately equal to $2/\eta$ instead of just smaller than $2/\eta$. Also, Jastrzebski et al. [2020] observed that there is a break-even point in the SGD's trajectory, and after this point there is a regularization effect on the loss curvature.

One recent well-known example is the phenomenon called "Edge of Stability" (EOS) (Cohen et al. [2021]). Based on the classical optimization theory, the learning rate η of gradient based method should be smaller than $2/\lambda$ so that the loss can decrease, where λ is the largest eigenvalue of the Hessian of the objective, also called "sharpness" in the literature. Otherwise, the loss diverges (even for simple quadratic functions). However, the empirical findings in Cohen et al. [2021] show that under various network settings, the EOS phenomena typically occurs along the gradient descent trajectory: (1) the sharpness first increases until it reaches $2/\eta$ (called "progressive sharpening") (2) the sharpness starts hovering around $2/\eta$ (the EOS regime) and (3) the loss non-monotonically decreases without diverging.

Although (1) seems to be consistent with traditional beliefs about optimization, a rigorous mathematical explanation for it is still open. Moreover, the phenomena (2) and (3) are more mysterious because they violate the $\eta < 2/\lambda$ "rule" in traditional optimization theory, yet the training loss does

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

not completely diverge. Instead, the loss may oscillate but still decrease in the long run, while the sharpness seems to be restrained from increasing.

In this paper, we aim to provide a theoretical explanation to the mystery of EOS. Towards the goal, we focus on the dynamics of these key quantities when EOS happens and attempt to theoretically and empirically explain these phenomena along the gradient descent trajectory.

Our contributions can be summarized as follows.

(Section 3.1) We analyze the typical sharpness behavior along the gradient descent trajectory when EOS happens, and propose a four-phase division of GD trajectory, based on the dynamics of some key quantities such as the loss and the sharpness, for further understanding this phenomenon.

(Section 3.2) We empirically identify the weight norm of the output layer as an effective indicator of the sharpness dynamics. We show that analyzing the dynamics of this representative can qualitatively explain the dynamics of sharpness. By assuming this relation, together with some additional simplifying assumptions and approximations, we can explain the dynamics of the sharpness, the loss and the output layer norm in each phase of EOS (Section 3.3). In this context, we also offers an interesting explanation for the non-monotonic loss decrement (also observed in Cohen et al. [2021], Xing et al. [2018]) (Section 3.4).

(Section 4) Following similar ideas, we provide a more rigorous proof for the progressive sharpening and EOS phenomena in a two-layer fully-connected linear neural network setting based on certain assumptions. The assumptions made here are either weaker or arguably less restrictive.

1.1 Related work

The structure of Hessian The Hessian matrix carries the second order information along the training trajectory. Several prior works have empirically found that the spectrum of Hessian has several "outliers" and a continuous "bulk" (Sagun et al. [2016, 2017], Papyan [2018, 2019]). Typically, each outlier corresponds to one class in multi-class classification. As we considered binary classification, there is typical one outlier (i.e., the largest eigenvalue). This is consistent with out Assumption 4.1. The Gauss-Newton decomposition of the Hessian has been used in several prior works Martens [2016], Bottou et al. [2018], Papyan [2018, 2019], and Papyan [2018] empirically showed that the outliers of Hessian can be attribute to a "G component", which is also known as Fisher Information Matrix (FIM) in Karakida et al. [2019a,b]. Also, Wu et al. [2020] characterize the leading Hessian eigenspace by weight matrix and theoretically prove the outliers structure in a random setting.

Neural Tangent Kernel Series of works prove that gradient descent can optimize over-parameterized neural network in NTK regime or lazy training regime (Jacot et al. [2018], Lee et al. [2019], Du et al. [2018, 2019], Arora et al. [2019], Chizat et al. [2019]). A major result in this regime is that when the neural network is wide enough, gradient flow can find the global optimal empirical minimizer very close to the initialization. Moreover, the Hessian does not change much in NTK regime. Our findings go beyond NTK setting to analyze the change of sharpness.

Edge of Stability regime The Edge of Stability phenomena is first formalized by Cohen et al. [2021]. Similar phenomena are identified in Jastrzebski et al. [2020] as the existence of the "break-even" point on SGD trajectory after which loss curvature gets regularized. Xing et al. [2018] observe that gradient descent eventually enters a regime where the iterates oscillate on the leading curvature direction and the loss drops non-monotonically. Very recently Ahn et al. [2022] echo the EOS phenomenon and consider the non-monotonic decreasing behavior of GD which they call unstable convergence. Also, Ma et al. [2022] proposes a special subquadratic landscape property and proves EOS based on this assumption. Lewkowycz et al. [2020] proposes a similar regime called "catapult phase" where loss does not diverge even if the largest Hessian eigenvalue is larger than $2/\eta$. Our work also operates in this regime and assumes that the training is not in the "divergent phase" in Lewkowycz et al. [2020]. Compared with Lewkowycz et al. [2020], we provide a more detailed analysis in more general cases along gradient descent trajectory.

2 Preliminaries

Notations: We denote the training dataset as $\{\mathbf{x}_i, y_i\}_{i=1}^n \subset \mathbb{R}^d \times \{1, -1\}$ and the neural network as $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$. The network $f(\theta, \mathbf{x})$ maps the input $\mathbf{x} \in \mathbb{R}^d$ and parameter $\theta \in \mathbb{R}^p$ to an output

in $\mathbb{R}$. In this paper, we mainly consider the case of binary classification with mean square error (MSE) loss $\ell(z, y) = (z - y)^2$.

Denote the input matrix as $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n) \in \mathbb{R}^{d \times n}$ and the label vector as $\boldsymbol{Y} = (y_1, y_2, ..., y_n) \in \mathbb{R}^n$. We let $\boldsymbol{F}(t) = (f(\mathbf{x}_1), f(\mathbf{x}_2), ..., f(\mathbf{x}_n)) \in \mathbb{R}^n$ and $\boldsymbol{D}(t) = \boldsymbol{F}(t) - \boldsymbol{Y}$ be the (output) prediction vector, and the residual vector respectively at time t. The training objective is:

$$\mathcal{L}(f(\theta)) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\theta, \mathbf{x}_i), y_i) = \frac{1}{n} \sum_{i=1}^{n} (f(\theta, \mathbf{x}_i), y_i)^2.$$

Hessian, Fisher information matrix and NTK: In this part, we apply previous works to illustrate that the largest eigenvalue of Hessian is almost the same as the largest eigenvlue of NTK, where the latter one is the definition we use in this paper. Detailed illustration can be found in Appendix F.

