Edge of Stability Training Dynamics

Itai Shapira
Harvard University

Abstract

Cohen et al. [2021] identified a distinct two-phase dynamics in the gradient descent trajectory within the loss landscape. Initially, the leading eigenvalue of the objective function's Hessian increases to $2/$step size (Progressive Sharpening), followed by a phase transition characterized by oscillations of the top eigenvalue around this critical value (Edge of Stability). This project reviews the empirical findings of Cohen et al. [2021] and provides an expository summary of two theoretical papers. The first, Damian et al. [2022], shows that the dynamics of gradient descent at the edge of stability can be captured by a cubic Taylor expansion. This forms a negative feedback loop, where, upon divergence of the iterates, the third-order term in the Taylor expansion of the loss function becomes predominant. Characterized by stages of increasing sharpness followed by self-correction, this cyclical process effectively serves as a self-correcting mechanism. The second paper, Li et al. [2022], focuses on the mechanisms of Progressive Sharpening. Analyzing a two-layer linear model, they argue that sharpness is primarily influenced by the norm of the weights in the output layer. The increase in sharpness hinges on the consistently negative inner product between the network's predictions and the residual, leading to an increase in the norm and, consequently, in the sharpness.

1 Introduction

Gradient descent and its simple variations provably reduce the objective function at every step, provided the largest eigenvalue of the Hessian, referred to here as the sharpness, is bounded by $2/\eta$ along the optimization trajectory, where η is the learning rate. Conversely, the path taken by gradient descent is provably diverging if the sharpness exceeds this threshold at any step. Thus, the stability of gradient descent is determined by whether the optimization path remains within $S_{\text{good}} := \{\theta \mid S(\theta) < \frac{2}{\eta}\}$, where $S(\theta)$ denotes the sharpness (see Definition 2.1), or lies entirely in its complement.

Cohen et al. [2021] observed that in the training of deep neural networks, the training loss exhibits a cart-before-the-horse dynamic: sharpness steadily increases along the gradient descent trajectory until it reaches the instability cutoff of $2/\eta$ ('Progressive Sharpening'). Once the trajectory crosses this critical threshold, a phase transition occurs, and the sharpness stably hovers around $2/\eta$, with the loss decreasing albeit non-monotonically ('Edge of Stability').

In this project, we review the empirical findings from Cohen et al. [2021] and provide an expository summary of two theoretical papers on this topic. The first, Damian et al. [2022], shows that the dynamics of gradient descent at the edge of stability follow a self-correcting mechanism where the divergence of the iterate in S_{bad} forces the sharpness down and the trajectory back to S_{good}. The second, Li et al. [2022], attempts to explain the progressive sharpening phenomenon by analyzing simple two-layer linear neural networks.

This presentation prioritizes clarity over exhaustive technical detail. To accurately yet succinctly convey the core arguments of the two theoretical papers, we present the results under slightly stronger assumptions. We will explicitly mention any assumptions added and succinctly describe how one might extend the argument more generally.

The rest of this paper is structured as follows: in Section 2, we provide background on the connection between sharpness and gradient descent. In Section 3, we cover the findings in Cohen et al. [2021] and discuss related research. In Sections 4 and 5, we cover the arguments of Damian et al. [2022] and Li et al. [2022] respectively.

2 Background

Throughout, we assume $\mathcal{L} : \mathbb{R}^p \longrightarrow \mathbb{R}$ is a three times continuously differentiable function we are trying to minimize. We consider the trajectory $(\theta_t)_{t=0}^T$ induced by running the gradient descent algorithm on $\mathcal{L}$ with initial point $\theta_0 \in \mathbb{R}^p$ and $\theta_{t+1} - \theta_t := -\eta \nabla \mathcal{L}(\theta_t)$. As a first-order optimization algorithm, the dynamics of gradient descent are fundamentally influenced by the curvature of the loss function $\mathcal{L}$. Specifically, these dynamics are closely tied to the evolution of sharpness:

Definition 2.1 (Sharpness of Loss Function)**.** *The sharpness at a point θ of $\mathcal{L}$ is defined as $S(\theta) := \lambda_{max}(\nabla^2 \mathcal{L}(\theta))$, where λ_{max} denotes the largest eigenvalue of the Hessian matrix $\nabla^2 \mathcal{L}(\theta)$. If this eigenvalue is unique, the corresponding eigenvector is denoted by $u(\theta)$.*

The following lemma shows that, provided the learning rate is sufficiently small relative to $S(\theta)$, the function $\mathcal{L}$ will decrease along the trajectory $(\theta_t)_{t=0}^T$:

Lemma 1 (Descent Lemma)**.** *Suppose that $S(\theta) \leq \ell$ for any θ in an open set that contains the trajectory $\{\theta_t\}_{t=0}^T$. Then, the following inequality holds:*

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t) \leq -\frac{\eta}{2}(2 - \eta\ell)\|\nabla \mathcal{L}(\theta_t)\|^2$$

Proof. Since $S(\theta) \leq \ell$, $\mathcal{L}$ is ℓ-smooth. Therefore, we have:

$$\begin{aligned}
\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t) &\leq \nabla \mathcal{L}(\theta_t)^\top (\theta_{t+1} - \theta_t) + \frac{\ell}{2}\|\theta_{t+1} - \theta_t\|^2 \\
&= -\eta \nabla \mathcal{L}(\theta_t)^\top \nabla \mathcal{L}(\theta_t) + \frac{\ell}{2}\eta^2 \|\nabla \mathcal{L}(\theta_t)\|^2 \\
&= -\eta \|\nabla \mathcal{L}(\theta_t)\|^2 + \frac{\ell}{2}\eta^2 \|\nabla \mathcal{L}(\theta_t)\|^2 \\
&= -\frac{\eta}{2}(2 - \eta\ell)\|\nabla \mathcal{L}(\theta_t)\|^2
\end{aligned}$$

