# CONVERGENCE OF LARGE-STEPSIZE SGD FOR THE CROSS-ENTROPY LOSS VIA STOCHASTIC LYAPUNOV STABILITY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Modern deep learning has been shown to operate using learning rates far larger than those justified by classical optimization theory. Most prior analyses of this phenomenon focus on deterministic gradient descent, leaving the stochastic setting largely unexplored. In this work, we provide sharp convergence guarantees for Stochastic Gradient Descent (SGD) applied to the multiclass cross-entropy loss, for both linear classifiers and two-layer neural networks. We show that the stochasticity of SGD may cause the dynamics to alternate between an *edge-of-stability regime* that is dominated by curvature-driven oscillations, and a *stable regime* in which the expected loss decreases at a controlled rate. Despite that, we prove that SGD stabilizes the dynamics, ensuring that the iterates return to stability in a fixed number of iterations and allowing convergence in the best-iterate sense even with large learning rates. Experiments validate our theoretical findings and illustrate the benefits of SGD in the large-stepsize regime.

## 1 INTRODUCTION

Modern machine learning models are trained at unprecedented scales and with stepsizes that lie far outside the regime covered by classical optimization theory. Yet, stochastic gradient descent (SGD) and its variants (Bottou, 2012; Kingma and Ba, 2017; Gould and Tanaka, 2024) remain the basic workhorses of deep learning. A striking empirical phenomenon, reported repeatedly across architectures, datasets, and training paradigms, is that state-of-the-art performance is typically achieved by using large learning rates, often far exceeding the stability thresholds predicted by smooth convex optimization. This phenomenon, commonly referred to as the Edge of Stability (EoS) (Cohen et al.), has attracted considerable attention in recent years and has emerged as a unifying perspective on the behavior of gradient methods in modern deep networks (Damian et al.; Cohen et al., 2024).

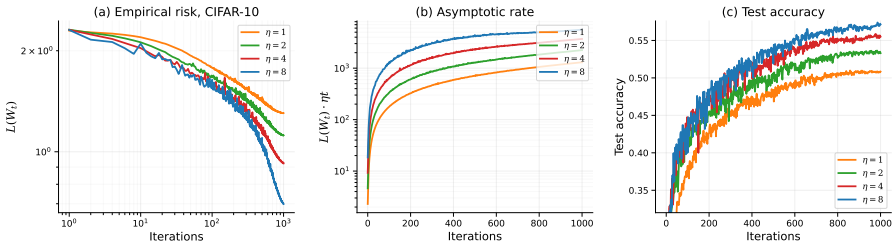


Figure 1: Experiment on CIFAR-10. An 8-layer NN with GELU activation function trained with SGD and large step sizes. The training loss decreases faster for large step sizes.

In the deterministic setting, the EoS phenomenon is now relatively well understood in linear logistic regression (Wu et al., 2024a). For Gradient Descent (GD) applied to the logistic regression setup

$$L(W) = \frac{1}{n}\sum_{i=1}^{n} \ell(Wx_i, y_i), \tag{1}$$



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the classical theory dictates that convergence with a monotonic decrease of the loss requires a step size $\eta$ satisfying $\lambda_{\max}(\nabla^2 L(W_t)) \leq \frac{2}{\eta}$, where $\lambda_{\max}(\nabla^2 L(W_t))$ denotes the maximum eigenvalue of the Hessian of the loss at the current iterate $W_t$. However, prior empirical results have demonstrated that GD often converges even when this condition is violated, entering a regime characterized by non-monotonic decrease of the loss and oscillatory dynamics. Moreover, in the case of the square objective $\lambda_{\max}(\nabla^2 L(W_t))$ hovers around the critical value $\frac{2}{\eta}$ (Cohen et al.; Damian et al.).

The aforementioned behavior was analyzed in a sequence of works (Wu et al., 2024b; Cai et al., 2024; Wu et al., 2025) culminating in the analysis of *large-stepsize GD* for logistic regression. A recurring conclusion is that the training dynamics can be divided in two regimes:

1. An early *EoS phase*, where the iterates experience curvature-driven oscillations, the loss is not monotonically decreasing, and the Hessian's top eigenvalue exceeds the stability boundary.
2. A late *stable phase*, in which the algorithm self-stabilizes, the iterates enter a region of the landscape where $\lambda_{\max}(\nabla^2 L(W_t)) \leq \frac{2}{\eta}$, and the loss decreases at a predictable rate.

A key insight is that, although the dynamics during the EoS phase appear erratic, their structure is surprisingly rigid: the oscillations persist only for a finite number of iterations, after which the system transitions to the stable regime where standard descent analysis becomes applicable.

These advances have significantly sharpened our understanding of large-stepsize GD. Nevertheless, the deterministic viewpoint does not capture the behavior of the stochastic counterpart of the algorithm (SGD), that is commonly used in practice. Indeed, stochasticity introduces several new layers of complexity. First, the notion of a stability threshold becomes ill-defined. Prior work (Wu et al., 2024a) only provides theoretical guarantees for the average loss of SGD in logistic regression, without establishing what happens when the iterates become stable. Second, even if the trajectory reaches a region that would be stable under full-batch dynamics, a stochastic update may drive the iterates back outside this region. Thus, rather than a single transition from an EoS phase to a stable phase, SGD may exhibit a more intricate evolution in which stable and unstable regimes alternate before the dynamics eventually settle. As a consequence, the deterministic EoS theory, based on precise control of the Hessian's largest eigenvalue does not extend in a straightforward or even meaningful way to the stochastic setting. This leaves a fundamental question open:

*What is the appropriate notion of stability for SGD with the cross-entropy loss and can we characterize the dynamics of SGD beyond the classical small learning rate regime?*

In this work, we aim to characterize the complex behaviour of the dynamics of SGD on the multi-class cross-entropy loss under arbitrarily large stepsizes and show that they exhibit three distinct regimes: an initial EoS regime, an intermediate regime in which stochastic fluctuations may repeatedly drive the trajectory in and out of stability, and a final with-high-probability stable regime. Our contributions can be described as follows:

- ⋆ We propose a notion of stability for SGD in the cross-entropy loss setting, by leveraging the stochastic Lyapunov stability theory for dynamical systems. An important characteristic is that the stochastic stability criterion reduces to the deterministic one in the full-batch limit and captures the precise balance between curvature and noise that determines whether SGD remains stable.
- ⋆ For the EoS setting, we provide convergence guarantees for SGD on the multi-class cross-entropy loss with arbitrarily large stepsizes $\eta > 0$, establishing the first theoretical guarantees for the multi-class setting.
- ⋆ For the intermediate regime, we prove that, despite the inherent stochasticity, SGD progressively stabilizes the trajectory, returning the dynamics back to stability after a fixed number of iterations and forcing the iterates to be stable with higher probability as the loss eventually decreases.
- ⋆ For the stable regime, we provide a refined analysis of the dynamics, showing that SGD enters the stable regime in a fixed number of iterations with high probability and establishing a rate of $\tilde{\mathcal{O}}\left(\frac{1}{t}\right)$ for the last iterate of convergence.
- ⋆ We provide a fine-grained analysis for the two-layer neural network dynamics of SGD in the large learning rate setting, establishing distinct EoS and stable regimes and characterizing the behaviour of the dynamics in each regime.
- ⋆ We provide extensive experimental validation of our theoretical results, showing the benefits of large step-sizes in practice.

## 2 SGD FOR CROSS-ENTROPY LOSS

We consider the multi-class classification problem (2) with $K$ classes, where $x_i \in \mathbb{R}^d$ is the input, $y_i \in \{1, \dots, K\}$ is the corresponding label, $f$ denotes the prediction function parameterized by weights $W = [w_1, \dots, w_K] \in \mathbb{R}^{d \times K}$, and $\ell(\cdot)$ is the cross-entropy loss function

$$L(W) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(f(x_i; W), y_i\big), \qquad \ell(f, y) = \log\Big( \sum_{j=1}^{K} e^{f_j} \Big) - f_y. \tag{2}$$

Our goal is to analyze the behavior of SGD in the large stepsize regime. For a mini-batch $\mathcal{B}_t \subseteq [n]$ of size $b \geq 1$, the SGD update at time $t \geq 0$ is given by

$$W_{t+1} = W_t - \eta\, g_t, \tag{SGD}$$

where $g_t = \frac{1}{b} \sum_{i \in \mathcal{B}_t} \nabla \ell(f(x_i; W_t), y_i)$ is the stochastic gradient oracle and $\eta > 0$ is the stepsize.

### 2.1 BLANKET ASSUMPTIONS

We, next, introduce the assumptions needed for presenting our theoretical results. We assume without loss of generality that inputs satisfy $\|x_i\|_2 \leq 1, \forall i \in [n]$. The next assumption regards the separability of the dataset, which is common in the literature (Wu et al., 2023; Cai et al., 2024) for this setting.

**Assumption 2.1** (Separable data)**.** There exists $W_* \in \mathbb{R}^{d \times K}$ with $\|W_*\|_F = 1$ and a margin $\gamma > 0$ such that for every $i \in [n]$ and every class $j \neq y_i$, it holds that $(W_* x_i)_{y_i} - (W_* x_i)_j \geq \gamma$.

### 2.2 CONVERGENCE GUARANTEES

We are now in a position to present our main convergence guarantees for the cross-entropy loss setting. We emphasize that the dynamics of SGD with large stepsizes are inherently non-monotonic: the loss may increase due to stochasticity, while curvature-driven oscillations may pose another obstacle in achieving convergence. Our first main result shows that, despite the inherent impediments, SGD is able to converge even with *arbitrarily large* stepsizes and in the presence of stochastic noise.

**Theorem 2.2.** Let Assumption 2.1 hold. The iterates of (SGD) with any step size $\eta > 0$ and batch size $b \geq 1$ satisfy for any $t > 0$ that

$$\min_{0 \leq k \leq t-1} \mathbb{E}\Big[L(W_k)\Big] \leq \frac{K - 1 + \ln^2(\gamma^2 \eta t) + \eta^2 (1 + \frac{1}{b})^2}{\gamma^2 \eta t}. \tag{3}$$

Theorem 2.2 establishes an upper bound on the best iterate of SGD with any step size $\eta > 0$. The theorem allows us to compute the number of iterations after which the loss is below a specific value. Interestingly, the result indicates a $\tilde{\mathcal{O}}\left(\frac{1}{t}\right)$ rate for the best iterate both for the case of small step sizes as well as even for the case of large step sizes. In the small learning rate regime, inequality (3) recovers the well known $\mathcal{O}\left(\frac{1}{t}\right)$ rate (Nacson et al., 2019b) for SGD last iterate of the algorithm.

Our result serves as an extension from the binary logistic setting considered in Wu et al. (2024a) to the multi-class cross-entropy loss, recovering the same rate of convergence in the case of $K = 2$. We highlight that even though the rate in Wu et al. (2024a) is established for the average loss, it can easily be algebraically converted to a best-iterate guarantee. Lastly, we note that our in expectation results can be converted to high probability results by leveraging Freedman's inequality.

### 2.3 STABILITY IN STOCHASTIC OPTIMIZATION

In the stochastic setting, non-monotonicity of the loss is no longer a sufficient signature of edge-of-stability behavior. Along the trajectory of SGD, the loss can increase for two different reasons. The first is simply the noise coming from the minibatch gradient: since SGD does not use the full gradient, the loss may occasionally increase even when the stepsize is small. The second reason is the use of a large stepsize: when the stepsize is too large compared to the local curvature of the loss, the iterates start to oscillate, which is the behavior typically associated with the edge of stability.