As shown in Papyan [2019], Martens [2016], Bottou et al. [2018], the Hessian can be decomposed into two components, where the term known as "Gauss-Newton matrix", G-term or Fisher information matrix (FIM) dominates the other term. Meanwhile, Karakida et al. [2019b] pointed out the duality between the FIM and a Gram matrix $\boldsymbol{M}$, defined as

$$\boldsymbol{M} = \frac{2}{n} \frac{\partial \boldsymbol{F}(\theta)}{\partial \theta} \frac{\partial \boldsymbol{F}(\theta)}{\partial \theta}^{\top} \tag{1}$$

It also known as the neural tangent kernel NTK (Karakida et al. [2019b,a]), which has been studied extensively in recent years (Du et al. [2018],Arora et al. [2019]). Note that in this paper, we do not assume the training is in NTK regime (or the lazy training regime Jacot et al. [2018]).

In this paper, we use θ_t to denote the parameter at iteration t (or time t) and the sharpness at time t as $\Lambda(t) = \Lambda(\theta_t)$. We similarly define $\boldsymbol{M}(t), \boldsymbol{F}(t), \boldsymbol{D}(t)$.

Here we show the gradient flow dynamics of the residual vector $\boldsymbol{D}(t)$:

$$\frac{\mathrm{d}\boldsymbol{D}(t)}{\mathrm{d}t} = \frac{\partial \boldsymbol{D}(t)}{\partial \theta} \frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t} = -\frac{\partial \boldsymbol{F}(t)}{\partial \theta} \frac{\partial \mathcal{L}(t)}{\partial \theta} = -\frac{2}{n} \frac{\partial \boldsymbol{F}(t)}{\partial \theta} \frac{\partial \boldsymbol{F}(t)}{\partial \theta}^{\top} \boldsymbol{D}(t) = -\boldsymbol{M}(t)\boldsymbol{D}(t) \tag{2}$$

3 A Four-phase Analysis of GD Dynamics

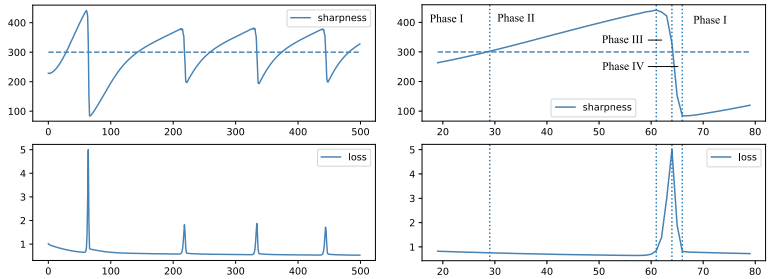
In this section, we study the dynamics of gradient descent and the change of sharpeness along the optimizaiton trajectory. We divide the whole training process into four phases, occurring repeatedly in the EOS regime. In Section 3.1, we introduce the four phases. In Section 3.2, we show empirically that the change of the norm of output layer weight vector almost coincides with the change of the sharpness. In Section 3.3, using this observation, we attempt to explain the dynamics of each phase, and provide mathematical explanation for the changes of the sharpness. In Section 3.4, we explain why the the loss decreases but non-monotonically. We admit that a completely rigorous theoretical explanation is still beyond our reach and much of our argument is based various simplifying assumptions, and is somewhat heuristic at some points.

3.1 A Four-phase Division

To further understand the properties along the trajectory when EOS happens, we study the behaviors of the loss and sharpness during the training process. As illustrated in Figure 1, We train a shallow neural network by gradient descent on a subset of 1,000 samples from CIFAR-10 (Krizhevsky et al. [2009]), using the MSE loss as the objective. Notice that the sharpness keep increasing while the loss is decreasing until the sharpness reaches $2/\eta$. Then the sharpness begins to oscillate around $2/\eta$ while the loss decreases non-monotonically. This is a typical sharpness behavior in the EOS regime, and consistent with the experiments in prior work Cohen et al. [2021].

We divide the training process into four phases according to the evolution of the loss, sharpness and their correlation, as shown in Figure 1. The four phases happens in a cyclic manner along the training trajectory. We first briefly describe the properties of each phase, and explain the dynamics in more details in Section 3.3.

Phase I: Sharpness $\Lambda < 2/\eta$. In this stage, all the eigenvalues of Gram matrix $\boldsymbol{M}$ are below the threshold $2/\eta$. In particular, using standard initialization, the training starts from this phase, and



(a) Sharpness and Loss (b) Phase Division Illustration

Figure 1: In Figure (a), we show the sharpness and the loss when training on a two-hidden-layer linear activated network by gradient descent. Shortly after the sharpness crosses $2/\eta$, it drops quickly back to some value below $2/\eta$. In Figure (b) we illustrate how the four phases are divided according to the evolution of the sharpness.

during this phase, the loss keeps decreasing and the sharpness keeps growing along the trajectory. This initial phase is called *progressive sharpening (PS)* in prior work Cohen et al. [2021]. Empirically, the behavior of GD trajectory (as well as the loss and the sharpness) are very similar to that of gradient flow, until the sharpness reaches $2/\eta$ (This phenomena is also observed in Cohen et al. [2021]. See Figure 5 or Appendix J.1 in Cohen et al. [2021]). We note that GD may come back to this phase from Phase IV later.

Phase II: Sharpness $\Lambda > 2/\eta$. In this phase, the sharpness exceeds $2/\eta$ and may keeps increasing. We will show that the fact that $\Lambda > 2/\eta$ makes $|\boldsymbol{D}^\top \boldsymbol{v}_1|$ (where $\boldsymbol{v}_1$ the first eigenvector of $\boldsymbol{M}$) increase exponentially (Lemma 3.2). This would quickly causes $\|\boldsymbol{D}\|$ to exceed $\|\boldsymbol{Y}\|$ in a few iterations, which leads the sharpness to start decreasing by proposition 3.1, hence the training process enters Phase III.