□

In other words, by considering a second-order approximation, Lemma 1 upper bounds the change in $\mathcal{L}$ at each step by a quadratic function in η. The lemma guarantees a decrease in loss at any step if $\eta < 2/\ell$, and the right-hand side is minimized when $\eta = 1/\ell$. This bound is tight, as the following example shows:

Example 2.1. *Consider a quadratic function $\mathcal{L}(\theta) = \frac{1}{2}\theta^\top A \theta$ where A is a positive semidefinite matrix.*

We can assume, without loss of generality, that A is diagonal. Let its eigenvalues be $(\lambda^1, \ldots, \lambda^p)$ in a descending order. Analyzing the trajectory of gradient descent, we have:

$$\theta_{t+1} = \theta_t - \eta \nabla \mathcal{L}(\theta_t) = (I - \eta A)\theta_t = (I - \eta A)^t \theta_0.$$

this dynamics evolve independently along each hessian eigenvector:

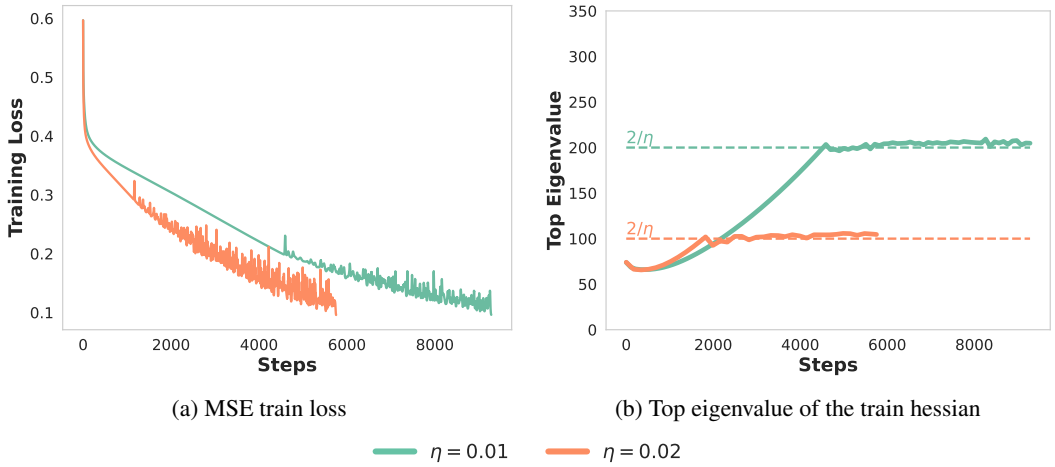
$$\forall k \in [p] \quad \theta_{t+1}^k = (1 - \eta\lambda^k)^t \theta_0^k.$$

If $\eta > \frac{2}{\lambda^1}$, the first coordinate diverges exponentially, demonstrating the tightness of the descent lemma's bound. For any i, if $\eta < \frac{2}{\lambda^i}$, the iterations θ_{t+1}^i converge exponentially to the optimum. Furthermore, for $\eta < \frac{1}{\lambda^i}$, this convergence is monotonic, while for $\frac{1}{\lambda^i} < \eta < \frac{2}{\lambda^i}$, the iterations oscillate as they converge. Directions of high curvature, corresponding to high eigenvalues, converges faster than directions of low curvature

This quadratic toy case helps us reason more generally about the behavior of gradient descent on neural networks. By considering the second-order Taylor approximation along the optimization trajectory,

if the sharpness at step t, $S(\theta_t)$, exceeds $2/\eta$, then the gradient descent iterates will diverge with exponentially increasing magnitude in the direction of each Hessian eigenvector whose eigenvalue is greater than $2/\eta$. As long as $S(\theta)$ is smaller than this threshold throughout the optimization trajectory, the loss provably drops in each iteration.

3 Edge of Stability



(a) MSE train loss

(b) Top eigenvalue of the train hessian

Figure 1: We reproduce the empirical observations from Cohen et al. [2021], where we trained an MLP network with two hidden layers of 200 units each and a smooth tanh activation function, on a randomly selected subset of 5,000 images from the CIFAR-10 dataset (Krizhevsky et al. [2009]), using the mean squared error (MSE) loss and full-batch gradient descent. During each step of the gradient descent optimization, we measure the training MSE loss (Figure 1a) and the top eigenvector of the training loss Hessian (Figure 1b, computed using the Lanczos algorithm, an adaption of the power method). Training is stopped once the train accuracy exceeds 99%.

In training deep neural networks using gradient descent and many of its simple variations, practitioners balance stability and convergence rate, while sharpness along the optimization path is a priori unknown. In a very comprehensive empirical study, Cohen et al. [2021] identify two stages in the dynamic of full-batch gradient descent on neural networks (see Figure 1):

- **Progressive Sharpening:** Initially, the training loss mostly decreases monotonically and $S(\theta_t)$ consistently increases along the training trajectory. This phase continues until the sharpness reaches the critical threshold of $2/\eta$. Typically, with standard initialization methods, the trajectory often starts from this phase. Moreover, the trajectory's behavior of increasing sharpness and decreasing loss closely parallels the dynamics seen in gradient flow.
- **Edge of Stability (EoS):** In this phase, the sharpness stably hovers around $2/\eta$ and the training loss exhibits a non-monotonic yet gradually decreasing behavior over long timescales.