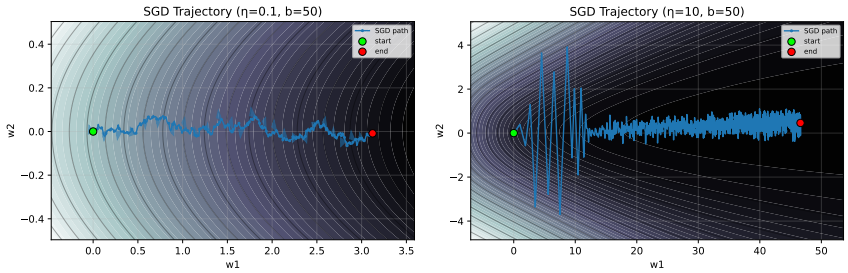


Figure 2: Two-dimensional trajectories of SGD with small and large stepsizes. Left: the trajectory of small stepsize SGD is affected by stochastic noise but does not oscillate. Right: in large stepsize SGD the stochastic noise is accompanied by genuine oscillations caused by the local curvature.

This makes the definition of stability less direct than in the deterministic setting. One could try to use the largest eigenvalue of the full-batch Hessian, or a similar quantity for the minibatch Hessian. However, any such definition has to allow for the fact that SGD may temporarily leave any given stable region because of minibatch noise. Thus, the stability criterion should capture whether the loss decreases on average over the mini-batches, rather than whether it decreases at every single step.

For this reason, we use a stochastic Lyapunov viewpoint (Kushner, 1967; Khasminskii, 2012; Mao, 2007). In deterministic optimization, a Lyapunov function $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, which in our case is the loss, decreases along the trajectory. This is exactly the role played by the usual descent lemma. For SGD, such a pointwise decrease is too strong, since every update contains noise. The natural replacement is that the loss decreases in conditional expectation. Namely, for a Lyapunov function $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, we say that the dynamics are Lyapunov stable whenever

$$\mathbb{E}\left[\mathcal{E}(W_{t+1}) - \mathcal{E}(W_t)\middle|\mathcal{F}_t\right] \leq 0. \tag{4}$$

We refer the interested reader for a dedicated comparison of the proposed stability criterion with other notions of stochastic stability to Appendix A.

### 2.4 The Stable Regime

In this section, we provide a refined analysis of convergence by studying the stable regime of SGD. Using the Lyapunov stability viewpoint, we define the stable set as

$$\mathcal{S} = \left\{W \in \mathbb{R}^{d\times K} : L(W) \leq \tilde{L}, \tilde{L} = \frac{1}{\eta c}\right\},$$

where $c = 8\left(1 + \frac{n}{b\min\{\gamma^2,1\}}\right)$. This definition captures the region in which the *expected loss* admits a one-step descent. In other words, inside $\mathcal{S}$ the stochastic updates may still fluctuate, but the loss decreases on average over the different mini-batches. The constant $c$ reflects the effect of minibatch noise through the batch size $b \geq 1$, and in the full-batch case the criterion recovers the deterministic stability condition up to constants.

With this definition in place, we can describe the behavior of SGD more precisely. The next theorem shows that the iterates enter the stable set with high probability and, as long as they remain inside it, the expected loss of the last iterate decays at a rate $\tilde{\mathcal{O}}(1/t)$.

**Theorem 2.3.** Let Assumption 2.1 hold. The iterates of (SGD) with any stepsize $\eta > 0$ and batch size $b \geq 1$ satisfy the following.

1. **Entrance to the stable regime.** For any $\delta \in (0,1)$, there exists $t_{in} \leq t_{max}(\delta)$ such that, with probability at least $1-\delta$, the SGD iterates enter the stable set $\mathcal{S}$ at time $t_{in}$.

2. **Convergence inside the stable regime.** If the iterates $W_k \in \mathcal{S}, \forall k \in [t_1, t-1]$ for some $t_1 > 0$, then it holds that

$$\mathbb{E}\Big[L(W_t)\Big] \quad \leq \quad \frac{8F(W_{t_1}) + 4\ln^2(\gamma^2\eta(t-t_1))}{7\gamma^2\eta(t-t_1)}, \tag{5}$$

where $F(W) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq y_i} e^{-\langle W(y_i)-W(j), x_i\rangle}$.

Theorem 2.3 gives a last-iterate guarantee for SGD after the dynamics have entered the stable regime. In the deterministic setting, GD first passes through an EoS phase and then enters a stable phase after some deterministic time bound $t_{in}^{GD}$. For SGD, however, the entrance time is necessarily probabilistic, because the trajectory depends on the sampled minibatches. Thus, the theorem shows that even in the presence of stochastic noise, the dynamics reach a region where the expected loss decreases at the standard $\tilde{\mathcal{O}}(1/t)$ rate.

However, there is an important difference from GD. In the deterministic case, once the iterates enter the stable regime, the dynamics remain there by construction. For SGD, this is no longer automatic: minibatch noise can temporarily push the iterates outside $\mathcal{S}$. This is the main additional difficulty in the stochastic setting and motivates the stochastic stabilization analysis developed in the next section.

### 2.5 Stabilization of SGD

We now show that SGD has an additional stabilizing property that is absent from the deterministic setup. Although mini-batch noise may occasionally push the iterates outside the stable set, this instability cannot persist indefinitely. Instead, the dynamics are pulled back toward the stable regime after a controlled number of iterations. In this sense, SGD is *progressively stabilizing* the trajectory:

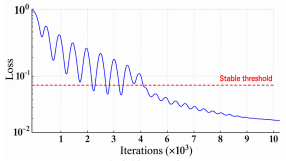

Figure 3: Stabilization property of SGD.

*even if the iterates escape the stable regime,*
*they return to stability in a fixed number of iterations.*

This property is the key reason why SGD can converge despite the noise introduced by minibatch gradients. The iterates may not remain in the stable set forever, but they revisit it often enough for the loss to keep decreasing over time. We make this statement precise in the following theorem.

**Theorem 2.4** (Stabilization of SGD)**.** Let Assumption 2.1 hold. The iterates of (SGD) with any stepsize $\eta > 0$ and batch size $b \geq 1$ satisfy the following.

1. **Stability is maintained with high probability.** If $W_t \in \mathcal{S}$, then with probability at least $1 - \delta'$ it holds that $W_{t+1} \in \mathcal{S}$, where $\delta' = \mathcal{O}\left(\exp\left(-\frac{b}{\eta L(W_t)^{3/2}}\right)\right)$.

2. **Return to stability.** Assume that the dynamics exit the stable regime at $t_{out} > 0$, and fix $\delta \in (0,1)$. Then, with probability at least $1-\delta$, the iterates of (SGD) return to the stable regime in at most $t_{\text{re}} = \tilde{\mathcal{O}}\left(\frac{\eta b^2}{\gamma^2 \delta \bar{L}}\right)$ number of steps.

Theorem 2.4 shows that the stable regime is robust under the stochastic perturbations of SGD. The first part states that, if the current iterate is already stable, then the next iterate remains stable with high probability. This probability improves as the loss becomes smaller and as the iterate moves farther away from the stability boundary. In particular, as $L(W_t) \to 0$, the probability of staying inside $\mathcal{S}$ tends to one.

The second part shows that even if SGD leaves the stable set, it returns after a bounded number of iterations with high probability. Thus, the stochastic noise may create short excursions outside $\mathcal{S}$, but it does not permanently destabilize the dynamics. The resulting picture is a repeated cycle: SGD enters the stable set, may occasionally exit because of minibatch noise, and then returns again. As training progresses and the loss decreases, the exits become less likely, and the dynamics spend more and more time in the stable regime.

This stochastic stabilization mechanism is one of the main differences between SGD and GD with large learning rates. In GD, once the iterates enter the stable phase, there is no stochastic noise that can push them out. In SGD, temporary exits are possible, but the geometry of the cross-entropy loss pulls the dynamics back inside the stable set. We emphasize that this mechanism is different from the self-stabilization phenomenon for the square loss studied by Damian et al., where stabilization comes from a higher-order Taylor term in the deterministic GD dynamics. Here, stabilization is instead driven by the self-bounded structure of the cross-entropy loss, which makes both curvature and stochastic noise decrease as the loss becomes small.

## 3 Edge of Stability for Two-Layer Neural Networks

An important question is whether the previous theoretical guarantees and insights extend to the dynamics of SGD when applied in neural networks. While the curvature of deep neural networks is more intricate, many architectures exhibit structural properties, such as near-homogeneity or controlled activation derivatives, that allow for a thorough analysis to be proceeded.

We consider two-layer neural networks in mean-field scaling. The predictor is given by

$$f(W;x) := \frac{1}{m}\sum_{j=1}^{m} a_j \phi(x^\top W^{(j)}), \quad W^{(j)} \in \mathbb{R}^{d\times K}, \quad j = 1,\dots,m, \tag{6}$$

where the coefficients $a_j \in \{\pm 1\}$ are fixed and $W = (W^{(j)})_{j=1}^m \in \mathbb{R}^{md\times K}$ are the trainable parameters. We train the network with SGD using the rescaled stepsize $\tilde{\eta} := m\eta$, namely

$$W_{t+1} = W_t - \tilde{\eta} g_t.$$

Throughout the section, we assume $\sup_i \|x_i\| \le 1$. This is only for simplicity of presentation and can always be enforced by normalizing the data.

We impose the following conditions on the activation function.

**Assumption 3.1** (Activation function conditions)**.** Let $\phi : \mathbb{R} \to \mathbb{R}$ be continuously differentiable and satisfy the following

1. **Derivative condition.** There exists $\alpha \in \left(\frac{1}{1+\gamma}, 1\right)$ such that $\alpha \le \phi'(z) \le 1$, where $\gamma > 0$ is the margin from Assumption 2.1.
2. **Smoothness.** There exists $\tilde{\beta} > 0$ such that, for all $x, y \in \mathbb{R}$, $|\phi'(x) - \phi'(y)| \le \tilde{\beta}|x - y|$.
3. **Near-homogeneity.** There exists $\kappa > 0$ such that, for every $z \in \mathbb{R}$, $|\phi(z) - \phi'(z)z| \le \kappa$.

Assumption 3.1 is satisfied by several standard activations after a mild leaky modification. The smoothness and near-homogeneity conditions are standard in analyses of optimization for two-layer networks, and are also used in prior work on large-stepsize gradient descent (Cai et al., 2024). The derivative condition ensures that every neuron continues to receive a non-trivial gradient during training; in particular, it rules out the degenerate case where some neurons become effectively frozen.

The lower bound $\alpha > 1/(1+\gamma)$ is the multiclass analogue of the positive derivative lower bound used in the binary setting by Cai et al. (2024). In binary classification, any positive lower bound on $\phi'$ is enough for the usual perceptron-type argument, because each sample contributes only one signed margin inequality. In the multiclass case, each sample must be separated from all $K-1$ incorrect classes, and the activation derivative may be evaluated at different values for different classes. This creates an additional error term of size $1-\alpha$, which is controlled precisely when $\alpha\gamma > 1-\alpha$, or equivalently $\alpha > 1/(1+\gamma)$.

This leads to the effective margin

$$\tilde{\gamma} := \alpha\gamma - (1-\alpha) = \alpha(1+\gamma) - 1. \tag{7}$$

The condition $\alpha > 1/(1+\gamma)$ is exactly what makes $\tilde{\gamma}$ positive. Thus, $\tilde{\gamma}$ plays the role of the margin in the two-layer network analysis. It can be interpreted as the original margin $\gamma$, discounted by the

smallest possible activation derivative, and corrected by the variation of the activation derivatives across classes. In the limit $\alpha \to 1$, this correction disappears and $\tilde{\gamma} \to \gamma$, recovering the usual margin for linear logistic regression.

The next lemma shows that this assumption is not restrictive: leaky versions of several common activations satisfy Assumption 3.1.

**Lemma 3.2** (Leaky activation functions satisfying Assumption 3.1)**.** For a margin $\gamma > 0$, let $c$ satisfy $1/(1+\gamma) < c < 1$. The following leaky variants of common activation functions satisfy Assumption 3.1.