Phase III: Sharpness $\Lambda > 2/\eta$ yet begins to gradually drop. Before Λ drops below $2/\eta$, lemma 3.2 still holds, so $|\boldsymbol{D}^\top \boldsymbol{v}_1|$ keeps increasing. Proposition 3.1 still holds and thus the sharpness keeps decreasing until it is below $2/\eta$, at which point we enter Phase IV. A distinctive feature in this phase is that the loss may increase due to the exponential increase of $|\boldsymbol{D}^\top \boldsymbol{v}_1|$.

Phase IV: Sharpness $\Lambda < 2/\eta$. When the sharpness is below $2/\eta$, $|\boldsymbol{D}^\top \boldsymbol{v}_1|$ begins to decrease quickly, leading the loss to decrease quickly. At the same time, the sharpness keeps oscillating and gradually decreasing for some iterations. This lasts until the loss decrease to a level that is around its value right before Phase III. The sharpness is still below $2/\eta$ and gets back to Phase I.

3.2 The Norm of the Output Layer Weight

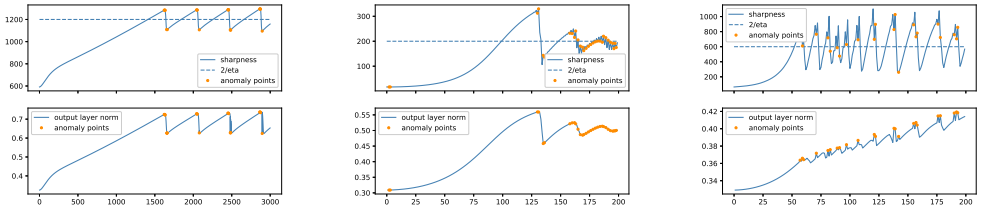
It is difficult to rigorously analyze the dynamics of the sharpness $\Lambda(t)$. In this subsection, we make an interesting observation, that the change of the norm of the output layer of the network (usually a fully-connected linear layer) is consistent with the change of the sharpness most of the time.

In particular, for a general neural network $f(\mathbf{x}) = \boldsymbol{A}^\top \boldsymbol{h}(\boldsymbol{W}, \mathbf{x})$, where $\boldsymbol{A} \in \mathbb{R}^m$ is the output layer weight and the feature extractor $\boldsymbol{h} : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}^m$ outputs a m-dimensional feature vector. $\boldsymbol{W} \in \mathbb{R}^p$ is the parameter vector or the extractor $\boldsymbol{h}$. Note that $\boldsymbol{M} = (\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{\theta}})^\top (\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{\theta}})$ can be decomposed as follows: $\boldsymbol{M} = \left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{\theta}}\right)^\top = \left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{A}}\right)\left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{A}}\right)^\top + \left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{W}}\right)\left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{W}}\right)^\top := \boldsymbol{M}_{\boldsymbol{A}} + \boldsymbol{M}_{\boldsymbol{W}}$ where the $(i,j)-$entry of $\boldsymbol{M}_{\boldsymbol{W}}$ is $(\boldsymbol{M}_{\boldsymbol{W}})_{ij} = \left\langle \frac{\partial f(\mathbf{x}_i)}{\partial \boldsymbol{W}}, \frac{\partial f(\mathbf{x}_j)}{\partial \boldsymbol{W}} \right\rangle = \boldsymbol{A}^\top \frac{\partial \boldsymbol{h}(\boldsymbol{W}, \mathbf{x}_i)}{\partial \boldsymbol{W}} \frac{\partial \boldsymbol{h}(\boldsymbol{W}, \mathbf{x}_j)}{\partial \boldsymbol{W}}^\top \boldsymbol{A}$.

In this expression, intuitively $\|\boldsymbol{A}\|$ should be positively related to $\|\boldsymbol{M}_{\boldsymbol{W}}\|$. We empirically observe that the part $\boldsymbol{M}_A = (\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{A}})(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{A}})^\top$ has a much smaller spectral norm compared to the whole Gram matrix $\boldsymbol{M}$ (see Figure 3(a)), which means $\|\boldsymbol{M}_{\boldsymbol{W}}\|$ dominates $\boldsymbol{M}_A$ in spectral norm. Therefore, $\|\boldsymbol{A}\|$ should be positively correlated with $\|\boldsymbol{M}\|$.

The benefit of analyzing $\|\boldsymbol{A}\|^2$ is that the gradient flow of $\|\boldsymbol{A}\|^2$ enjoys the following clean formula:

$$\frac{d\|\boldsymbol{A}\|^2}{dt} = -2\left(\frac{\partial \mathcal{L}}{\partial \boldsymbol{A}}\right)^\top \boldsymbol{A} = -\frac{2}{n}\boldsymbol{D}^\top \left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{A}}\right) \boldsymbol{A} = -\frac{4}{n}\boldsymbol{D}^\top \boldsymbol{F} \tag{3}$$



(a) Two-layer Linear Network (b) Deep 5-layer tanh network (c) Resnet18

Figure 2: Sharpness and A norm are highly correlated; *Anomaly points* (orange points) in the figure is the iterations when $\text{sign}(\Lambda(t+1) - \Lambda(t)) \neq \text{sign}(\|\boldsymbol{A}(t+1)\| - \|\boldsymbol{A}(t)\|)$. More discussions are in section 5 and Appendix B.

In this work, we do experiments on two-layer linear network, fully connected deep neural networks, Resnet18, and all of them has this output layer structure. From Figure 3(a), we can observe that **the output layer norm $\|\boldsymbol{A}\|^2$ and the sharpness Λ change in the same direction** most of the time along the gradient descent trajectory, i.e., they both increase or decrease at the same time. We note that they may change in different directions very occasionally around the time when $\|\boldsymbol{A}(t+1)\|^2 - \|\boldsymbol{A}(t)\|^2$ changes signs (see the experiments in Figure 2).