That is, the training loss exhibits a cart-before-the-horse dynamic: so long as the sharpness is small enough for gradient descent to be stable, sharpness along the trajectory is consistently increasing. Once the trajectory crosses the critical threshold, a phase transition occurs, leading to a self-correction which deviates from the previously mentioned relationship and effectively prevents divergence. This behavior, marked by a continuous increase in sharpness followed by self-stabilization, has been observed across a variety of popular architectures and training datasets.

3.1 Related Work

Some aspects of Progressive Sharpening and EoS dynamics, such as the precise conditions under which these phenomena occur, their existence under various optimization algorithms (e.g. Adam), and

their connection to generalization performance, are still not fully understood. Theoretical explanations for these phenomena remain elusive. In the subsequent sections, we will provide an explanation for the second phase, EoS, under mild assumptions. To offer better context, we will first briefly review the current research in this area.

Progressive Sharpening Observations. Progressive sharpening has been observed empirically by several studies. Li et al. [2020] observed that by using normalization and weight decay, training to near-zero loss inevitably leads to high sharpness. Complementing this, Cohen et al. [2021] demonstrated that the EoS phenomenon can occur even in the absence of normalization. Gilmer et al. [2021] extends these observations to stochastic gradient descent, examining the impact of learning rates, architectural choices, and initialization on loss sharpness.

Empirical Observations of EoS Dynamics. Xing et al. [2018] reported that gradient descent eventually enters a regime characterized by oscillations in the leading curvature direction, accompanied by non-monotonic loss reduction. Similarly, Wu et al. [2018] found that the solutions attained by gradient descent exhibit sharpness values approximately equal to $2/\eta$. Jastrzebski et al. [2020] identified a "break-even" point in the SGD trajectory, beyond which there is a regularization effect on loss curvature. Additionally, Lewkowycz et al. [2020] argued that gradient descent could 'catapult' into flatter regions if the loss landscape around initialization is too sharp, proposing the "catapult phase" as a regime parallel to EoS.

Non-Monotonic Loss in Various Settings. The non-monotonic behavior of loss has been noted in diverse settings, e.g. Jastrzebski et al. [2020], Xing et al. [2018], Lewkowycz et al. [2020], Li et al. [2022] and Arora et al. [2018].

Conditions and Characterizations of EoS. Ahn et al. [2022a] discusses unstable convergence in gradient descent. Arora et al. [2022] connects modifications in gradient descent, such as normalized GD or GD with weight decay, to minimizing gradient flow sharpness. Lyu et al. [2022] examines GD entering EoS in normalized loss scenarios, emphasizing the sharpness reduction effect. Similarly, Ma et al. [2022] shows that subquadratic growth around minimizers leads to 2-periodic trajectories in GD, contributing to our understanding of EoS dynamics.

EoS in Specific Classes of Objectives. Research focusing on specific classes of objectives further elucidates EoS mechanisms. Chen and Bruna [2022] investigates periodic behavior in models like two-layer scalar networks, while Agarwala et al. [2022] examines squared quadratic functions. Ahn et al. [2022b] studies 2-dimensional objectives, providing bounds on final sharpness as η decreases. This exploration is complemented by Ma et al. [2022], who describe a quasi-static heuristic for GD trajectories with oscillating components. Moreover, Zhu et al. [2022] and Li et al. [2022] analyze simpler models. For instance, Zhu et al. [2022] examines a two-dimensional loss function $(x, y) \mapsto (x^2y^2 - 1)^2$.

4 Self-Stabilization Dynamics

In this section, we aim to provide an overview of the study Damian et al. [2022]. This work seeks to explain the dynamics of the gradient descent trajectory at the EoS using the cubic Taylor expansion of the loss function, $\mathcal{L}$. Let S_{bad} denote the set of unstable points:

$$S_{\text{bad}} := \{\theta \mid S(\theta) \geq 2/\eta\}$$

and we denote its complement as S_{good}. We assume that the gradient descent trajectory encounters instability at some point. For notational simplicity, we shift time so that for θ_0, it holds that $S(\theta_0) = 2/\eta$.

Following empirical observations, we consider the case where only the leading eigenvalue of the Hessian exceeds the critical value of $2/\eta$, hence instability occurs along the top eigenspace only. We further assume that gradient descent tends to increase sharpness further, i.e., gradient flow naturally increases sharpness, i.e., $\langle \nabla S(\theta_0), \nabla \mathcal{L}(\theta_0) \rangle < 0$. Our goal now is to demonstrate that the trajectory $\{\theta_t\}_{t=0}^{T}$ will stably hover between S_{bad} and S_{good}. We show that as the iterates increase, the cubic term of $\mathcal{L}$ acts to decrease the curvature, effectively restoring stability and allowing the trajectory to exit S_{bad}.

4.1 Setup

First, as mentioned, we assume that EoS occurs in a single unstable direction:

Assumption 1 (Eigengap)**.** *For some absolute constant $c < 2$, for all t, we have $\lambda_2(\nabla^2\mathcal{L}(\theta_t)) < c/\eta$.*