- **GELU, Softplus, SiLU.** Let $\phi$ be GELU, Softplus, or SiLU. Then, $\tilde{\phi}(x) := cx + (1-c)\phi(x)$ satisfies Assumption 3.1 with $\alpha = c$, $\tilde{\beta} = 4(1-c)$, and $\kappa = 1-c$.
- **Huberized ReLU.** Let $\phi$ be the Huberized $\mathrm{ReLU}_h$. Then, $\tilde{\phi}(x) := cx + (1-c)\phi(x)$ satisfies Assumption 3.1 with $\alpha = c$, $\tilde{\beta} = (1-c)/h$, and $\kappa = (1-c)h/2$.
- **Tanh, Sigmoid.** The leaky tanh $\tilde{\phi}(x) := cx + (1-c)\tanh(x)$ and the leaky sigmoid $\tilde{\phi}(x) := cx + \frac{1-c}{1+e^{-x}}$ both satisfy Assumption 3.1 with $\alpha = c$, $\tilde{\beta} = 1-c$, and $\kappa = 1-c$.

Having introduced the assumptions on the activation function, we now state the main convergence result for two-layer neural networks. The theorem shows that the large-stepsize convergence guarantees obtained in the linear case continue to hold despite the nonlinearity of the model.

**Theorem 3.3.** Let Assumptions 2.1 and 3.1 hold. The iterates of (SGD) for the two-layer neural network in (6), with any stepsize $\tilde{\eta} := m\eta$ and batch size $b > 1$, satisfy

$$\min_{0 \le t \le T-1} \mathbb{E}\Big[L(W_t)\Big] \quad \le \quad \frac{K - 1 + 2\ln^2(\tilde{\gamma}^2 \eta T) + 8\kappa^2 + \eta^2\left(1 + \frac{1}{b}\right)^2}{\tilde{\gamma}^2 \eta T}, \tag{8}$$

where $\tilde{\gamma} = \alpha\gamma - (1-\alpha)$.

Theorem 3.3 shows that SGD converges for any stepsize $\eta > 0$ for two-layer neural networks. More precisely, the best iterate achieves a $\widetilde{\mathcal{O}}(1/T)$ convergence rate. This guarantee holds without requiring the trajectory to be in the stable regime; it remains valid even during the edge-of-stability phase, where the loss may be non-monotone and the iterates may oscillate.

To the best of our knowledge, this is the first convergence guarantee for SGD on two-layer neural networks trained with the cross-entropy loss in the large-stepsize regime. The closest related result is the deterministic GD analysis of Cai et al. (2024), which studies two-layer networks under the binary logistic loss. Our result recovers the same asymptotic rate up to constants, while allowing stochastic gradients and multiclass cross-entropy loss.

### 3.1 The Stable Regime

We, next, provide a fine-grained analysis of the dynamics by introducing a stable regime for the iterates similarly to the cross-entropy loss case. As in the cross-entropy loss, the appropriate notion of stability is not based on pointwise monotonicity of the realized loss, since the SGD updates remain stochastic. Instead, we use the Lyapunov viewpoint and define the stable regime as the region where the expected loss admits a one-step descent.

For the two-layer network, this leads to the stable set (refer to Appendix D.5 for the exact derivation)

$$\begin{aligned} \mathcal{S}_{NN} &= \left\{W_t \in \mathbb{R}^{md \times K} : L(W_t) \le \tilde{L}_{NN}\right\}, \\ \tilde{L}_{NN} &= \min\left\{\frac{1}{8\eta(1+\tilde{\beta})\left(1 + \frac{2n}{b\min\{\tilde{\gamma}^2, 1\}}\right)}, \frac{1}{2ne^{\kappa+2}}\right\}. \end{aligned}$$

We now show that SGD enters this stable regime with high probability and that, once inside it, the last iterate enjoys the same sublinear convergence rate.

**Theorem 3.4.** Let Assumptions 2.1 and 3.1 hold. The iterates of (SGD) for the two-layer neural network in (6), with any stepsize $\tilde{\eta} := m\eta$ and batch size $b > 1$, satisfy the following.

1. **Entrance to the stable regime.** For any $\delta \in (0,1)$, there exists $t_{in} \leq t_{max}(\delta)$ such that, with probability at least $1-\delta$, the SGD iterates enter the stable set $\mathcal{S}_{NN}$.
2. **Convergence inside the stable regime.** If the iterates of SGD satisfy $W_k \in \mathcal{S}_{NN}, \forall k \in [t_1, t-1]$, for some $t_1 > 0$, it holds that

$$\mathbb{E}[L(W_t)] \;\leq\; 2\frac{F(W_{t_1}) + \ln^2(\tilde{\gamma}^2\eta\,(t-t_1)) + \kappa^2}{\tilde{\gamma}^2\,\eta\,(t-t_1)},$$

where $F(W) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq y_i} e^{-\langle z_{y_i} - z_j, x_i\rangle}$, $z_i = f(W;x)_i$.

Theorem 3.4 shows that, after entering the stable regime, the last iterate of SGD satisfies a $\widetilde{\mathcal{O}}(1/t)$ convergence rate. This is the nonlinear analogue of the stable-regime guarantee obtained for the linear cross-entropy model. It also matches the rate known for deterministic GD with logistic regression in two-layer networks (Cai et al., 2024), while allowing for stochastic gradients and multiclass cross-entropy loss.

### 3.2 Stabilization in Two-Layer Neural Networks

Interestingly, the stabilization mechanism persists beyond the linear cross-entropy loss setting. Even in the nonlinear two-layer network setting, SGD has the same qualitative behavior: as the loss decreases, the iterates become increasingly likely to remain in the stable set $\mathcal{S}_{NN}$, and even if mini-batch noise temporarily forces them outside, the dynamics return to stability after a controlled number of iterations with high probability.

**Theorem 3.5** (Stabilization of SGD). Let Assumptions 2.1 and 3.1 hold. The iterates of (SGD) with any stepsize $\tilde{\eta} := m\eta$ and batch size $b \geq 1$ satisfy the following.

1. **Maintaining stability.** If $W_t \in \mathcal{S}_{NN}$, then with probability at least $1-\delta'$ it holds that $W_{t+1} \in \mathcal{S}_{NN}$, where $\delta' = \mathcal{O}\left(\exp\left(-\frac{b}{\eta L(W_t)^{3/2}}\right)\right)$.
2. **Return to stability.** If the dynamics exit the stable regime $\mathcal{S}_{NN}$ at time $t_{out} > 0$, then for any $\delta \in (0,1)$, with probability at least $1-\delta$, they return to $\mathcal{S}_{NN}$ in at most $t_{\text{re}} = \tilde{\mathcal{O}}\left(\frac{(\kappa^2+\eta^2)b^2}{\gamma^2\eta\delta\bar{L}}\right)$ number of steps.

## 4 Experiments

We now complement our theoretical results with a series of experiments illustrating the behavior of SGD in the large step-size regime. We conduct three set of experiments to validate our theoretical results. In the first set of experiments, we run SGD with a linear model on a separable synthetic dataset and plot the loss for different step sizes. Each class of points contains $N = 50$ points from $\mathcal{N}(\mu_i, \sigma^2 I)$, where $\mu_i = (\frac{\pi i}{3}, 1), i \in [2], \sigma = 0.05$ are the parameters of the $i-$th class.

In Figure 4, we observe that the loss incurs significant spikes at the start and non-monotonic decrease, indicating the distinct unstable and stable regimes. In addition, we validate the sublinear rate of decrease in the stable regime, validating the results of Theorem 2.3. Lastly, we observe that for larger batch sizes the loss incurs less noise and less oscillations.

The second set of experiments focuses on validating our theory for Neural Networks (NN). In particular, we run SGD on a two-layer NN with leaky softplus activation on the MNIST dataset. In Figure 5, we observe that the loss spikes at the initial unstable regime and then decreases non-monotonically. We, also, verify that the loss converges at an $\mathcal{O}\left(\frac{1}{t}\right)$ rate in the stable regime. Lastly, we observe that larger stepsizes provide faster convergence and, if not better, similar test accuracy.

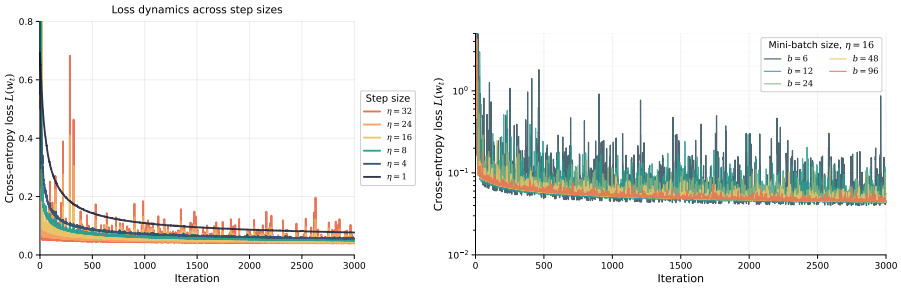


Figure 4: Cross-entropy loss with different stepsizes and batch sizes.

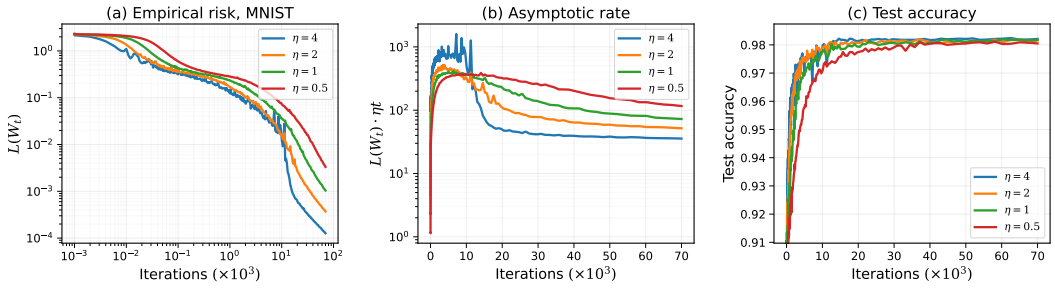


Figure 5: Experiment on MNIST dataset. The dynamics of SGD for training a two-layer NN with leaky softplus activation.

The third set of experiments focuses on the CIFAR-10 dataset, where we test whether our results hold even for larger networks than the two-layer NNs considered in the theoretical results. More specifically, we run SGD on an 8-layer NN with GELU activation for different stepsizes. In figure 1, we observe that the loss decreases non-monotonically at a $\mathcal{O}\left(\frac{1}{t}\right)$ rate. Large stepsizes provide faster decrease of the loss and seem to be synergetic in achieving higher test accuracy. We have not tried to test exhaustively the effect of the stepsize on the test accuracy, as this remains beyond the scope of this paper.

## 5 Conclusion

In this work, we developed a theoretical framework for analyzing SGD for the cross-entropy loss with large learning rates. Unlike deterministic GD, non-monotonicity of the loss in SGD is caused either due to mini-batch noise or due to curvature-driven oscillations, and therefore stability must be defined through a stochastic Lyapunov criterion. Using this viewpoint, we proved convergence guarantees for multiclass cross-entropy loss with arbitrary large step sizes and showed that the dynamics enter a stable regime where the last iterate converges at a sublinear rate. We, further, established the stabilization mechanism of SGD: even when stochastic noise temporarily forces the iterates outside the stable set, the dynamics return to stability in a controlled number of iterations. Finally, we extended this picture to two-layer neural networks and validated the established results experimentally, showing the benefits of SGD with large learning rates.

### AI USE STATEMENT

In this work, we used generative AI tools for helping in editing the manuscript, brainstorming, as well as searching for related literature. Any mathematical claim, in which generative AI has assisted, has been carefully inspected by the authors. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES

Jason M Altschuler and Pablo A Parrilo. Acceleration by random stepsizes: Hedging, equalization, and the arcsine stepsize schedule. *arXiv*, 2024.

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.

Jinho Bok and Jason M Altschuler. Accelerating proximal gradient descent via silver stepsizes. *arXiv*, 2024.

Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: tricks of the trade: second edition*, pages 421–436. Springer, 2012.

Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. In *International Conference on Learning Representations*, 2018.