3.3 Detailed Analysis of Each Phase

In this section, we explain the dynamics of each phase in more details. We consider the dynamics of $\boldsymbol{D}$ and $\|\boldsymbol{A}\|^2$ by their first order approximations. According to (2) and (3) and ignoring higher order terms, we have the following approximate update rules of $\boldsymbol{D}$ and $\|\boldsymbol{A}\|^2$:

$$\boldsymbol{D}(t+1) - \boldsymbol{D}(t) \approx -\eta \boldsymbol{M}(t)\boldsymbol{D}(t), \ \|\boldsymbol{A}(t+1)\|^2 - \|\boldsymbol{A}(t)\|^2 \approx -\frac{4\eta}{n}\boldsymbol{D}(t)^\top \boldsymbol{F}(t) \tag{4}$$

In each phase, we attempt to explain the main driving force of the change of the sharpness and the loss. For different phases, we may need some different assumptions to simplify the arguments. Most of the assumptions are quite consistent with the experiments or the findings in the literature. Some of them are somewhat stronger, and we also discuss how to relax them. All the proof can be found in Appendix E.

Assumption 3.1. *If the training does not diverge, there exists some constant B_Λ, such that $0 < \Lambda(t) \leq \frac{B_\Lambda}{\eta}$ for all t.*

This assumption states that there must be an upper bound of the sharpness. Actually in Lewkowycz et al. [2020], they proved that $4/\eta$ is a upper bound of the sharpness in two-layer linear network with one datapoint, otherwise the training process would diverge. They empirically found that similar upper bounds exist also for nonlinear activations, albeit with somewhat larger constants B_Λ.

Phase I: In this phase, we show that $\boldsymbol{D}(t)^\top \boldsymbol{F}(t) < 0$ under certain assumptions (detailed shortly) on the spectral properties of $\boldsymbol{M}(t)$ (see Lemma 3.1 below). By equation (4), we have $\|\boldsymbol{A}(t+1)\|^2 - \|\boldsymbol{A}(t)\|^2 > 0$, implying that the sharpness $\Lambda(t)$ also increases based on our observation in Section 3.2. This phase stops if $\Lambda(t)$ grows larger than $2/\eta$.

We assume the output vector $\boldsymbol{F}(t)$ is initialized to be small (this is true if we use very small initial weights). For simplicity, we assume $\boldsymbol{F}(0) = 0$ in the following argument. Next, we make some assumptions on the eigenvectors $\{\boldsymbol{v}_i(t)\}_{i=1}^n$ of $\boldsymbol{M}(t)$, as well as the gradient descent trajectory.

Assumption 3.2. *The set of eigenvectors $\{\boldsymbol{v}_i\}_{i=1}^n$ of $\boldsymbol{M}(t)$ are fixed throughout this phase.*

Assumption 3.3. *When $\Lambda(t) < 2/\eta$, $\boldsymbol{D}(t)$ follows the gradient flow trajectory: $\frac{\mathrm{d}\boldsymbol{D}(t)}{\mathrm{d}t} = -\boldsymbol{M}(t)\boldsymbol{D}(t)$.*

Assumption 3.2 on the eigenvectors is somewhat strong, and the eigenvectors corresponding to small eigenvalues may change drastically in our experiments. It is possible to relax this assumption with some weaker assumptions on the spectrum of $\boldsymbol{M}$ (see the proof of Theorem 1 for 2-layer linear network). Moreover, we note that $\boldsymbol{v}_1(t)$ (the eigenvectors corresponding to the largest eigenvalue)

changes much slowly and the inner product of its initially direction and its direction at the end of the phase is also large (See Appendix D.).

Assumption 3.3 holds empirically (see Figure 5 or Appendix J.1 in Cohen et al. [2021]) when the networks are continuously differentiable. We include these experiment details in our Appendix D.

Lemma 3.1. *For all t in Phase I, under Assumption 3.2 and 3.3, it holds that $\boldsymbol{D}(t)^\top \boldsymbol{F}(t) < 0$.*

Now we can see, the sharpness will keep increasing until it reaches $2/\eta$ or the loss converges to 0. In the former case the training process enters Phase II, while the latter case is also possible when η is really small.

Phase II: When the training process just enters Phase II, the sharpness keeps increasing. We show shortly that $\boldsymbol{D}(t)^\top \boldsymbol{v}_1(t)$ starts to increase geometrically, and this will cause the sharpness to stop increasing at some point, thus entering Phase III.

Assumption 3.4. *There is a small constant $\epsilon_2 \geq 0$ such that $\langle \boldsymbol{v}_1(t), \boldsymbol{v}_1(t+1) \rangle \geq 1 - \epsilon_2$ for all t in this phase.*

Assumption 3.5. *Assume $\boldsymbol{D}(t)^\top \boldsymbol{v}_1(t) \geq c\epsilon_2 \|\boldsymbol{D}(t)\|$ for some $c > 1$ for some $t = t_0$ at the beginning of this phase.*

Assumption 3.4 says the direction of the first eigenvector changes slowly[1], which is a weaker assumption than Assumption 3.2. Assumption 3.5 says $\boldsymbol{D}(t)$ has a non-negligible component in the direction of $\boldsymbol{v}_1$. Since ϵ_2 is small, this is not a strong assumption as some small perturbation (due to discrete updates) would make the assumption hold for some $c > 1$.

Lemma 3.2. *Suppose Assumption 3.4 and 3.5 hold during this phase (with constants $\epsilon_2 > 0$ and $c > 1$). If $\Lambda(t) = (2+\tau)/\eta$ and $\tau > \frac{1}{1-\epsilon_2-1/c} - 1$, then $\boldsymbol{D}(t)^\top \boldsymbol{v}_1(t)$ increases geometrically with factor $(1+\tau)(1-\epsilon_2-1/c) > 1$ for $t \geq t_0$ in this phase.*

Since $\boldsymbol{D}(t)^\top \boldsymbol{v}_1(t)$ increases geometrically, $\|\boldsymbol{D}\| \geq \boldsymbol{D}(t)^\top \boldsymbol{v}_1(t)$ will exceed $\|\boldsymbol{Y}\|$ eventually. Next, we prove the following proposition which states that when this happens, $\boldsymbol{D}(t)^\top \boldsymbol{F}(t) > 0$. As a consequence, $\|\boldsymbol{A}\|$ decreases, and the sharpness decreases at the same time based on our assumption.