This implies that $S(\theta)$ is differentiable. Let u be the leading eigenvector of $\nabla^2\mathcal{L}(\theta_0)$. In any small neighborhood of θ_0 within S_{bad}, the displacement $\theta_t - \theta_0$ diverges along u, with increasing values of $|u^\top(\theta_t - \theta_0)|$. We track the movement of the iterates in this direction:

$$x_t := u^\top(\theta_t - \theta_0)$$

Analyzing the dynamics of gradient descent solely on $t \mapsto \mathcal{L}(\theta_0 + t \cdot u)$ does not suffice to capture the dynamics of (θ_t) in the u direction. Specifically, we show that the movement of the components of $\nabla S(\theta_t)$ in the subspace spanned by other eigenvectors affects the dynamics of x by altering the sharpness. Let $\Pi_u^\perp$ be the projection operator onto the orthogonal complement of the subspace spanned by u. We track displacement along the $\Pi_u^\perp \nabla S(\theta)$ direction:

$$y_t := (\Pi_u^\perp \nabla S(\theta_0))^\top(\theta_t - \theta_0)$$

Hence, the sharpness at time t is decomposed into the initial $2/\eta$, y_t, and the change along u:

$$S(\theta_t) \approx \frac{2}{\eta} + \nabla S(\theta_0)^\top(\theta_t - \theta_0) = \frac{2}{\eta} + \langle \nabla S(\theta_0), u\rangle u + y_t$$

We focus only on cases where sharpness increases as the algorithm progresses. Hence, we introduce the assumption of progressive sharpening, although it is not essential for the subsequent analysis:

Assumption 2 (Progressive Sharpening)**.** $\alpha(\theta_0) := -\nabla\mathcal{L}(\theta_0)^\top \nabla S(\theta_0) > 0$

Consider the Taylor approximation of $\nabla\mathcal{L}$ up to the quadratic term around θ_0:

$$P_{\nabla\mathcal{L}(\theta_t),2}(\theta_t) := \nabla\mathcal{L}(\theta_0) + \nabla^2\mathcal{L}(\theta_0)(\theta_t - \theta_0) + \frac{1}{2}\nabla^3\mathcal{L}(\theta_0)(:, \theta_t - \theta_0, \theta_t - \theta_0) \qquad ((1))$$

Here, $\nabla^3\mathcal{L}(\theta_0)(:, \theta_t - \theta_0, \theta_t - \theta_0)$ denotes the operation of contracting the third-order derivative tensor $\nabla^3\mathcal{L}(\theta_0)$ over its last two indices with the vector $\theta_t - \theta_0$.

Next, decompose the change in trajectory into $\theta_t - \theta_0 = x_t u + \Pi_u^\perp(\theta_t - \theta_0)$:

$$\begin{aligned}
P_{\nabla\mathcal{L}(\theta_t),2}(\theta_t) &= \nabla\mathcal{L}(\theta_0) + \left(x_t\nabla^2\mathcal{L}(\theta_0)u + \nabla^2\mathcal{L}(\theta_0)\Pi_u^\perp(\theta_t - \theta_0)\right) \\
&\quad + \frac{1}{2}\nabla^3\mathcal{L}(\theta_0)(:, \theta_t - \theta_0, \theta_t - \theta_0) \\
&= \nabla\mathcal{L}(\theta_0) + \frac{2}{\eta}x_t u + \frac{1}{2}x_t^2\nabla^3\mathcal{L}(\theta_0)(:, u, u) \\
&\quad + \frac{1}{2}\nabla^3\mathcal{L}(\theta_0)(:, \Pi_u^\perp(\theta_t - \theta_0), \Pi_u^\perp(\theta_t - \theta_0)) \\
&\quad + x_t\nabla^3\mathcal{L}(\theta_0)(:, \Pi_u^\perp(\theta_t - \theta_0), u)
\end{aligned}$$

The cubic term is dominated by $\nabla^3\mathcal{L}(\theta_0)(:, u, u)$, which we can express in terms of change in sharpness:

Lemma 2 (Self-Stabilization Property)**.** *For any $\theta \in \mathbb{R}^p$, where $u(\theta)$ denotes the top eigenvector of $\nabla^2\mathcal{L}(\theta)$:*

$$\nabla S(\theta) = \nabla^3\mathcal{L}(\theta)(:, u(\theta), u(\theta))$$

Proof. Since $\mathcal{L}$ is thrice continuously differentiable, by definition, the tensor contraction of $\nabla^3\mathcal{L}(\theta)$ along $(u(\theta), u(\theta))$ yields the gradient of $\frac{\partial^2\mathcal{L}}{\partial u^2}(\theta)$, which is equivalent to $u(\theta)^\top\nabla^2\mathcal{L}(\theta)u(\theta) = S(\theta)$. □

We add another simplifying assumption to eliminate the constant component of $P_{\nabla\mathcal{L}(\theta_t),2}$ along the direction of u:

Assumption 3. $\langle \nabla \mathcal{L}(\theta_0), u \rangle = 0$

i.e., $\nabla \mathcal{L}(\theta_0) = \Pi_u^{\perp}(\nabla \mathcal{L}(\theta_0))$. This assumption is relaxed in the paper by expanding the Taylor expansion of $\mathcal{L}$ around the projection of θ_0 onto $S_{\text{bad}} \cap \{\theta \mid \nabla \mathcal{L}(\theta_0)^\top u = 0\}$, and then bounding the error.