Yuhang Cai, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter Bartlett. Large stepsize gradient descent for non-homogeneous two-layer networks: Margin improvement and fast optimization. *Advances in Neural Information Processing Systems*, 37:71306–71351, 2024.

Jeremy 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*.

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.

Jeremy M. Cohen, Alex Damian, Ameet Talwalkar, Zico Kolter, and Jason D. Lee. Understanding optimization in deep learning with central flows, 2024. URL https://arxiv.org/abs/2410.24206.

Alex Damian, Eshaan Nichani, and Jason D Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. *ICLR 2023*.

Konstantinos Emmanouilidis, René Vidal, and Nicolas Loizou. Stochastic extragradient with random reshuffling: Improved convergence for variational inequalities. In *International Conference on Artificial Intelligence and Statistics*, pages 3682–3690. PMLR, 2024.

Anastasios Kyrillidis Fangshuo Liao, Afroditi Kolomvaki. Sgd at the edge of stability: The stochastic sharpness gap. 2026.

Rhys Gould and Hidenori Tanaka. Continuous-time analysis of adaptive optimization and normalization, 2024.

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

Desmond J. Higham. Mean-square and asymptotic stability of the stochastic theta method. *SIAM Journal on Numerical Analysis*, 38(3):753–769, 2000.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798. PMLR, 2019.

Rafail Khasminskii. *Stochastic Stability of Differential Equations*. Springer, 2 edition, 2012. doi: 10.1007/978-3-642-23280-0.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Harold J. Kushner. *Stochastic Stability and Control*, volume 33 of *Mathematics in Science and Engineering*. Academic Press, 1967.

Xuerong Mao. *Stochastic Differential Equations and Applications*. Horwood Publishing, 2 edition, 2007.

Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Communications and Control Engineering. Springer London, 1993a. doi: 10.1007/978-1-4471-3267-7.

Sean P. Meyn and Richard L. Tweedie. Stability of markovian processes iii: Foster–lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, 25(3):518–548, 1993b.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019a.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3051–3059. PMLR, 2019b.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

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.

Jingfeng Wu, Peter L. Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency, 2024a.

Jingfeng Wu, Peter L Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 5019–5073. PMLR, 2024b.

Jingfeng Wu, Pierre Marion, and Peter Bartlett. Large stepsizes accelerate gradient descent for regularized logistic regression. *arXiv preprint arXiv:2506.02336*, 2025.

# SUPPLEMENTAL MATERIAL

## CONTENTS


**1 Introduction** **1**

**2 SGD for Cross-entropy Loss** **3**

2.1 Blanket Assumptions . . . . . 3

2.2 Convergence Guarantees . . . . . 3

2.3 Stability in Stochastic Optimization . . . . . 3

2.4 The Stable Regime . . . . . 4

2.5 Stabilization of SGD . . . . . 5

**3 Edge of Stability for Two-Layer Neural Networks** **6**

3.1 The Stable Regime . . . . . 7

3.2 Stabilization in Two-layer Neural Networks . . . . . 8

**4 Experiments** **8**

**5 Conclusion** **9**

**Appendix** **12**

**A Stochastic Lyapunov Stability and the Stable Regime** **13**

**B Additional Related Work** **14**

**C Proof for Cross-entropy Loss** **16**

C.1 Notation . . . . . 16

C.2 Preparatory Lemmas . . . . . 16

C.3 Variance Bound . . . . . 20

C.4 Proofs for EoS Regime . . . . . 21

C.5 Proofs for Stable Regime . . . . . 22

C.6 Proofs for Stochastic Stabilization Mechanism . . . . . 27

**D Proofs for Two-Layer Neural Networks** **31**

D.1 Notation . . . . . 31

D.2 Preparatory Lemmas . . . . . 31

D.3 Proof of Lemma 3.2 for Leaky Activations . . . . . 40

D.4 Proofs for the EoS Regime . . . . . 41

D.5 Proofs for the Stable Regime . . . . . 43

D.6 Proofs for the Stochastic Stabilization Mechanism . . . . . 48

# A Stochastic Lyapunov Stability and the Stable Regime

We provide additional context for the stability notion used in the main paper. The classical stability criterion for deterministic GD is based on monotonic decrease of a Lyapunov function, typically the objective itself. Indeed, if $W_{t+1} = W_t - \eta \nabla L(W_t)$, then the standard descent lemma implies

$$L(W_{t+1}) - L(W_t) \leq -\eta \|\nabla L(W_t)\|^2 + \frac{\eta^2}{2} \lambda_{\max}(\nabla^2 L(W_t)) \|\nabla L(W_t)\|^2.$$

Thus, a sufficient condition for one-step descent is

$$\eta \lambda_{\max}(\nabla^2 L(W_t)) \leq 2.$$

This is the usual deterministic stability threshold. In deterministic edge-of-stability analyses, the violation of this inequality is directly related to curvature-driven oscillations and non-monotone behavior of the loss (Cohen et al., 2021; Damian et al.; Wu et al., 2024b).

For SGD, however, pointwise monotonicity of the realized loss is too strong and is not the right stability requirement. The update

$$W_{t+1} = W_t - \eta g_t, \qquad \mathbb{E}[g_t \mid \mathcal{F}_t] = \nabla L(W_t),$$

contains mini-batch noise. Therefore, even when the stepsize is small and the curvature is benign, the realized loss $L(W_{t+1})$ may be larger than $L(W_t)$ for a particular mini-batch. Consequently, non-monotonicity alone does not distinguish between two qualitatively different phenomena: harmless stochastic fluctuations and genuine curvature-driven instability. A stability notion for SGD must therefore average over the randomness of the next mini-batch while keeping the current iterate fixed.

The natural replacement for deterministic Lyapunov decrease is stochastic Lyapunov decrease. Given the filtration $\mathcal{F}_t$ generated by the trajectory up to time $t \geq 0$, a nonnegative function $\mathcal{E}$ is a stochastic Lyapunov function if it satisfies a one-step drift condition of the form

$$\mathbb{E}[\mathcal{E}(W_{t+1}) - \mathcal{E}(W_t) \mid \mathcal{F}_t] \leq 0.$$

This conditional drift condition is the discrete-time analogue of the Lyapunov criteria used in stochastic stability theory (Kushner, 1967; Khasminskii, 2012; Mao, 2007). Related Foster–Lyapunov drift conditions are also standard in the stability and recurrence theory of Markov chains (Meyn and Tweedie, 1993a). The condition says that $\mathcal{E}$ is not required to decrease for every realization of the stochastic update, but it must decrease on average over the algorithmic randomness at the current point. In our setting, the most appropriate Lyapunov function is the cross-entropy loss itself, namely $\mathcal{E}(W) = L(W)$. This choice is natural for two reasons. First, the goal of the dynamics is to drive the training loss to zero. Second, in the separable cross-entropy setting the parameters need not converge to a finite minimizer; instead, their norm may diverge while the loss goes to zero (Soudry et al., 2018). Hence, stability notions based on distance to a fixed equilibrium point are *not* well aligned with the geometry of the problem.

This Lyapunov stability viewpoint motivates, as discussed in Section 2.3, the definition of the stable set

$$\mathcal{S} := \left\{ W \in \mathbb{R}^{d \times K} : L(W) \leq \frac{1}{\eta c} \right\}.$$

Inside $\mathcal{S}$, SGD may still exhibit random upward fluctuations of the realized loss, but the loss decreases in conditional expectation. This is precisely the stochastic analogue of deterministic Lyapunov stability.

It is useful to contrast this criterion with other standard notions of stochastic stability. Stability in probability requires that trajectories starting near an equilibrium remain near it with high probability; this is the classical stochastic analogue of Lyapunov stability and is commonly formulated for stochastic differential equations and Markov processes (Khasminskii, 2012; Kushner, 1967). Almost-sure stability strengthens this requirement by demanding pathwise stability with probability one, namely that the trajectory remains close to the equilibrium for almost every realization of the noise (Mao, 2007; Khasminskii, 2012). Moment stability requires decay, or at least uniform control, of quantities such as $\mathbb{E}\|W_t - W_\star\|^p$, with the cases $p = 1$ and $p = 2$ corresponding to mean and mean-square stability, respectively (Mao, 2007; Higham, 2000). Exponential moment stability further asks for such moments to decay at an exponential rate, and is often used in the numerical analysis of

stochastic differential equations (Higham, 2000). Finally, Foster–Lyapunov or positive-recurrence criteria require a negative drift outside a suitable compact set and are used to prove recurrence, ergodicity, or existence of invariant distributions for Markov chains and Markov processes (Meyn and Tweedie, 1993a;b). These criteria are powerful in settings with a finite attracting equilibrium, a stationary distribution, or a recurrent compact set. However, they are less suited to separable cross-entropy minimization, where the relevant limiting behavior is not convergence of $W_t$ to a finite point, but rather convergence of $L(W_t)$ to zero while the parameters may diverge in norm.

For this reason, the conditional Lyapunov-drift criterion is the correct notion for our purposes. It is a pointwise criterion, since it evaluates stability at the current iterate; it is genuinely stochastic, since it takes expectation with respect to the next mini-batch; and it is expressed in terms of the loss, which is the relevant convergent quantity in separable cross-entropy minimization. Furthermore, when the mini-batch coincides with the full dataset, the stochastic gradient $g_t$ reduces to the full gradient $\nabla L(W_t)$, and the conditional drift condition recovers the standard deterministic descent criterion, up to constant factors. In this sense, the stable regime notion introduced above extends the deterministic edge-of-stability threshold while also accounting for the fluctuations caused by mini-batch sampling.

## B Additional Related Work

In this appendix, we provide a more detailed discussion of related work. Our work is connected to several lines of research: edge-of-stability dynamics, large-stepsize optimization for logistic and cross-entropy losses as well as implicit bias in separable classification.

**Edge of stability.** The edge-of-stability phenomenon was systematically documented by Cohen et al. (2021), who observed that gradient descent on neural networks often operates near the classical stability threshold predicted by the largest eigenvalue of the Hessian. Classical smooth optimization theory suggests that, for deterministic gradient descent, monotone descent of the loss is guaranteed when the stepsize is smaller than a curvature-dependent threshold. However, empirical evidence shows that modern neural network training often uses stepsizes that exceed this threshold, while the loss still decreases over longer time scales.

Damian et al. proposed a self-stabilization mechanism for deterministic gradient descent at the edge of stability. Their analysis shows that, in certain settings, the dynamics themselves can reduce the effective instability and bring the trajectory back toward a stable region. Relatedly, Cohen et al. (2024) studied edge-of-stability dynamics through the lens of central flows, further clarifying the structure of large-stepsize training. These works focus primarily on deterministic gradient descent. In contrast, our work studies stochastic gradient descent, where non-monotonicity of the loss has two possible sources: curvature-driven oscillations caused by large stepsizes, and random fluctuations caused by mini-batch sampling. This distinction makes the stochastic setting qualitatively different from the deterministic one and motivates a stability notion based on conditional Lyapunov drift rather than pointwise monotone descent.

Another related although distinct mechanism in SGD is the stochastic sharpness gap (Fangshuo Liao, 2026), where the sharpness of SGD equilibrates at a level slightly smaller than $2/\eta$ depending on the batch size. The aforementioned phenomenon is observed primarily in the square loss and differs from the stochastic stabilization mechanism that drives the stochastic dynamics of the cross-entropy loss, making the curvature and noise decrease asymptotically to zero. The unification of all mechanisms into a principled framework remains an interesting open question left for future work.

**Large-stepsize methods for logistic and cross-entropy losses.** A recent line of work has provided rigorous guarantees for large-stepsize gradient descent in classification problems with logistic-type losses. Wu et al. (2024b) showed that, for logistic regression, large stepsizes can accelerate optimization and that non-monotonicity of the loss can be beneficial rather than harmful. Their analysis demonstrates that deterministic gradient descent may pass through a non-monotone edge-of-stability phase before entering a stable regime where the loss decreases at a controlled rate. Cai et al. (2024) extended this perspective to non-homogeneous two-layer networks under binary logistic loss, showing that large stepsizes can improve margin growth and optimization speed. More recent work has continued to refine the theory of large-stepsize training for logistic-type objectives (Wu et al., 2025).