Proposition 3.1. *If $\|\boldsymbol{D}(t)\| > \|\boldsymbol{Y}\|$, then $\boldsymbol{D}(t)^\top \boldsymbol{F}(t) > 0$.*

Proof. This proposition can be proved by simply noticing that $\boldsymbol{D}(t)^\top \boldsymbol{F}(t) = \boldsymbol{D}(t)^\top (\boldsymbol{D}(t) - \boldsymbol{Y}(t)) \geq \|\boldsymbol{D}\|^2 - \|\boldsymbol{D}\|\|\boldsymbol{Y}\| > 0$ when $\|\boldsymbol{D}(t)\| \geq \|\boldsymbol{Y}\|$. □

Phase III: The sharpness is still larger than $2/\eta$, but it starts decreasing. Meanwhile, the loss continues to increase rapidly due to Lemma 3.2. Eventually, the sharpness will fall below $2/\eta$ and then the training process enters phase IV.

For the same reasoning in Phase II, if the sharpness stays above $2/\eta$, then we can have arbitrarily large loss. According to Proposition 3.1, if the loss is large enough, the sharpness keeps decreasing.

Now we show that if the sharpness stays above $2/\eta$, $\|\boldsymbol{A}(t)\|^2$ will decrease by a significant amount. This partially explains that the sharpness should also decrease significantly until it drops below $2/\eta$ (instead of decreasingly converging to a value above $2/\eta$ without ever entering the next phase). In fact, using (3) and noticing that $\boldsymbol{D}(t)^\top \boldsymbol{F}(t) \geq \|\boldsymbol{D}(t)\|^2 - \|\boldsymbol{D}(t)\|\|\boldsymbol{Y}\| > (\|\boldsymbol{D}(t)\| - \|\boldsymbol{Y}\|)^2$, we can obtain the following proposition.

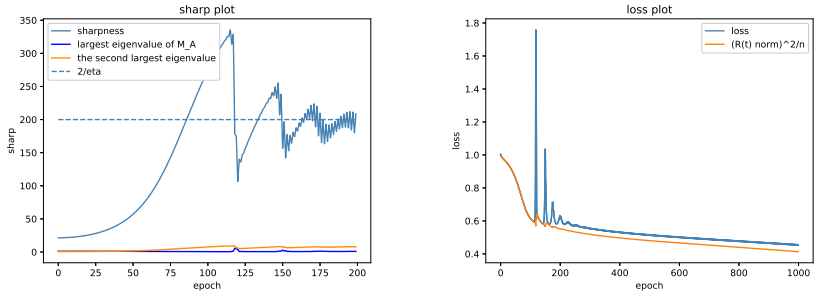
Proposition 3.2. *If $\|\boldsymbol{D}(t)\| > \|\boldsymbol{Y}\|$, then $\|\boldsymbol{A}(t+1)\|^2 - \|\boldsymbol{A}(t)\|^2 < (\|\boldsymbol{D}(t)\| - \|\boldsymbol{Y}\|)^2$*

From the above argument, we can see that if $\boldsymbol{D}(t)^\top \boldsymbol{v}_1(t)$ is larger than $\|\boldsymbol{Y}\|$, then $\boldsymbol{D}(t)^\top \boldsymbol{v}_1(t)$ does not decrease in Phase III, and according to Proposition 3.2, $\|\boldsymbol{A}(t)\|^2$ decreases significantly, implying the sharpness drops below $2/\eta$ eventually.

Remark. The fact that the sharpness can surely drop below $2/\eta$ in this phase can be proved more rigorously in Section 4 for the two layer linear setting. See Theorem 2.

Phase IV: First, since the training process has just left phase III, $\boldsymbol{D}(t)^\top \boldsymbol{F}(t)$ is still positive and large, hence $\|\boldsymbol{A}(t)\|^2$ keeps decreasing and the sharpness decreases as well. Since the sharpness

[1] This assumption holds in our experiments (In a 5-layer tanh network with width $m = 200$, $\eta = 2/200$, during training 800 iterations, the maximum value of ϵ_2 is about 0.0093).



(a) $\lambda_{\max}(\boldsymbol{M})$ (sharpness), $\lambda_{\max}(\boldsymbol{M_A})$, $\lambda_2(\boldsymbol{M})$. (b) The loss ($\|\boldsymbol{D}(t)\|^2/n$) and $\|\boldsymbol{R}(t)\|^2/n$

Figure 3: In Figure (a), we verify the assumptions in Section 3.2 and 3.3 empirically. The orange line: the second largest eigenvalue is far below $1/\eta$, illustrating Assumption 3.6. The blue line: the largest eigenvalue of $\boldsymbol{M_A}$ is negligible compared to the sharpness. In Figure (b), we show that even though the total loss $\mathcal{L}(t) = \|\boldsymbol{D}(t)\|^2/n$ oscillates considerably, $\|\boldsymbol{R}(t)\|^2/n$ decreases more steadily.

stays below $2/\eta$, the loss decreases due to the following lemma (with $\boldsymbol{u}$ replaced by $\boldsymbol{D}(t)$, proof see Appendix E).

Lemma 3.3. *If $\Lambda(t) < 2/\eta$, then for any vector $\boldsymbol{u} \in \mathbb{R}^n$, $\|\boldsymbol{u}^\top(\boldsymbol{I} - \eta\boldsymbol{M}(t))\| \le (1-\eta\alpha)\|\boldsymbol{u}\|$, where $\alpha = \min\{2/\eta - \Lambda(t), \lambda_{\min}(\boldsymbol{M}(t))\}$. In particular, replacing $\boldsymbol{u}$ with $\boldsymbol{D}(t)$, we can see $\|\boldsymbol{D}(t+1)\| \le (1-\eta\alpha)^2\|\boldsymbol{D}(t)\|$.*

Next we prove that $\boldsymbol{D}(t)^\top\boldsymbol{F}(t)$ will become negative eventually, which indicates that $\|\boldsymbol{A}(t)\|^2$ and hence the sharpness will grow again. Since the sharpness is below $2/\eta$, $\boldsymbol{D}(t)^\top\boldsymbol{v}_1(t)$ does not blow up. Moreover, the overall loss decreases due to Lemma 3.3. In fact, as we show in the next subsection, $\boldsymbol{D}(t)$ excluding the $\boldsymbol{v}_1(t)$ component closely follows a gradient flow trajectory with sharpness always under $1/\eta$. So it decreases as in Phase I for this component, and as the $\boldsymbol{v}_1(t)$ component of $\boldsymbol{D}$ also decreases, $\boldsymbol{D}(t)^\top\boldsymbol{F}(t)$ will become negative again, which means $\|\boldsymbol{A}\|$ and the sharpness will grow again. Then, the training process re-enters Phase I and a new cycle begins.