Let β be the norm of the gradient of the sharpness at θ_0: $\beta = \|\Pi_u^{\perp} \nabla S(\theta_0)\|^2$. To bound the error from approximating $\nabla \mathcal{L}$ with the second-order Taylor polynomial, we also define ρ_4 such that $\|\nabla^4 \mathcal{L}(\theta)\|_{op} \leq \rho_4$ and $r^3 = \max_t\{x_t^3, y_t^3\}$. We express the updates y_t in terms of the Taylor expansion $P_{\nabla \mathcal{L}(\theta_0),2}(\theta_t)$:

$$\begin{aligned} y_{t+1} - y_t &= \langle \Pi_u^{\perp}(\nabla S(\theta_0)), \theta_{t+1} - \theta_t \rangle = -\eta \langle \Pi_u^{\perp}(\nabla S(\theta_0)), \nabla \mathcal{L}(\theta_t) \rangle \\ &= -\eta \langle \nabla S(\theta_0), \Pi_u^{\perp}(P_{\nabla \mathcal{L}(\theta_0),2}(\theta_t)) \rangle + O(\eta \sqrt{\beta} \rho_4 r^3) \\ &= \eta\alpha - 2x_t \langle \Pi_u^{\perp}(\nabla S(\theta_0)), u \rangle - \frac{\eta}{2} x_t^2 \beta + O(\eta \sqrt{\beta} \rho_4 r^3) \\ &= \eta \left(\alpha - \frac{x_t^2 \beta}{2} \right) + O(\eta \sqrt{\beta} \rho_4 r^3) \end{aligned} \tag{(2)}$$

Similarly,

$$\begin{aligned} x_{t+1} &= x_t + \langle u, \theta_{t+1} - \theta_t \rangle = x_t - \eta \langle u, \nabla \mathcal{L}(\theta_t) \rangle \\ &= x_t - \eta \underbrace{\langle u, \nabla \mathcal{L}(\theta_0) \rangle}_{=0,\ \text{A.3}} - 2x_t \langle u, u \rangle - \frac{\eta}{2} x_t^2 \langle u, \nabla S(\theta_0) \rangle - \eta x_t \langle \Pi_u^{\perp}(\nabla S(\theta_0)), u \rangle + O(\eta \rho_4 r^3) \\ &= -x_t - \eta x_t y_t + O(\eta \rho_4 r^3) \\ &= -(1 + \eta y_t) x_t + O(\eta \rho_4 r^3) \end{aligned} \tag{(3)}$$

Finally, for our estimations of x_t and y_t, we informally assume that all higher-order terms are negligible, implying $x_t, y_t > \Omega(\eta \rho_4 r^3)$. Combining Equations ((2)) and ((3)), we define:

Definition 4.1. *We define the Predicted Edge of Stability dynamics as the discrete-time dynamical system given by:*

$$(x_{t+1}, y_{t+1}) := (-(1 + \eta y_t) x_t, (\alpha\eta - \frac{\eta}{2} \beta x_t^2) + y_t)$$

with initial conditions $y_0 = 0$ and $x_0 > 0$.

Under the assumptions above, the sequence $\{(x_t, y_t)\}_{t=0}^T$ describes the gradient descent dynamics at the edge of stability along the two directions of interest, up to an error term of the fourth order.

4.2 Analyzing the Dynamics

We now turn to qualitatively analyze the behavior of the sequence in Definition 4.1. The trajectory (x_t) behaves as the trajectory of gradient descent on a quadratic function with sharpness $\frac{2}{\eta} + y_t$ (recall Example 2.1). Hence, the trajectory (x_t) is stable if and only if $y_t < 0$. When $y_t > 0$, x_t diverges exponentially with increasing absolute values. The sequence (y_t) has a linear component $\eta\alpha$ and another component proportional to $-x_t^2$.

Qualitatively, we identify four distinct stages (see Figure 2):

1. **Progressive Sharpening**: When $|x_t|$ is small, y_t is dominated by the linear component: $y_t \approx \eta\alpha \cdot t$. By Assumption 2, the sharpness begins to increase at a rate proportional to η.
2. **Destabilization**: As $y_t > 0$, (x_t) begins to diverge, behaving like:

$$x_{t+1} = \prod_{i=0}^{t} (-1)^i (1 + \eta y_i) x_0$$

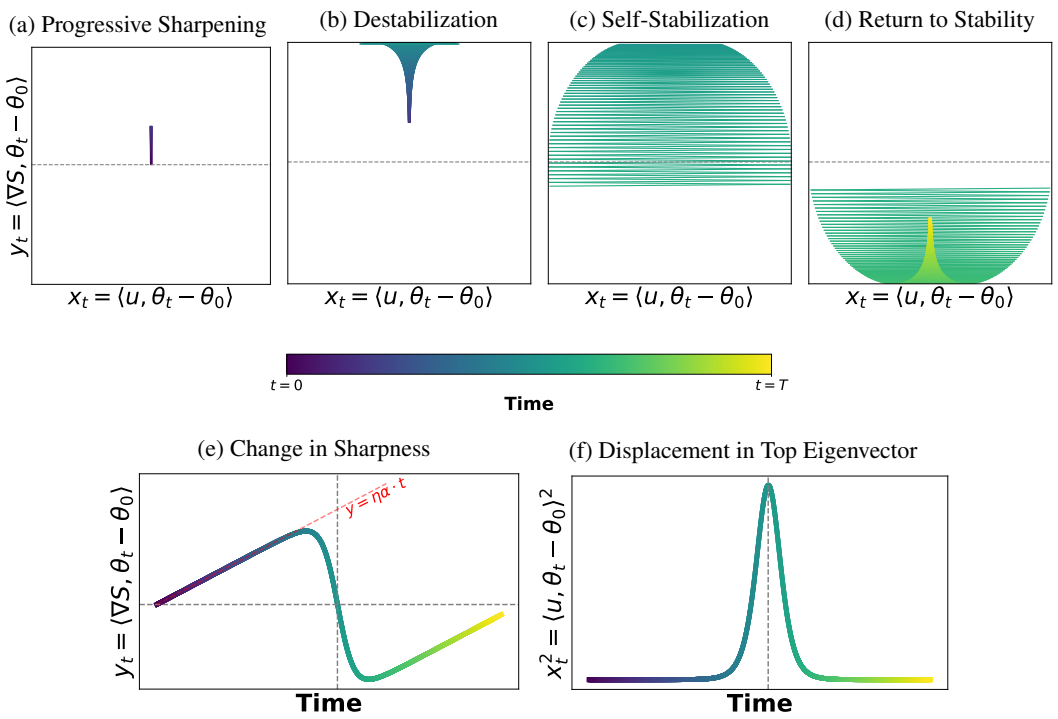


Figure 2: The dynamics at the edge of stability: we simulated the temporal progression of the predicted sequence $\{(x_t, y_t)\}_{t=1}^{T}$ with the initial conditions $y_0 = 0$ and $x_0 > 0$; to visualize four phases. In the initial phase (Figure 2a), the sharpness surpasses the critical threshold ($y_t > 0$). x_t is still small and hence the change in sharpness is proportional to $\eta\alpha$. Subsequently, the displacement along the principal eigenvector of the Hessian begins to diverge, with x_t exhibiting increasingly higher values of alternating signs (Figure 2b). Once x_t becomes sufficiently large, sharpness begins to decrease until we return to stability (Figure 2c). Finally, as y_t assumes negative values (Figure 2d), x_t begins to decrease in absolute value and the dynamics revert back to Progressive Sharpening.
In Figures 2e and 2f illustrate the self-correcting behavior of this dynamical system: when x_t^2 assumes larger values, y_t is pushed to negative values as a result of the cubic term in of $\nabla\mathcal{L}(\theta_t)$.

with alternating sign at each step and exponentially increasing values. The loss also increases rapidly:

$$\mathcal{L}(\theta_t) - \mathcal{L}(\theta_0) \approx \nabla\mathcal{L}(\theta_0)(\theta_t - \theta_0) + \frac{1}{2}(\theta_t - \theta_0)^\top \nabla^2\mathcal{L}(\theta_0)(\theta_t - \theta_0) \approx \frac{1}{\eta}x_t^2$$

3. **Self-Stabilization**: As x_t^2 grows larger, the quadratic component of y_t becomes more significant. y_t peaks when $x_t = \sqrt{\frac{2\alpha}{\beta}}$. Afterwards, the sharpness begins to decline. While the iterates diverge in the u direction, the strength of the movement in the $-\nabla S(\theta)$ direction forces the sharpness down. Here, α serves as a destabilizing force that locally increases sharpness, while β stabilizes.
4. **Return to Stability**: Eventually, y_t becomes negative, and the sharpness returns to a stable level. As x_t diminishes, the $\alpha\eta \cdot t$ component of y_t dominates once more, leading back to progressive sharpening.

In essence, the dynamics analyzed here form a negative feedback loop, aligning with the empirical observations made by Cohen et al. [2021]. As the algorithm enters regions of excessive sharpness, marked by the divergence of the iterates, the third-order term in the Taylor expansion of the loss function becomes predominant. This term acts to steer the trajectory back to S_{good}. This cyclical process, characterized by stages of increasing sharpness followed by self-correction, effectively

serves as a self-correcting mechanism, hence the term 'self-stabilization'. It ensures that the trajectory of gradient descent remains stable, oscillating between S_{good} and S_{bad}.

5 Linear models

In this section, we provide an overview of Li et al. [2022]. The objective is to demonstrate that sharpness increases along the gradient descent trajectory until it reaches a critical threshold, within the context of a two-layer linear neural network model. The central argument unfolds as follows: Initially, it is asserted that the sharpness is primarily influenced by the norm of the weights in the second (output) layer. Subsequently, we analyze the update equation for the norms of these second-layer weights during the training process. A key element of this update rule is the negative inner product between the network's predictions and the residual, which is identified as the main factor influencing sharpness changes. The escalation in sharpness is contingent on this inner product remaining negative. As long as this condition persists, both the norm of the second-layer weight and the sharpness continue to increase.

5.1 Setup

Consider a two-layer neural network $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ with linear activation functions, given by:

$$f(x) = \frac{1}{\sqrt{m}} A^\top W x,$$

where $W \in \mathbb{R}^{m\times d}$ are the weights of the hidden layer, and $A \in \mathbb{R}^m$ is the weight vector for the output layer. Let θ denote the set of all trainable parameters, $\theta = (W, A)$. The training data matrix is represented as $X \in \mathbb{R}^{d\times n}$, and the corresponding label vector is $Y \in \mathbb{R}^n$, with each $y_i = \pm 1$ for $i \in [n]$. Assume $X^\top X$ has rank r, and decompose $\Sigma := X^\top X$ into an orthonormal basis $\{v_i\}$ of the eigenvectors of $\Sigma = \sum_{i=1}^r \lambda_i v_i v_i^\top$, where v_i is associated with the i-th largest eigenvalue λ_i of $X^\top X$. It is assumed that the data is normalized such that $\|X^\top X\| = \Theta(n)$.

Define $F = ((f(x_i))_{i\in[n]}$ as the vector of outputs and D as the vector of residuals during training, $D = F - Y$. The subscript t is used to denote the state of variables at time t (e.g., D_t, F_t, etc.). The mean square error (MSE) is considered as the loss function during training, defined as $\mathcal{L}(A, W) = \frac{1}{n}\|D\|^2$.