Our work is closest in spirit to this line of research, but differs in three important ways. First, we study SGD rather than full-batch GD. This requires separating curvature-driven instability from mini-batch noise, since the realized loss of SGD may be non-monotone even in regimes where the expected loss decreases. Second, we consider the multiclass cross-entropy loss rather than only the binary logistic loss. The multiclass setting introduces interactions between all incorrect classes through the softmax probabilities, and each example must be separated from every competing class. Third, our stability criterion is stochastic: the stable regime is defined as the region where the cross-entropy loss decreases in conditional expectation over the next mini-batch. Thus, our results can be viewed as a stochastic analogue of the deterministic large-stepsize theory for logistic and cross-entropy losses.

**Large-stepsize methods for Convex Optimization** A recent line of works studies how acceleration can be achieved by using merely large stepsizes in GD. More specifically, Altschuler and Parrilo (2025) show that GD can be accelerated in strongly convex problems by using the silver stepsize schedule. Another work (Grimmer, 2024) studies the effect of long steps in ensuring descent not per iteration of the optimization algorithm but along multiple iterations of the trajectory. For separable convex problems, random stepsizes are shown to accelerate the rate of convergence of GD (Altschuler and Parrilo, 2024), while an extension (Bok and Altschuler, 2024) to proximal GD is also available for smooth convex optimization.

**Implicit bias in separable classification.** Our analysis is also related to the literature on the implicit bias of gradient methods for separable classification. For logistic and cross-entropy losses on separable data, the empirical risk does not have a finite minimizer. Instead, the training loss converges to zero while the norm of the parameters diverges. At the same time, the direction of the parameters may converge to a maximum-margin solution. This phenomenon was established for linear logistic regression by Soudry et al. (2018), and subsequent works refined the convergence rates and extended the analysis to broader settings (Ji and Telgarsky, 2019; Nacson et al., 2019a).

This behavior is important for the stability notion used in our work. Many classical notions of stability are formulated around convergence to, or recurrence near, a finite equilibrium point. Such notions are not directly aligned with separable cross-entropy minimization, where the relevant limiting object is not a finite parameter vector. Instead, the quantity that converges is the loss. For this reason, we use the cross-entropy loss itself as the Lyapunov function. The resulting stable regime is not a neighborhood of a finite minimizer; it is a region in parameter space where the conditional drift of the loss is non-positive. This loss-based viewpoint matches the geometry of separable classification more naturally than equilibrium-based stability criteria.

**Stochastic optimization.** For SGD, Nacson et al. (2019b) proved convergence to zero loss with a fixed non-vanishing learning rate for homogeneous linear classifiers trained on linearly separable data, under both sampling with and without replacement. It is an interesting question to study large step size SGD with random reshuffling (Emmanouilidis et al., 2024) or other without replacement sampling strategies. For neural networks trained with SGD and small step sizes, Brutzkus et al. (2018) showed that SGD can learn overparameterized two-layer networks on linearly separable data. In our case, we extend the two-layer neural network analysis for SGD in the setting of large step sizes.

# C Proof for Cross-entropy Loss

## C.1 Notation

For each sample $(x_i, y_i)$, let

$$\begin{aligned} z_i &= W^\top x_i, \\ p_i &= \mathrm{softmax}(z_i), \end{aligned}$$

where $p_i(j) = \frac{e^{z_i(j)}}{\sum_{\ell=1}^K e^{z_i(\ell)}}$. The gradient and Hessian of the cross-entropy loss evaluated at the sample $(x_i, y_i)$ are given by

$$\nabla \ell_i(W) = (p_i - e_{y_i}) x_i^\top, \tag{9}$$

$$\nabla^2 \ell_i(W) = (\mathrm{diag}(p_i) - p_i p_i^\top) \otimes (x_i x_i^\top). \tag{10}$$

Define the potential functions $F, G : \mathbb{R}^{d\times K} \to \mathbb{R}_{\geq 0}$ with

$$F(W) = \frac{1}{n} \sum_{i=1}^n \sum_{j \neq y_i} e^{-\langle W(y_i) - W(j), x_i \rangle}, \tag{11}$$

$$G(W) = \frac{1}{n} \sum_{i=1}^n (1 - p_i(y_i)). \tag{12}$$

Since $\log(1+u) \leq u, \forall u \in (0, 1]$, it follows that

$$G(W) \leq L(W) \leq F(W). \tag{13}$$

## C.2 Preparatory Lemmas

We provide the analysis of SGD for the multi-class cross-entropy loss with any step size $\eta > 0$. The proof of Theorems 2.2, 2.3, 2.4 relies on a sequence of lemmas that establish key properties of the dynamics and the geometry of the loss landscape. We first introduce the aforementioned lemmas and then present the main theoretical results.

**Lemma C.1** (Perceptron-type inequality). Let Assumption (2.1) hold. Then, for $W \in \mathbb{R}^{d\times K}$ it holds

$$\langle \nabla L(W), W^* \rangle_F \leq -\gamma\, G(W). \tag{14}$$

*Proof.* From (9), we have that

$$\langle \nabla L(W), W^* \rangle_F \overset{(9)}{=} \frac{1}{n} \sum_{i=1}^n \langle (p_i - e_{y_i}) x_i^\top, W^* \rangle_F = \frac{1}{n} \sum_{i=1}^n \langle p_i - e_{y_i}, W^{*\top} x_i \rangle_F.$$

Letting $\Delta_i = W^{*\top} x_i$, we have that

$$\langle p_i - e_{y_i}, \Delta_i \rangle_F = \sum_{j \neq y_i} p_i(j) \big(\Delta_i(j) - \Delta_i(y_i)\big).$$

By Assumption 2.1, we have that $\Delta_i(y_i) - \Delta_i(j) \geq \gamma$, and thus we obtain for $i \in [n]$ that

$$\langle p_i - e_{y_i}, \Delta_i \rangle \leq -\gamma \sum_{j \neq y_i} p_i(j) = -\gamma(1 - p_i(y_i)). \tag{15}$$

Summing for $i = 1, ..., n$ and dividing by $n$, we get

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \langle p_i - e_{y_i}, \Delta_i \rangle &\leq -\frac{\gamma}{n} \sum_{i=1}^n (1 - p_i(y_i)) \\ \iff \langle \nabla L(W), W^* \rangle_F &\leq -\gamma\, G(W). \end{aligned} \tag{16}$$

□

**Lemma C.2.** For every $W \in \mathbb{R}^{d\times K}$, it holds $\forall i \in [n]$ that

$$\left\| \frac{1}{b}\sum_{i=1}^{b} \nabla \ell(W^\top x_i) \right\|_F \leq \frac{\sqrt{2}}{b}\sum_{i=1}^{b} \ell(W^\top x_i) \leq \frac{\sqrt{2}n}{b} L(W), \tag{17}$$

$$\|\nabla L(W)\|_F \leq \sqrt{2}\, G(W) \leq \sqrt{2}\, L(W). \tag{18}$$

*Proof.* It holds that $\nabla \ell(W^\top x_i) = (p_i - e_{y_i})x_i^\top, \nabla L(W) = \frac{1}{n}\sum_{i=1}^{n}(p_i - e_{y_i})x_i^\top$, where $p_i = \text{softmax}(W^\top x_i)$.
We, first, bound the individual gradients. Since $(p_i - e_{y_i})x_i^\top$ is a rank-one matrix, we have

$$\begin{aligned} \|\nabla \ell(W^\top x_i)\|_F &= \|(p_i - e_{y_i})x_i^\top\|_F \\ &= \|p_i - e_{y_i}\|_2 \, \|x_i\|_2 \\ &\leq \|p_i - e_{y_i}\|_2. \end{aligned} \tag{19}$$

Next, it holds that

$$\begin{aligned} \|p_i - e_{y_i}\|_2^2 &= \sum_{j=1}^{K} \big(p_i(j) - \mathbb{1}_{\{j=y_i\}}\big)^2 \\ &= \big(1 - p_i(y_i)\big)^2 + \sum_{j\neq y_i} p_i(j)^2 \\ &\leq \big(1 - p_i(y_i)\big)^2 + \left(\sum_{j\neq y_i} p_i(j)\right)^2 \\ &= \big(1 - p_i(y_i)\big)^2 + \big(1 - p_i(y_i)\big)^2 \\ &= 2\big(1 - p_i(y_i)\big)^2. \end{aligned} \tag{20}$$

Therefore, we get

$$\|p_i - e_{y_i}\|_2 \leq \sqrt{2}\big(1 - p_i(y_i)\big). \tag{21}$$

Combining (19) and (21), we obtain

$$\|\nabla \ell(W^\top x_i)\|_F \leq \sqrt{2}\big(1 - p_i(y_i)\big), \qquad \forall i \in [n]. \tag{22}$$

Since $\ell(W^\top x_i) = -\log p_i(y_i) \geq 1 - p_i(y_i)$ and $L(W) = \frac{1}{n}\sum_{i=1}^{n} \ell(W^T x_i)$, it holds for any $b \geq 1$ that

$$\begin{aligned} \left\| \frac{1}{b}\sum_{i=1}^{b} \nabla \ell(W^\top x_i) \right\|_F &\leq \frac{1}{b}\sum_{i=1}^{b} \|\nabla \ell(W^\top x_i)\|_F \\ &\leq \frac{\sqrt{2}}{b}\sum_{i=1}^{b} \big(1 - p_i(y_i)\big) \\ &\leq \frac{\sqrt{2}}{b}\sum_{i=1}^{b} \ell(W^\top x_i) \\ &\leq \frac{\sqrt{2}n}{b} L(W). \end{aligned}$$

We now turn to the full gradient. From triangle inequality, we have that

$$\begin{aligned}\|\nabla L(W)\|_F &= \Big\|\frac{1}{n}\sum_{i=1}^{n}(p_i - e_{y_i})x_i^\top\Big\|_F \\ &\le \frac{1}{n}\sum_{i=1}^{n}\|(p_i - e_{y_i})x_i^\top\|_F \\ &= \frac{1}{n}\sum_{i=1}^{n}\|\nabla\ell(W^\top x_i)\|_F \\ &\le \frac{\sqrt{2}}{n}\sum_{i=1}^{n}\big(1 - p_i(y_i)\big) \\ &= \sqrt{2}\,G(W). \end{aligned} \tag{23}$$

Finally, since $G(W) \le L(W)$, we conclude that

$$\|\nabla L(W)\|_F \le \sqrt{2}\,G(W) \le \sqrt{2}\,L(W). \tag{24}$$

This completes the proof. □

**Lemma C.3.** For all $W \in \mathbb{R}^{d\times K}$, it holds that

$$\|\nabla^2 L(W)\|_{\text{op}} \le 2L(W). \tag{25}$$

*Proof.* From (10) and using the fact that the operator norm of a Kronecker product satisfies $\|A\otimes B\| = \|A\|\,\|B\|$, we obtain

$$\begin{aligned}\|\nabla^2\ell_i(W)\|_{\text{op}} &= \|\text{diag}(p_i) - p_ip_i^\top\|_{\text{op}}\|x_ix_i^\top\|_{\text{op}} \\ &\overset{\|x\|_2\le 1}{\le} \|\text{diag}(p_i) - p_ip_i^\top\|_{\text{op}}. \end{aligned} \tag{26}$$

For the positive semidefinite matrix $A_i := \text{diag}(p_i) - p_ip_i^\top$, it holds that

$$\begin{aligned}\|A_i\|_{\text{op}} &\le tr(H_i) = 1 - p_i^2 = 1 - p_i(y_i)^2 - \sum_{j\ne i}p_i(j)^2 \\ &\le 1 - p_i(y_i)^2 = (1 - p_i(y_i))(1 + p_i(y_i)) \\ &\le 2(1 - p_i(y_i)) \end{aligned}$$

Using the inequality $-\log z \ge 1 - z, \forall z > 0$ and the fact that $\ell_i(W) = -\log p_i(y_i)$ , we have that

$$1 - p_i(y_i) \le \ell_i(W). \tag{27}$$

Hence, we obtain

$$\|A_i\|_{\text{op}} \le 2(1 - p_i(y_i)) \le 2\,\ell_i(W). \tag{28}$$

Combining (26) and (28), we obtain

$$\|\nabla^2\ell_i(W)\|_{\text{op}} \le 2\,\ell_i(W). \tag{29}$$

Summing over $i = 1, \dots, n$ and dividing by $n$, we have that

$$\|\nabla^2 L(W)\|_{\text{op}} \le \frac{1}{n}\sum_{i=1}^{n}\|\nabla^2\ell_i(W)\| \le \frac{2}{n}\sum_{i=1}^{n}\ell_i(W) = 2\,L(W). \tag{30}$$

□

**Lemma C.4.** The iterates of SGD with any step size $\eta > 0$ satisfy that

$$\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\Big[G(W_k)\Big] \leq \frac{\sqrt{K-1} + 2\ln\big(\gamma^2\eta t\big) + 2\eta\left(1+\frac{1}{b}\right)}{\eta\gamma^2 t}$$

where $G(W) = \frac{1}{n}\sum_{i=1}^{n}(1 - p_i(y_i))$.