3.4 Explaining Non-monotonic Loss Decrement

In this subsection, we attempt to explain the non-monotonic decrease of the loss during the entire GD trajectory. See Figure 3(b). To this end, we decompose $\boldsymbol{D}$ into the $\boldsymbol{v}_1$-component $\boldsymbol{v}_1\boldsymbol{v}_1^\top\boldsymbol{D}$ and the remaining part $\boldsymbol{R}$ defined as $\boldsymbol{R}(t) := (\boldsymbol{I} - \boldsymbol{v}_1(t)\boldsymbol{v}_1(t)^\top)\boldsymbol{D}(t)$. Below, we prove that $\boldsymbol{R}(t)$ is not affected much by the exponential growth of the loss (Proposition 3.5) in Phase II and III, and almost keeps its trajectory as in Phase I. All the proof of Lemmas can be found in Appendix E.

The arguments in this subsection need Assumption 3.1, Assumption 3.4, both very consistent with the experiments. We need an additional assumption on the spectrum of $\boldsymbol{M}$.

Assumption 3.6. *All $\boldsymbol{M}(t)$'s eigenvalues except $\Lambda = \lambda_{\max}(\boldsymbol{M}(t))$ are no larger than $1/\eta$ for all t.*

Assumption 3.6 states that the largest eigenvalue is an outlier in the spectrum. Empirically, it is true. See Figure 3(b). Similar fact is also mentioned in Sagun et al. [2016, 2017].

First, we let B_D be an upper bound of $\boldsymbol{D}(t)$, i.e., for all t, $\|\boldsymbol{D}(t)\| \le B_D$. In the two layer linear network case, we can have an explicit form of B_D (see Conclusion 3 in Theorem 2). Recall in Assumption 3.1, B_Λ is the upper bound of $\eta\Lambda$.

Lemma 3.4. *Suppose Assumption 3.4 holds. $\boldsymbol{R}(t)$ satisfies the following:*

$$\boldsymbol{R}(t+1) = (I - \eta\boldsymbol{M})\boldsymbol{R}(t) + \boldsymbol{e}_1(t), \quad \textit{where} \quad \|\boldsymbol{e}_1(t)\| \le 3\sqrt{\epsilon_2}\|\boldsymbol{D}(t)\|(B_\Lambda - 1)$$

Lemma 3.5. *Define an auxiliary sequence $\boldsymbol{R}'(t)$ by $\boldsymbol{R}'(0) = \boldsymbol{R}(0)$, $\boldsymbol{R}'(t+1) = (\boldsymbol{I} - \eta\boldsymbol{M}(t))\boldsymbol{R}'(t)$. If Assumption 3.1, Assumption 3.4, Assumption 3.6 hold, and for any time t there exists a quantity $\lambda_r > 0$, such that the smallest eigenvalue of $\boldsymbol{M}(t)$, i.e. $\lambda_{\min}(\boldsymbol{M}(t)) > \lambda_r$, then there exists a constant $c_r > 0$ such that $\|\boldsymbol{R}(t) - \boldsymbol{R}'(t)\| \le c_r\frac{B_D(B_\Lambda - 1)\sqrt{\epsilon_2}}{\eta\lambda_r}$.*

From the definition of $\boldsymbol{R}'(t)$, we can use the same argument as in Phase I to see that $\boldsymbol{R}'(t)^\top(\boldsymbol{R}'(t) + \boldsymbol{Y}(\boldsymbol{I} - \boldsymbol{v}_1(t)\boldsymbol{v}_1(t)^\top)) < 0$ (the left hand side is simple $\boldsymbol{D}^\top\boldsymbol{F}$ with $\boldsymbol{v}_1$ components removed). Since $\|\boldsymbol{R}(t) - \boldsymbol{R}'(t)\|$ is sufficiently small, we have similar result for $\boldsymbol{R}(t)$.

Now, in light of Lemma 3.2 and Lemma 3.5, we arrive at an interesting explanation of the phenomena of non-monotonic decrease of the loss. Basically, $\boldsymbol{D}$ can be decomposed into the $\boldsymbol{v}_1$-component $\boldsymbol{v}_1\boldsymbol{v}_1^\top\boldsymbol{D}$ and the remaining part $\boldsymbol{R} = (\boldsymbol{I} - \boldsymbol{v}_1\boldsymbol{v}_1^\top)\boldsymbol{D}$. The $\boldsymbol{v}_1$-component may increase geometrically during the EOS (Lemma 3.2), but the behavior of the remaining part $\boldsymbol{R}(t)$ is close to $\boldsymbol{R}'(t)$, which follows the simple updating rule $\boldsymbol{R}'(t+1) = (\boldsymbol{I} - \eta\boldsymbol{M}(t))\boldsymbol{R}'(t)$, so Lemma 3.3 implies that the $\boldsymbol{R}$ part almost keeps decreasing during the entire trajectory (here Lemma 3.3 applies with $\boldsymbol{u}$ replaced by $\boldsymbol{R}'(t)$, noticing that the eigenvalues except the first are well below $2/\eta$). Hence, the non-monotonicness or the oscillation of the loss is mainly due to the $\boldsymbol{v}_1$-component of $\boldsymbol{D}$, and the rest part $\boldsymbol{R}$ is optimized in the classical regime (step size well below 2/(the operator norm)) and hence steadily decreases. See Figure 3(b).

4 A Theoretical Analysis for 2-Layer Linear NN

In this section, we aim to provide a more rigorous explanation of the EOS phenomenon in two-layer linear networks. The proof ideas follows almost the same intuition as the proof in Section 3.3. In particular, we can remove or replace the assumptions in Section 3.3 by arguably weaker assumptions. Due to space limit, we state our main theoretical results and elaborate their relation with the proofs in Section 3.3. The detailed settings and proof can be found in Appendix C. Reasoning about the assumptions can be found in Appendix D.