Denote by G the Gauss-Newton matrix of $\mathcal{L} \circ f$ at θ:

$$n^{-1} \sum_{i=1}^{n} \frac{\partial f(x_i,\theta)}{\partial \theta}^\top \frac{\partial f(x_i,\theta)}{\partial \theta}$$

Empirical evidence (e.g., Papyan [2019]) shows that the sharpness, the top eigenvalue of $\nabla^2_\theta \mathcal{L}(f(\theta))$, is roughly the top eigenvalue of G: $||G|| \approx ||\nabla^2_\theta \mathcal{L}(f)||$. Let M denote the Gram matrix, defined as

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

Notice that G and M have the same non-zero eigenvalues. Thus, we assume for this analysis that considering M alone is sufficient:

Assumption 4. *For all t, $S(\theta_{t+1}) > S(\theta_t)$ if and only if $\lambda_1(M_{t+1}) > \lambda_1(M_t)$.*

For simplicity, $S(\theta_t)$ is is redefined as $\lambda_1(M_t)$. In our case,

$$M_t = \frac{2}{mn} \left(||A_t||^2 \Sigma + X^\top W_t^\top W_t X \right)$$

That is, M_t has two components: the first is proportional to Σ with a magnitude proportional to the norm of the second-layer weights, and the second is the covariance matrix of the second layer's pre-activation.

5.2 Simplifying the Problem

Next we show that under mild assumptions, it suffices to show that $v_1^\top M_t v_1$ is increasing with t to show progressive sharpening.

Assumption 5. *There exists some constant $c > 0$ such that for any t:*

$$||X^\top W_t W_t X - \frac{m}{d}\Sigma|| \leq \frac{cn}{2}$$

This assumption could be weakened, as shown in the paper. Moreover, this has been validated empirically (see Appendix D.2.2 in their paper). This means that $M_t \approx \frac{2}{mn}(||A_t||^2 + \frac{m}{d})\Sigma$.

Lemma 3. $S(\theta_t) \geq v_1^\top M_t v_1 \geq S(\theta_t) - \frac{2c}{m}$

Proof. Clearly $S(\theta_t) \geq v_1^\top M_t v_1$. Conversely, let u the leading eigenvector of M and Γ_t the matrix $\frac{2}{mn}(X^\top W_t W_t X - \frac{m}{d}\Sigma)$. Notice that now $M_t - \Gamma_t$ is a scalar multiplication of Σ. Then:

$$\begin{aligned} v_1^\top M_t v_1 &= v_1^\top (M_t - \Gamma) v_1 + v_1^\top \Gamma_t v_1 \overset{A.5}{\geq} u^\top (M_t - \Gamma_u) u + \frac{c}{m} \\ &\geq u^\top M_t u + \frac{2c}{m} = S(\theta_t) + \frac{2c}{m} \end{aligned}$$

□

Hence, we reduce the problem to show that $v_1^\top M_t v_1$ is increasing.

5.3 Analyzing the Dynamics

In this subsection we analyze the progression of $(D_t)_{t\in[T]}$, $(||A_t||^2)_{t\in[T]}$ and $(M_t)_{t\in[T]}$. The update rule under gradient descent is:

$$\begin{aligned} A_{t+1} - A_t &= -\eta \frac{\partial \mathcal{L}(F(\theta), Y)}{\partial A_t} = -\frac{2\eta}{n\sqrt{m}} W_t X D_t \\ W_{t+1} - W_t &= -\eta \frac{\partial \mathcal{L}(F(\theta), Y)}{\partial W_t} = -\frac{2\eta}{n\sqrt{m}} A_t D_t^\top X^\top \end{aligned}$$

Lemma 4 (The Dynamics of the Error)**.** *The update rule for the residual vector D_t is given by:*

$$D_{t+1} - D_t = -\eta D_t^\top M_t + \frac{4\eta^2}{n^2 m} \langle F_t, D_t \rangle D_t^\top \Sigma$$

Proof. We expand $D_{t+1}^\top - D_t^\top$ as follows:

$$\begin{aligned} D_{t+1}^\top - D_t^\top &= \frac{1}{\sqrt{m}}(A_{t+1}^\top W_{t+1} - A_t^\top W_t) X \\ &= \frac{1}{\sqrt{m}}((A_{t+1}^\top - A_t^\top) W_{t+1} + A_t^\top (W_{t+1} - W_t)) X \\ &= \frac{1}{\sqrt{m}}((A_{t+1}^\top - A_t^\top) W_t + A_t^\top (W_{t+1} - W_t)) X + \frac{1}{\sqrt{m}}(A_{t+1}^\top - A_t^\top)(W_{t+1} - W_t) X \\ &= -\frac{2\eta}{nm}\left(D_t^\top X^\top W_t^\top W_t + ||A_t||^2 D_t^\top X^\top \right) X + \frac{4\eta^2}{n^2 m\sqrt{m}} (W_t X D_t)^\top A_t D_t^\top X^\top X \\ &= -\eta D_t^\top M_t + \frac{4\eta^2}{n^2 m} \langle F_t, D_t \rangle D_t^\top \Sigma \end{aligned}$$

□

By assuming overparameterization $m > n$ (along with the assumption that $||\Sigma|| = \Theta(n)$), we ignore the second term and assume: $D_t^\top = D_0^\top \prod_{t' \in [t]} (I - \eta M_{t'})$. Moreover, we assume symmetric

initialization such that the output vector is initialized $F_0 = 0$ and thus $D_0 = -Y$, leading to an analogue of Example 2.1:

$$D_t = -Y \prod_{t' \in [t]} (I - \eta M_{t'}) \tag{(4)}$$

Consider the norm of the second layer, $||A_t||^2$:

$$\begin{aligned} ||A_{t+1}||^2 - ||A_t||^2 &= ||A_t - \eta \frac{\partial \mathcal{L}}{\partial A}(A_t)||^2 - ||A_t||^2 \\ &= -2\eta A_t^\top \frac{\partial \mathcal{L}}{\partial A} + \eta^2 ||\frac{\partial \mathcal{L}}{\partial A}||^2 \\ &= -\frac{4\eta}{n}\langle F_t, D_t \rangle + \eta^2 ||\frac{\partial \mathcal{L}}{\partial A}||^2 \end{aligned}$$