*Proof.* From Lemma C.1, the gradient of the softmax loss satisfies

$$\langle \nabla L(W_t), W^* \rangle_F \;\leq\; -\gamma\, G(W_t), \tag{31}$$

where $G(W_t) = \frac{1}{n}\sum_{i=1}^{n}(1 - p_i(y_i))$. Using the (SGD) update $W_{t+1} = W_t - \eta g_t$, we get

$$\langle W_{t+1}, W^* \rangle_F \;=\; \langle W_t, W^* \rangle_F - \eta \langle g_t, W^* \rangle_F$$

Taking expectation condition on the filtration $\mathcal{F}_t$ and using that $\mathbb{E}\left[g_t \middle| \mathcal{F}_t\right] = \nabla L(W_t)$, we have that

$$\begin{aligned}\mathbb{E}\left[\langle W_{t+1}, W^* \rangle_F \middle| \mathcal{F}_t\right] &= \langle W_t, W^* \rangle_F - \eta \langle \nabla L(W_t), W^* \rangle_F \\ &\overset{(31)}{\geq} \langle W_t, W^* \rangle_F + \eta\gamma G(W_t)\end{aligned} \tag{32}$$

where at the last step we have used Lemma C.1. Taking expectation again and using the tower law of expectation, we get

$$\mathbb{E}\Big[\langle W_{t+1}, W^* \rangle_F - \langle W_t, W^* \rangle_F\Big] \;\geq\; \eta\gamma\, \mathbb{E}\Big[G(W_t)\Big]$$

Unrolling the recursion and multiplying by $\frac{1}{t}$, we obtain

$$\frac{1}{t}\,\mathbb{E}\Big[\langle W_t, W^* \rangle_F\Big] \;\geq\; \frac{\eta\gamma}{t}\sum_{k=0}^{t-1}\mathbb{E}\Big[G(W_k)\Big], \tag{33}$$

where we have used the fact that $W_0 = \mathbf{0}$. Since $\|W^*\|_F = 1$ and from applying the Cauchy-Schwarz inequality it holds $\|W_t\|_F \geq \langle W_t, W^* \rangle_F$, we obtain

$$\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\Big[G(W_k)\Big] \leq \frac{\mathbb{E}\Big[\|W_t\|_F\Big]}{\eta\gamma t} \tag{34}$$

We, next, bound the $\mathbb{E}\Big[\|W_t\|_F\Big]$. Applying the triangle inequality, we have that

$$\|W_t\|_F \leq \|W_t - U\|_F + \|U\|_F \tag{35}$$

Thus, it suffices to bound the terms $\|W_t - U\|_F, \|U\|_F$. From Proposition C.7, for any decomposition $U = U_1 + U_2$ with $U_2 = \frac{\eta}{\gamma}\left(1+\frac{1}{b}\right)W^*$, it holds that

$$\begin{aligned}\frac{\mathbb{E}\Big[\|W_t - U\|_F^2\Big]}{2\eta t} + \frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\Big[L(W_k)\Big] &\leq L(U_1) + \frac{\|W_0 - U\|_F^2}{2\eta t} \\ \Rightarrow \mathbb{E}\Big[\|W_t - U\|_F^2\Big] &\leq 2\eta t L(U_1) + \|U\|_F^2,\end{aligned} \tag{36}$$

where we have used the fact that $W_0 = \mathbf{0}$ and the non-negativity of the loss $L$. Selecting $U_1 = \alpha W^*$ with $\alpha = \frac{\ln(\gamma^2\eta t)}{\gamma}$. Then, it holds

$$L(U_1) \;\leq\; (K-1)e^{-\alpha\gamma} \;=\; \frac{K-1}{\gamma^2\eta t} \tag{37}$$

Thus, substituting into (36) and using Jensen inequality, we have that

$$\mathbb{E}\Big[\|W_t - U\|_F\Big] \leq \sqrt{\mathbb{E}\Big[\|W_t - U\|_F^2\Big]} \leq \sqrt{2\eta t L(U_1) + \|U\|_F^2} \leq \sqrt{\frac{K-1}{\gamma^2} + \|U\|_F^2} \tag{38}$$

We, next, bound the norm $\|U\|_F$. It holds that

$$\|U\|_F^2 = \left\|\alpha W^* + \frac{\eta}{\gamma}\left(1+\frac{1}{b}\right)W^*\right\|_F^2 \leq \frac{1}{\gamma^2}\left[\ln^2(\gamma^2\eta t) + \eta^2\left(1+\frac{1}{b}\right)^2\right]$$
$$\implies \quad \|U\|_F \leq \frac{1}{\gamma}\left[\ln(\gamma^2\eta t) + \eta\left(1+\frac{1}{b}\right)\right]. \tag{39}$$

By the triangle inequality,

$$\|W_t\|_F \leq \|W_t - U\|_F + \|U\|_F \;\leq\; \frac{\sqrt{2}}{\gamma} + 2\|U\|_F \;\leq\; \frac{\sqrt{K-1} + 2\ln(\gamma^2\eta t) + 2\eta\left(1+\frac{1}{b}\right)}{\gamma} \tag{40}$$

where we have used the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$ and the bound on $\|U\|_F$ from (39). Substituting (40) into (34) gives

$$\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\Big[G(W_k)\Big] \leq \frac{\sqrt{K-1} + 2\ln(\gamma^2\eta t) + 2\eta\left(1+\frac{1}{b}\right)}{\eta\gamma^2 t} \tag{41}$$

□

**Lemma C.5.** The stochastic gradient oracle $g_t$ of the cross-entropy loss satisfies

$$\mathbb{E}\left[\|g_t\|^2\middle|\mathcal{F}_t\right] \leq \left(1+\frac{n}{b\gamma^2}\right)\|\nabla L(W_t)\|^2$$

*Proof.* From the bias-variance decomposition, it holds

$$\begin{aligned}\mathbb{E}\left[\|g_t\|^2\middle|\mathcal{F}_t\right] &= \|\nabla L(W_t)\|_F^2 + \mathbb{E}\left[\|g_t - \nabla L(W_t)\|_F^2\middle|\mathcal{F}_t\right] \\ &= \|\nabla L(W_t)\|_F^2 + \frac{1}{b}\operatorname{Var}(\nabla\ell_i(W_t) \mid \mathcal{F}_t) \\ &\leq \|\nabla L(W_t)\|_F^2 + \frac{1}{b}\,\mathbb{E}\left[\|\nabla\ell_i(W_t)\|_F^2\middle|\mathcal{F}_t\right]\end{aligned}$$

as $g_t$ is the average of $b$ i.i.d. samples (conditionally on $\mathcal{F}_t$). Substituting the definition of $\mathbb{E}\left[\|\nabla\ell_i(W_t)\|_F^2\middle|\mathcal{F}_t\right]$ and using the inequality $\sum_{i=1}^n\|\nabla\ell_i(W_t)\|_F^2 \leq \frac{n^2}{\gamma^2}\|\nabla L(W_t)\|_F^2$, from Nacson et al. (2019b), we obtain

$$\mathbb{E}\left[\|g_t\|^2\middle|\mathcal{F}_t\right] \leq \|\nabla L(W_t)\|_F^2 + \frac{1}{nb}\sum_{i=1}^n\|\nabla\ell_i(W_t)\|_F^2 \;\leq\; \left(1+\frac{n}{b\gamma^2}\right)\|\nabla L(W_t)\|_F^2$$

□

### C.3 Variance Bound

As a stochastic optimization algorithm, SGD is leveraging a stochastic mini-batch gradient at each iteration instead of the full-batch gradient. Controlling the noise induced by the stochastic oracles in the dynamics is of critical importance for analyzing the convergence. We can bound the variance of the stochastic mini-batches by utilizing the following lemma.

**Lemma C.6** (Variance bound)**.** Let $g(W) = \frac{1}{b}\sum_{i\in\mathcal{B}_t}\nabla\ell(f(x_i;W),y_i)$ denote the mini-batch gradient. Then, it holds that

$$\mathbb{E}\Big[\|g(W) - \nabla L(W)\|^2\Big] \;\leq\; \frac{2}{b}\,G(W), \tag{42}$$

where $G(W) := \frac{1}{n}\sum_{i=1}^n\Big[1 - \operatorname{softmax}(Wx_i)_{y_i}\Big]$.

*Proof.* We begin by decomposing the variance of the stochastic oracle. For uniform sampling of minibatch $B_t$ of size $b \geq 1$, it holds that

$$\mathbb{E}\left[\|g_t - \nabla L(W_t)\|_F^2 \mid \mathcal{F}_t\right] = \frac{1}{b}\operatorname{Var}_i\left(\nabla \ell_i(W_t)\right) \leq \frac{1}{b}\mathbb{E}_i\left[\|\nabla \ell_i(W_t)\|_F^2\right], \tag{43}$$

where we used that $\operatorname{Var}(Z) \leq \mathbb{E}[\|Z\|^2]$ for any random variable $Z$. For each $i \in [n]$, by the definition of $\nabla \ell_i(W_t)$ we have

$$\|\nabla \ell_i(W_t)\|_F^2 = \|(p_i - e_{y_i})x_i^\top\|_F^2 = \|p_i - e_{y_i}\|_2^2 \|x_i\|_2^2 \leq \|p_i - e_{y_i}\|_2^2, \tag{44}$$

since $\|x_i\|_2 \leq 1$. We next expand the norm, obtaining

$$\|p_i - e_{y_i}\|_2^2 = (1 - p_i(y_i))^2 + \sum_{j \neq y_i} p_i(j)^2 \leq (1 - p_i(y_i))^2 + \left(\sum_{j \neq y_i} p_i(j)\right)^2 \leq 2(1 - p_i(y_i))^2, \tag{45}$$

where we used that and $\sum_{j \neq y_i} p_i(j) = 1 - p_i(y_i)$. Thus, from (44), (45) we have

$$\|\nabla \ell_i(W_t)\|_F^2 \leq 2(1 - p_i(y_i))^2 \leq 2(1 - p_i(y_i)). \tag{46}$$

Summing for $i \in [n]$ and dividing by $\frac{1}{n}$ gives

$$\mathbb{E}_i[\|\nabla \ell_i(W_t)\|_F^2] \leq 2\,G(W_t). \tag{47}$$

Substituting (47) into (43), we obtain the final

$$\mathbb{E}\left[\|g_t - \nabla L(W_t)\|_F^2 \mid \mathcal{F}_t\right] \leq \frac{2}{b}G(W_t). \tag{48}$$