4.1 Setting and basic notations

Model: In this section, we study a two-layer neural network with linear activation, i.e. $f(x) = \sum_{q=1}^m \frac{1}{\sqrt{m}} a_q \boldsymbol{w}_q x = \frac{1}{\sqrt{m}}\boldsymbol{A}^\top\boldsymbol{W}x$ where $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_m]^\top \in \mathbb{R}^{m\times d}$, $\boldsymbol{A} = [a_1, a_2, ..., a_m] \in \mathbb{R}^n$.

Input distribution: For simplicity, we assume $y_i = \pm 1$ for all $i \in [n]$, and $\|\mathbf{X}^\top\mathbf{X}\|_2 = \Theta(n)$. We assume $\mathbf{X}^\top\mathbf{X}$ has rank $r \ll n$, and we decompose $\mathbf{X}^\top\mathbf{X}$ and $\boldsymbol{Y}$ according to the orthonormal basis $\{\boldsymbol{v}_i\}$, the eigenvectors of $\mathbf{X}^\top\mathbf{X}$: $\mathbf{X}^\top\mathbf{X} = \sum_{i=1}^r \lambda_i \boldsymbol{v}_i\boldsymbol{v}_i^\top$, $\boldsymbol{Y} = \sum_{i=1}^r (\boldsymbol{Y}^\top\boldsymbol{v}_i)\boldsymbol{v}_i := \sum_{i=1}^r z_i\boldsymbol{v}_i$ where $\boldsymbol{v}_i$ is the eigenvector corresponding to the i-th largest eigenvalue λ_i of $\mathbf{X}^\top\mathbf{X}$. $z_i = \boldsymbol{Y}^\top\boldsymbol{v}_i$ is the projection of $\boldsymbol{Y}$ onto the direction $\boldsymbol{v}_i$. Also, we assume global minimum $(\boldsymbol{A}^*, \boldsymbol{W}^*)$ exists.

Update rule: We write explicitly the GD dynamics of $\boldsymbol{D}(t)$: $\boldsymbol{D}(t+1) = (\boldsymbol{I}_n - \eta\boldsymbol{M}^*(t))\boldsymbol{D}(t)$, where $\boldsymbol{M}^*(t) = \frac{2}{m\eta}(\|\boldsymbol{A}(t)\|^2\mathbf{X}^\top\mathbf{X} + \mathbf{X}^\top\boldsymbol{W}^\top(t)\boldsymbol{W}(t)\mathbf{X}) - \frac{4\eta}{n^2 m}(\boldsymbol{D}(t)^\top\boldsymbol{F}(t))\mathbf{X}^\top\mathbf{X}$ is the Gram matrix combined with second order terms.

4.2 Main Theorem and The Proof Sketch

Now we state the assumptions and our main theorem.

Phase I and Progressive Sharpening:

Assumption 4.1. *There exists some constant $\chi > 1$, s.t. for all $i \in [r-1]$, $\lambda_i(\mathbf{X}^\top\mathbf{X}) \le \chi\lambda_{i+1}(\mathbf{X}^\top\mathbf{X})$. Moreover, $\lambda_1(\mathbf{X}^\top\mathbf{X}) \ge 2\lambda_2(\mathbf{X}^\top\mathbf{X})$.*

Assumption 4.2. *There exists $\kappa = \Omega(r^{-1})$ such that $\min_{i\in[r]}\{z_i/\sqrt{n}\} \ge \kappa$.*

The first assumption is about the eigenvalue spectrum of $\mathbf{X}^\top\mathbf{X}$. The second assumes that all component $z_i = \boldsymbol{Y}^\top\boldsymbol{v}_i$ are not too small. [2]

Theorem 1. *Suppose Assumption 4.1, Assumption 4.2 hold. Suppose the smallest nonzero eigenvalue $\lambda_r = \lambda_r(\mathbf{X}^\top\mathbf{X}) > 0, \lambda_1 = \lambda_{\max}(\mathbf{X}^\top\mathbf{X}) = c_1 n$. Then for any $\epsilon > 0$, if $m = \Omega(\frac{c_1 n^2}{\lambda_r^2})$, and $n = \Omega(\frac{\lambda_r^2}{\kappa^4\epsilon^2})$ we have the following properties for $t = 1, 2, ..., t_0 - 1$ where t_0 is the time when $\|D(t)\|^2 \le O(\epsilon^2)$ or $\lambda_{\max}(\boldsymbol{M}^*(t)) > 1/\eta$ for the first time:*

[2] It guarantees the gap between two adjacent eigenvalues is not very large, and there is a gap between the largest and the second largest eigenvalue. Note the second part of the assumption is a relaxed version of Assumption 3.6. In our CIFAR-10 1k-subset with samples' mean subtracted, $\lambda_1/\lambda_2 = \chi \approx 3$

- $\Lambda(t+1) - \Lambda(t) > 0$.
- $\|\mathbf{\Gamma}(t)\| = O(1/m)$, *where* $\mathbf{\Gamma}(t) = \frac{2}{mn}(\mathbf{X}^\top \boldsymbol{W}^\top(t)\boldsymbol{W}(t)\mathbf{X} - \mathbf{X}^\top \boldsymbol{W}^\top(0)\boldsymbol{W}(0)\mathbf{X})$.

In the proof of this theorem, we justify Assumption 3.2 we made in the previous section by proving the Gram matrix $\boldsymbol{M}(t) \approx \frac{2}{mn}(\|\boldsymbol{A}(t)\|^2 + \frac{m}{d})\mathbf{X}^\top\mathbf{X}$ with other terms bounded. That shows all $\boldsymbol{M}(t)$ approximately share the same set of eigenvectors as $\mathbf{X}^\top\mathbf{X}$. In our proof, we also prove more rigorously $\|\boldsymbol{A}(t)\|^2$ is an indicator of the sharpness in this simpler setting.