Lemma 5 (The Dynamics of the Gram Matrix)**.** *The update rule of M_t is,*

$$\begin{aligned} M_{t+1} - M_t = &-\frac{4\eta}{n^2 m}\left(2\langle F_t, D_t \rangle \Sigma + F_t D_t^\top \Sigma - \Sigma D_t F_t^\top\right) \\ &+ \frac{8\eta^2}{n^3 m^2}\left(||W_t X D_t||^2 \Sigma + ||A_t||^2 \Sigma D_t D_t^\top \Sigma\right) \end{aligned}$$

The change in M_t is decomposed into two terms: the first which is the result of the change in $||A_t||^2$, and a second which is dominated by the first. This shows that the dynamics of the sharpness is closely realted to the dynamics of $||A_t||^2$, i.e. dependent on $-\langle D_t, F_t \rangle$.

Proof. By the update rule of A_t and W_t we have:

$$\begin{aligned} \frac{mn}{2}(M_{t+1} - M_t) &= (||A_{t+1}||^2 - ||A_t||^2) X^\top X + X^\top (W_{t+1} W_{t+1}^\top - W_t^\top W_t) X \\ &= \left(-\frac{4\eta}{n}\langle D_t, F_t \rangle + \eta^2 ||\frac{\partial \mathcal{L}}{\partial A}||^2\right) X^\top X \\ &+ X^\top ((W_{t+1}^\top - W_t^\top) W_t + W_t^\top (W_{t+1} - W_t)) X \\ &+ X^\top (W_{t+1}^\top - W_t^\top)(W_{t+1} - W_t) X \\ &= -\frac{4\eta}{n}\langle D_t, F_t \rangle \Sigma + \frac{4\eta^2}{n^2 m}||W_t X D_t||^2 \Sigma \\ &- \frac{2\eta}{n}(F_t D_t^\top \Sigma + \Sigma D_t F_t^\top) + \frac{4\eta}{n^n m}||A_t||^2 \Sigma D_t D_t^\top \Sigma \end{aligned}$$

and the result follows. □

Let v_i be an eigenvector of $X^\top X$ with eigenvalue $\lambda_i > 0$. Then, using Lemma 5:

$$\begin{aligned} v_i^\top (M_{t+1} - M_t) v_i &= -\frac{8\lambda_i \eta}{n^2 m}\left(\langle F_t, D_t \rangle + \langle v_i, F_t \rangle \langle D_t, v_i \rangle\right) \\ &+ \frac{8\lambda_i \eta^2}{n^3 m^2}\left(\underbrace{||W_t X D_t||^2}_{\geq 0} + \underbrace{\lambda_i ||A_t||^2 \langle v_i, D_t \rangle^2}_{\geq 0}\right) \\ &\geq -\frac{8\lambda_i \eta}{n^2 m}\left(\langle F_t, D_t \rangle + \langle v_i, F_t \rangle \langle D_t, v_i \rangle\right) \end{aligned}$$

The primary component in the update equation of the second-layer weight norm is the negative inner product between the network's predictions and the residual, $\langle F_t, D_t \rangle$. As long as the inner product between predictions and residual is consistently negative, it leads to an increase in the second-layer weight norm, thus resulting in progressive sharpening. By Equation (4) and $F_t = Y + D_t$, we get that as long as $\lambda_1(M_t) < 2/\eta$, $\langle D_t, F_t \rangle < 0$.

References

A. Agarwala, F. Pedregosa, and J. Pennington. Second-order regression models exhibit progressive sharpening to the edge of stability. 2022. URL `https://arxiv.org/abs/2210.04860`.

K. Ahn, J. Zhang, and S. Sra. Understanding the unstable convergence of gradient descent. 2022a. ArXiv, abs/2204.01050.

Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via the" edge of stability". *arXiv preprint arXiv:2212.07469*, 2022b.

S. Arora, Z. Li, and A. Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 948–1024. PMLR, 2022.

Sanjeev Arora, Zhiyuan Li, and Kaifeng Lyu. Theoretical analysis of auto rate-tuning by batch normalization. *arXiv preprint arXiv:1812.03981*, 2018.

L. Chen and J. Bruna. On gradient descent convergence beyond the edge of stability. 2022. ArXiv, abs/2206.04172.

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. In *International Conference on Learning Representations*, 2021.

Alex Damian, Eshaan Nichani, and Jason D Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. *arXiv preprint arXiv:2209.15594*, 2022.

Justin Gilmer, Behrooz Ghorbani, Ankush Garg, Sneha Kudugunta, Behnam Neyshabur, David Cardoze, George Dahl, Zachary Nado, and Orhan Firat. A loss curvature perspective on training instability in deep learning. *arXiv preprint arXiv:2110.04369*, 2021.

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.

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

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.

Zhiyuan Li, Kaifeng Lyu, and Sanjeev Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*, 33:14544–14555, 2020.

Zhouzi Li, Zixuan Wang, and Jian Li. Analyzing sharpness along gd trajectory: Progressive sharpening and edge of stability. *arXiv preprint arXiv:2207.12678*, 2022.

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.

Chao Ma, Daniel Kunin, Lei Wu, and Lexing Ying. Beyond the quadratic approximation: the multiscale structure of neural network loss landscapes. *arXiv preprint arXiv:2204.11326*, 2022.

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

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.

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

Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. *arXiv preprint arXiv:2210.03294*, 2022.