□

### C.4 Proofs for EoS Regime

**Lemma C.7.** Let $U = U_1 + U_2$ with $U_1 \in \mathbb{R}^{d \times K}, U_2 = \frac{\eta}{\gamma}\left(1 + \frac{1}{b}\right) W^*$. Then, for all $t \geq 1$, it holds that

$$\frac{\mathbb{E}\left[\|W_t - U\|_F^2\right]}{2\eta t} + \frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[L(W_k)\right] \leq L(U_1) + \frac{\|W_0 - U\|_F^2}{2\eta t} \tag{49}$$

*Proof.* From the update rule (SGD), we have that

$$\begin{aligned}\|W_{t+1} - U\|_F^2 &= \|W_t - U\|_F^2 + 2\eta\langle g_t, U - W_t\rangle_F + \eta^2\|g_t\|_F^2 \\ &= \|W_t - U\|_F^2 + 2\eta\langle g_t, U_1 - W_t\rangle_F + \eta^2\left(\frac{2}{\eta}\langle g_t, U_2\rangle + \|g_t\|_F^2\right)\end{aligned} \tag{50}$$

Taking expectation condition on the filtration $\mathcal{F}_t$ and using the unbiasedness property of the stochastic oracles, we have

$$\begin{aligned}\mathbb{E}\left[\|W_{t+1} - U\|_F^2 \middle| \mathcal{F}_t\right] &= \|W_t - U\|_F^2 + 2\eta\langle \nabla L(W_t), U_1 - W_t\rangle_F \\ &\quad + \eta^2\left(\frac{2}{\eta}\langle \nabla L(W_t), U_2\rangle + \mathbb{E}\left[\|g_t\|_F^2 \middle| \mathcal{F}_t\right]\right)\end{aligned} \tag{51}$$

Using Lemmas C.1, C.6 with the selected $U_2 = \frac{\eta}{\gamma}\left(1 + \frac{1}{b}\right) W^*$, we have

$$\begin{aligned}\frac{2}{\eta}\langle \nabla L(W_t), U_2\rangle + \mathbb{E}\left[\|g_t\|_F^2 \middle| \mathcal{F}_t\right] &\overset{(15)}{\leq} -2\left(1 + \frac{1}{b}\right)G(W_t) + \mathbb{E}\left[\|g_t\|_F^2 \middle| \mathcal{F}_t\right] \\ &= -2\left(1 + \frac{1}{b}\right)G(W_t) + \|\nabla L(W_t)\|_F^2 \\ &\quad + \mathbb{E}\left[\|g_t - \nabla L(W_t)\|_F^2 \middle| \mathcal{F}_t\right] \\ &\overset{\text{Lemma C.6}}{\leq} -2\left(1 + \frac{1}{b}\right)G(W_t) + \|\nabla L(W_t)\|_F^2 + \frac{2}{b}G(W_t) \\ &\overset{\text{Lemma C.1}}{\leq} -2\left(1 + \frac{1}{b}\right)G(W_t) + 2\left(1 + \frac{1}{b}\right)G(W_t) \\ &\leq 0,\end{aligned} \tag{52}$$

Thus, substituting (52) into (51) gives

$$\mathbb{E}\left[\|W_{t+1}-U\|_F^2\middle|\mathcal{F}_t\right] \quad \leq \quad \|W_t-U\|_F^2+2\eta\langle\nabla L(W_t),U_1-W_t\rangle_F$$

Using the fact that $L$ is convex, we have that

$$\mathbb{E}\left[\|W_{t+1}-U\|_F^2\middle|\mathcal{F}_t\right] \quad \leq \quad \|W_t-U\|_F^2+2\eta\left[L(U_1)-L(W_t)\right]$$

Taking expectation again and using the tower law of expectation, we obtain

$$\mathbb{E}\Big[\|W_{t+1}-U\|_F^2\Big] \quad \leq \quad \mathbb{E}\Big[\|W_t-U\|_F^2\Big]+2\eta\,\mathbb{E}\Big[L(U_1)-L(W_t)\Big]$$

Summing for $k=0,...,t-1$ and dividing by $2\eta t$, we have that

$$\frac{\mathbb{E}\Big[\|W_t-U\|_F^2\Big]}{2\eta t}+\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\Big[L(W_k)\Big]\leq L(U_1)+\frac{\|W_0-U\|_F^2}{2\eta t}$$

□

PROOF OF THEOREM 2.2

*Proof.* From Lemma C.7, we have that for all $t\geq 1$ it holds

$$\frac{\mathbb{E}\Big[\|W_t-U\|_F^2\Big]}{2\eta t}+\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\Big[L(W_k)\Big]\leq L(U_1)+\frac{\|U\|_F^2}{2\eta t} \tag{53}$$

For $U_1=\alpha W^*$, $U_2=\frac{\eta}{\gamma}\left(1+\frac{1}{b}\right)W^*$, $\alpha=\frac{\ln\left(\gamma^2\eta t\right)}{\gamma}$, it holds that

$$L(U_1) \quad \leq \quad F(U_1)=(K-1)e^{-\alpha\gamma}=\frac{K-1}{\gamma^2\eta t}, \tag{54}$$

$$\|U\|_F^2 \quad \leq \quad 2\|U_1\|_F^2+2\|U_2\|_F^2\leq\frac{2\ln^2(\gamma^2\eta t)}{\gamma^2}+\frac{2\eta^2(1+\frac{1}{b})^2}{\gamma^2}. \tag{55}$$

Substituting (54), (55) into (53) and rearranging the terms, we obtain

$$\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\Big[L(W_k)\Big]\leq\frac{K-1+\ln^2(\gamma^2\eta t)+\eta^2(1+\frac{1}{b})^2}{\gamma^2\eta t}$$

□

### C.5 PROOFS FOR STABLE REGIME

**Lemma C.8.** If for some $t\geq 0$ the iterates of (SGD) satisfy

$$L(W_t)\leq\frac{1}{8\eta\left(1+\frac{n}{b\min\{\gamma^2,1\}}\right)},$$

then

$$\mathbb{E}\left[L(W_{t+1})-L(W_t)\middle|\mathcal{F}_t\right]\leq 0, \tag{56}$$

and hence $W_t\in\mathcal{S}$.

*Proof.* Recall that the minibatch gradient is $g_t=\frac{1}{b}\sum_{j\in\mathcal{B}_t}\nabla\ell_j(W_t)$. Fix $i\in[n]$ and define the logit increment

$$\Delta s_i:=(W_{t+1}-W_t)x_i=-\eta g_t x_i.$$

Since $\|x_i\|_2 \le 1$, we have

$$\|\Delta s_i\|_2 \le \eta \|g_t\|_F \le \frac{\eta}{b} \sum_{j \in B_t} \|\nabla \ell_j(W_t)\|_2 \overset{(17)}{\le} \frac{\sqrt{2}\eta n}{b} L(W_t). \tag{57}$$

where we have used (17) from Lemma C.2. Let $\phi_i(\alpha) := \ell_i(W_t - \alpha\eta g_t)$ for $\alpha \in [0,1]$. From Taylor's theorem, there exists $\theta_i \in (0,1)$ such that

$$\ell_i(W_{t+1}) = \ell_i(W_t) - \eta \langle \nabla \ell_i(W_t), g_t \rangle + \frac{1}{2} (\Delta s_i)^\top H_i(\theta_i)(\Delta s_i), \tag{58}$$

where $H_i(\theta_i) = \operatorname{diag}(p_i^\theta) - p_i^\theta (p_i^\theta)^\top$, and $p_i^\theta := \operatorname{softmax}\big((W_t - \theta_i \eta g_t) x_i\big)$. Since $H_i(\theta_i)$ is positive semidefinite, we have that

$$(\Delta s_i)^\top H_i(\theta_i)(\Delta s_i) \le \|H_i(\theta_i)\|_{\text{op}} \|\Delta s_i\|_2^2.$$

Hence, from (57) we obtain

$$\ell_i(W_{t+1}) \le \ell_i(W_t) - \eta \langle \nabla \ell_i(W_t), g_t \rangle + \frac{\eta^2}{2} \|H_i(\theta_i)\|_{\text{op}} \|g_t\|_F^2. \tag{59}$$

For the softmax cross-entropy loss, if $u \in \mathbb{R}^K$ and $p = \operatorname{softmax}(u)$ it holds that

$$\operatorname{diag}(p) - pp^\top \succeq 0 \qquad \text{and} \qquad \|\operatorname{diag}(p) - pp^\top\|_{\text{op}} \le \operatorname{tr}\big(\operatorname{diag}(p) - pp^\top\big) = 1 - \|p\|_2^2.$$

Therefore, we have that

$$\|H_i(\theta_i)\|_{\text{op}} \le 1 - \|p_i^\theta\|_2^2 \le 2\big(1 - p_i^\theta(y_i)\big).$$

Since $-\ln a \ge 1 - a, \forall a \in (0,1]$, we have that $1 - p_i^\theta(y_i) \le \ell_i(W_t - \theta_i \eta g_t)$ and thus

$$\|H_i(\theta_i)\|_{\text{op}} \le 2\, \ell_i(W_t - \theta_i \eta g_t).$$

In the stable-regime, the loss is small enough so that the logits move by at most a constant amount. In particular, it holds that $\|\Delta s_i\|_\infty \le 1$. Then, the softmax probabilities along the segment between $W_t$ and $W_{t+1}$ change by at most an absolute constant factor, and hence

$$\ell_i(W_t - \theta_i \eta g_t) \le e^2 \ell_i(W_t).$$

Therefore, we have that

$$\|H_i(\theta_i)\|_{\text{op}} \le 2e^2 \ell_i(W_t) \le 16\, \ell_i(W_t),$$

where we used $e^2 < 8$. Substituting into (59), we get

$$\ell_i(W_{t+1}) \le \ell_i(W_t) - \eta \langle \nabla \ell_i(W_t), g_t \rangle + 8\eta^2 \ell_i(W_t) \|g_t\|_F^2. \tag{60}$$

Summing (60) over $i = 1, \dots, n$ and dividing by $n$, we have

$$L(W_{t+1}) - L(W_t) \le -\eta \langle \nabla L(W_t), g_t \rangle + 8\eta^2 L(W_t) \|g_t\|_F^2. \tag{61}$$

Taking conditional expectation with respect to $\mathcal{F}_t$ and using $\mathbb{E}\left[g_t \middle| \mathcal{F}_t\right] = \nabla L(W_t)$, we obtain

$$\mathbb{E}\left[L(W_{t+1}) - L(W_t) \middle| \mathcal{F}_t\right] \le -\eta \|\nabla L(W_t)\|_F^2 + 8\eta^2 L(W_t)\, \mathbb{E}\left[\|g_t\|_F^2 \middle| \mathcal{F}_t\right]. \tag{62}$$

Using Lemma C.5, we can bound the last term by $\mathbb{E}\left[\|g_t\|^2 \middle| \mathcal{F}_t\right] \le \left(1 + \frac{n}{b\gamma^2}\right) \|\nabla L(W_t)\|^2$ and thus have

$$\mathbb{E}\left[L(W_{t+1}) - L(W_t) \middle| \mathcal{F}_t\right] \le -\eta \left[1 - 8\eta \left(1 + \frac{n}{b\gamma^2}\right) L(W_t)\right] \|\nabla L(W_t)\|_F^2. \tag{63}$$

For $L(W_t) \le \frac{1}{8\eta\left(1 + \frac{n}{b\gamma^2}\right)}$, we have that

$$\mathbb{E}\left[L(W_{t+1}) - L(W_t) \middle| \mathcal{F}_t\right] \le 0.$$

Taking full expectation and applying the tower property of expectation, we get

$$\mathbb{E}\Big[L(W_{t+1}) - L(W_t)\Big] \le 0.$$

Combining the two requirements for the loss $L(W_t) \le \min\left\{\frac{1}{8\eta\left(1 + \frac{n}{b\gamma^2}\right)}, \frac{b}{\sqrt{2}n\eta}\right\}$ it suffices to have $L(W_t) \le \frac{1}{8\eta\left(1 + \frac{n}{b \min\{\gamma^2, 1\}}\right)}$. □

**Lemma C.9.** If the iterates satisfy $W_k \in \mathcal{S}, \forall k \in [t_1, t)$, then it holds that

$$\mathbb{E}\Big[L(W_t)\Big] \quad \leq \quad \frac{8F(W_{t_1}) + 4\ln^2(\gamma^2\eta(t-t_1))}{7\gamma^2\eta(t-t_1)} \tag{64}$$

where $F(W) = \frac{1}{n}\sum_{i=1}^n \sum_{j\neq y_i} e^{-\langle W(y_i) - W(j), x_i\rangle}$.