Edge of Stability (Phase II - IV):

Assumption 4.3. *There exists some constant $c_2 > 0$, such that $\|\mathbf{\Gamma}(t)\| \leq \frac{c_2}{m}$.*

This assumption is based on Theorem 1. In Theorem 1, we state that in the progressive sharpening phase, $\|\mathbf{\Gamma}(t)\|$ has an upper bound of $O(1/m)$. Now in the EOS phase, we assume that $\|\mathbf{\Gamma}(t)\|$ grows larger by at most constant factor. We note that Assumption 3.2 and 3.4 in Section 3.3 can both reduced to this assumption.

Assumption 4.4. *There exists some constant $\beta > 0$, such that $\Lambda \leq \frac{4}{\eta}(1-\beta)$.*

Similar with Assumption 3.1, we need to assume an upper bound of the sharpness. The reason why this assumption is reasonable is the same as Assumption 3.1.

Assumption 4.5. *There exist some constant c_3 such that $|\boldsymbol{D}(t)^\top v| > c_3\sqrt{n}/m$ for some $t = t_0$ at the beginning of phase II.*

This assumption is in the same spirit of Assumption 3.5 with the only change of the bound in terms of m and n. Now, we are ready to state our theorems in this stage.

Theorem 2. *Suppose the smallest nonzero eigenvalue $\lambda_r = \lambda_r(\mathbf{X}^\top\mathbf{X}) > 0, \lambda_1 = \lambda_{\max}(\mathbf{X}^\top\mathbf{X})$. Under Assumption 4.3,4.4, 4.5, and $\lambda_1(\mathbf{X}^\top\mathbf{X}) \geq 2\lambda_2(\mathbf{X}^\top\mathbf{X})$ in Assumption 4.1, there exists constants c_4, c_5, c_6 such that if $n > c_6\lambda_r\eta, m > \max\{\frac{c_4 d^2 n^3}{\lambda_r^2\eta^4\lambda_1}, c_5\eta\}$, then:*

- *There exists $\rho = O(1)$ which is related to c_3 such that if $\Lambda(t_0) > \frac{2}{\eta}(1+\rho)$ for some t_0, there must exist some $t_1 > t_0$ such that $\Lambda(t_1) < \frac{2}{\eta}(1+\rho)$.*
- *If $\Lambda(t), \Lambda(t+1) > \frac{2}{\eta}(1+\rho)$, then there is a constant $c_7 > 0$ (related to c_3) such that $|\boldsymbol{D}(t+1)^\top \boldsymbol{v}_1| > |\boldsymbol{D}(t)^\top \boldsymbol{v}_1|(1+c_7)$.*
- *Define $\boldsymbol{R}'(t) := (\boldsymbol{I} - \eta\boldsymbol{M}^*(t))\boldsymbol{R}'(t-1)$. $\|\boldsymbol{R}(t) - \boldsymbol{R}'(t)\| = O(\frac{n^2 d}{\eta^2\lambda_r\sqrt{m\lambda_1}})$.*

We can conclude the following from Theorem 2: (1) The first statement of theorem states that if the progressive sharpening phase causes the sharpness to grow over $2/\eta$, then the sharpness eventually goes below $2/\eta$. This illustrates the regularization effect of gradient descent on the sharpness, which is the goal of Phase III in Section 3.3. (2) The second states that $|\boldsymbol{D}(t)^\top v|$ geometrically increases in Phase II and III. Note that we proved a similar Lemma 3.2 for Phase II in the more general setting in Section 3.3. (3) The third conclusion shows the distance between $\boldsymbol{R}(t)$'s trajectory and $\boldsymbol{R}'(t)$ is small (similar to that in Phase IV of Section 3.3).

5 Discussions and Open Problems

In this section, we discuss the limitation of our theory and some related findings. First, our argument crucially relies on the assumption that $\|\boldsymbol{A}\|$ changes in the same direction as Λ does most of time. Here, we elaborate more on this point. Seeing from a longer time scale, $\|\boldsymbol{A}\|^2$ and the sharpness may have very different overall trends (See Figure (c) in 2). Moreover, in later iterations, the sharpness may non-monotonically increases or decreases more frequently while $\|\boldsymbol{A}\|^2$ is monotone (See Figure (b) in 2). While we believe the change of $\|\boldsymbol{A}\|$ is a major driving force of the change of the sharpness, other factors (such as other layers) must be taken into consideration for a complete understanding and explanation of the sharpness dynamics. We also carry out some experiments that reveals some interesting relation between the inner layers and the sharpness, which is not yet reflected in our theory. Due to space limit, we defer it to Appendix D.

We conclude with some open problems. It would be very interesting to remove some of our assumptions or replace them (especially those related to the spectrum of $\boldsymbol{M}$) by weaker or more natural assumptions on the data or architectures. Extending our results in Section 4 to deeper neural networks with nonlinear activation function is a challenging open problem.

References

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. *arXiv preprint arXiv:2204.01050*, 2022.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. *arXiv preprint arXiv:2002.09572*, 2020.

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, 2019a.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Pathological spectra of the fisher information metric and its variants in deep neural networks. *arXiv preprint arXiv:1910.05992*, 2019b.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

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.

Chao Ma, Lei Wu, and Lexing Ying. The multiscale structure of neural network loss functions: The effect on optimization and origin. *arXiv preprint arXiv:2204.11326*, 2022.

James Martens. *Second-order optimization for neural networks*. University of Toronto (Canada), 2016.

Vardan Papyan. The full spectrum of deepnet hessians at scale: Dynamics with sgd training and sample size. *arXiv preprint arXiv:1811.07062*, 2018.

Vardan Papyan. Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet hessians. *arXiv preprint arXiv:1901.08244*, 2019.

Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Lei Wu, Chao Ma, et al. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

Yikai Wu, Xingyu Zhu, Chenwei Wu, Annie Wang, and Rong Ge. Dissecting hessian: Understanding common structure of hessian in neural networks. *arXiv preprint arXiv:2010.04261*, 2020.

Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with sgd. *arXiv preprint arXiv:1802.08770*, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3.3 and 4, we included all assumptions.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix C.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Appendix A.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix A.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix A.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 3.3 and Appendix A.
   (b) Did you mention the license of the assets? [Yes] See Appendix A.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] The data is downloaded through torchvision.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]