*Proof.* From the update rule of (SGD) we have that

$$\|W_{t+1} - U\|_F^2 \quad = \quad \|W_t - U\|_F^2 - 2\eta\langle g(W_t), W_t - U\rangle_F + \eta^2\|g(W_t)\|_F^2.$$

Taking expectation condition on the filtration $\mathcal{F}_t$ and using the unbiased property $\mathbb{E}\left[g(W_t)\middle|\mathcal{F}_t\right] = \nabla L(W_t)$, we have

$$\mathbb{E}\left[\|W_{t+1} - U\|_F^2\middle|\mathcal{F}_t\right] \quad = \quad \|W_t - U\|_F^2 - 2\eta\langle \nabla L(W_t), W_t - U\rangle_F + \eta^2\,\mathbb{E}\left[\|g(W_t)\|_F^2\middle|\mathcal{F}_t\right] \tag{65}$$

Since $L$ is convex, we get

$$\begin{aligned} & L(U) \geq L(W_t) + \langle \nabla L(W_t), U - W_t\rangle_F \\ \iff \quad & -\langle \nabla L(W_t), W_t - U\rangle_F \leq L(U) - L(W_t) \end{aligned} \tag{66}$$

Substituting (66) into (65), we obtain

$$\|W_{t+1} - U\|_F^2 \quad \leq \quad \|W_t - U\|_F^2 + 2\eta\big(L(U) - L(W_t)\big) + \eta^2\,\mathbb{E}\left[\|g(W_t)\|_F^2\middle|\mathcal{F}_t\right]$$

Taking expectation on both sides and using the tower law of expectation, we get

$$\mathbb{E}\Big[\|W_{t+1} - U\|_F^2\Big] \quad \leq \quad \mathbb{E}\Big[\|W_t - U\|_F^2\Big] + 2\eta\,\mathbb{E}\Big[L(U) - L(W_t)\Big] + \eta^2\,\mathbb{E}\Big[\|g(W_t)\|_F^2\Big]. \tag{67}$$

Using the bias-variance decomposition and Lemma C.6, we have that

$$\begin{aligned} \eta^2\,\mathbb{E}\left[\|g(W_t)\|_F^2\middle|\mathcal{F}_t\right] \quad &= \quad \eta^2\|\nabla L(W_t)\|_F^2 + \eta^2\,\mathbb{E}\left[\|g(W_t) - \nabla L(W_t)\|_F^2\middle|\mathcal{F}_t\right] \\ &= \quad \eta^2\|\nabla L(W_t)\|_F^2 + \frac{\eta^2}{b}\,\mathrm{Var}(\nabla \ell_i(W_t) \mid \mathcal{F}_t) \\ &\leq \quad \eta^2\|\nabla L(W_t)\|_F^2 + \frac{\eta^2}{b}\,\mathbb{E}\left[\|\nabla \ell_i(W_t)\|_F^2\middle|\mathcal{F}_t\right] \\ &\leq \quad \eta^2\|\nabla L(W_t)\|_F^2 + \frac{\eta^2}{nb}\sum_{i=1}^n \|\nabla \ell_i(W_t)\|_F^2 \end{aligned}$$

Using the fact that $\sum_{i=1}^n \|\nabla \ell_i(W_t)\|_F^2 \leq \frac{n^2}{\gamma^2}\|\nabla L(W_t)\|_F^2$, we have

$$\begin{aligned} \eta^2\,\mathbb{E}\left[\|g(W_t)\|_F^2\middle|\mathcal{F}_t\right] \quad &\leq \quad \eta^2\left(1 + \frac{n}{b\gamma^2}\right)\|\nabla L(W_t)\|_F^2 \\ &\overset{\text{Lemma } C.2}{\leq} \quad 2\eta^2\left(1 + \frac{n}{b\gamma^2}\right)L(W_t)^2 \\ &\leq \quad \frac{\eta}{4}L(W_t). \end{aligned} \tag{68}$$

where at the last step we have used the fact that $L(W_t) \leq \frac{1}{4\eta C}$. Thus, (67) becomes

$$\begin{aligned} \mathbb{E}\left[\|W_{t+1} - U\|_F^2\middle|\mathcal{F}_t\right] \quad &\leq \quad \|W_t - U\|_F^2 + 2\eta[L(U) - L(W_t)] + \frac{\eta}{4}L(W_t) \\ &\leq \quad \|W_t - U\|_F^2 + 2\eta L(U) - \frac{7\eta}{4}L(W_t) \end{aligned} \tag{69}$$

Taking expectation again, using the tower law of expectation and rearranging the terms, we obtain

$$\mathbb{E}\Big[L(W_t)\Big] \leq \frac{8}{7}L(U) + \frac{4\,\mathbb{E}\Big[\|W_t - U\|_F^2 - \|W_{t+1} - U\|_F^2\Big]}{7\eta}. \tag{70}$$

Summing over $k = t_1, \ldots, t-1$ and dividing by $(t - t_1)$, we have that:

$$\begin{aligned}\frac{1}{t-t_1}\sum_{k=t_1}^{t-1}\mathbb{E}\Big[L(W_k)\Big] &\leq \frac{8}{7}L(U) + \frac{4}{7\eta\,(t-t_1)}\sum_{k=t_1}^{t-1}\left(\|W_k - U\|_F^2 - \|W_{k+1} - U\|_F^2\right) \\ &= \frac{8}{7}L(U) + \frac{4\left(\|W_{t_1} - U\|_F^2 - \|W_t - U\|_F^2\right)}{7\eta\,(t-t_1)} \end{aligned} \tag{71}$$

Letting $U = W_{t_1} + U_1$ with $U_1 = \alpha W^*$ and $\alpha = \frac{\ln\left(\gamma^2\eta(t-t_1)\right)}{\gamma}$ we have that

$$\|W_{t_1} - U\|_F^2 = \|U_1\|_F^2 = \frac{\ln^2(\gamma^2\eta(t-t_1))}{\gamma^2} \tag{72}$$

and

$$\begin{aligned} L(U) &\leq F(U) = \frac{1}{n}\sum_{i=1}^{n} e^{-\langle U, x\rangle_F} \\ &= \frac{1}{n}\sum_{i=1}^{n} e^{-\langle U_1, X\rangle_F} e^{-\langle W_s, X\rangle_F} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{e^{-\langle W_s, X\rangle_F}}{\gamma^2\eta(t-t_1)} \\ &= \frac{F(W_s)}{\gamma^2\eta(t-t_1)} \end{aligned} \tag{73}$$

Substituting (72), (73) into (71), we obtain

$$\min_{t_1 \leq k \leq t-1}\mathbb{E}\Big[L(W_k)\Big] \leq \frac{8F(W_{t_1}) + 4\ln^2(\gamma^2\eta(t-t_1))}{7\gamma^2\eta(t-t_1)} \tag{74}$$

Applying Lemma C.8 for the stable region $\mathcal{S}_{[t_1, t-1]}$ we have that

$$\mathbb{E}\Big[L(W_t)\Big] \leq \mathbb{E}\Big[L(W_{t-1})\Big] \leq \ldots \leq \mathbb{E}\Big[L(W_{t_1})\Big]$$

and thus from (74) we have that

$$\mathbb{E}\Big[L(W_t)\Big] \leq \min_{t_1 \leq k \leq t-1}\mathbb{E}\Big[L(W_k)\Big] \leq \frac{8F(W_{t_1}) + 4\ln^2(\gamma^2\eta(t-t_1))}{7\gamma^2\eta(t-t_1)}$$

□

**Lemma C.10.** Let $W_0 = \mathbf{0}$. There exists $t_{\text{in}} \leq t_{\max}(\delta)$ such that with probability at least $1 - \delta$, we have that $L(W_{t_{\text{in}}}) \leq \tilde{L}$, where

$$t_{\max}(\delta) \geq \frac{1}{\gamma^2}\max\left\{\frac{8\left(K - 1 + 2\eta(2 + \frac{1}{b})\right)}{\eta\,\delta\,\tilde{L}},\ \frac{32}{\eta\,\delta\,\tilde{L}}\ln\left(\frac{32}{\eta\,\delta\,\tilde{L}}\right)\right\}.$$

*Proof.* From Lemma C.4 applied at $t = t_{\max}(\delta)$, we have that

$$\frac{1}{t_{\max}}\sum_{k=0}^{t_{\max}-1}\mathbb{E}[G(W_k)] \leq \frac{2\left(K - 1 + 2\ln\left(\gamma^2\eta t_{\max}\right) + 2\eta(1 + \frac{1}{b})\right)}{\eta\,\gamma^2\,t_{\max}}. \tag{75}$$

Using inequality $\ln\left(\gamma^2\eta t_{\max}\right) = \ln\left(\gamma^2 t_{\max}\right) + \ln(\eta) \leq \ln\left(\gamma^2 t_{\max}\right) + \eta$, we obtain

$$\frac{1}{t_{\max}}\sum_{k=0}^{t_{\max}-1}\mathbb{E}[G(W_k)] \leq \frac{2\left(K - 1 + 2\ln\left(\gamma^2 t_{\max}\right) + 2\eta(2 + \frac{1}{b})\right)}{\eta\,\gamma^2\,t_{\max}}. \tag{76}$$

For $\delta \in (0, 1)$, we select $t_{\max}$ so that the right-hand side of (76) is at most $\delta\,\tilde{L}/2$. We, next, verify the formula for $t_{\max}$:

- For $t \geq \frac{8\big(K-1+2\eta(2+\frac{1}{b})\big)}{\eta\,\delta\,\tilde{L}\,\gamma^2}$, it holds that $\frac{2\big(K-1+2\eta(2+\frac{1}{b})\big)}{\eta\,\gamma^2\,t} \;\leq\; \frac{\delta\,\tilde{L}}{4}$.
- For $t \geq \frac{32}{\eta\,\delta\,\tilde{L}\,\gamma^2}\,\ln\!\Big(\frac{32}{\eta\,\delta\,\tilde{L}}\Big)$, it holds that $\gamma^2\,t \;\geq\; \frac{32}{\eta\,\delta\,\tilde{L}}\,\ln\!\Big(\frac{32}{\eta\,\delta\,\tilde{L}}\Big)$, which is a sufficient condition (see Lemma G.5 in Cai et al. (2024)) for $\frac{4\,\ln(\gamma^2\,t)}{\eta\,\gamma^2\,t} \;\leq\; \frac{\delta\,\tilde{L}}{4}$.

Thus, it suffices to select $t_{\max}(\delta)$ such that

$$t_{\max}(\delta) \;\geq\; \frac{1}{\gamma^2}\,\max\left\{\frac{8\big(K-1+2\eta(2+\frac{1}{b})\big)}{\eta\,\delta\,\tilde{L}},\;\; \frac{32}{\eta\,\delta\,\tilde{L}}\,\ln\!\Big(\frac{32}{\eta\,\delta\,\tilde{L}}\Big)\right\}. \tag{77}$$

Thus, we have that

$$\frac{1}{t_{\max}}\sum_{k=0}^{t_{\max}-1}\mathbb{E}[G(W_k)] \;\leq\; \frac{\delta\,\tilde{L}}{2}. \tag{78}$$

Hence, there exists $t_{\text{in}} \leq t_{\max}(\delta)$ such that

$$\mathbb{E}[G(W_{t_{\text{in}}})] \;\leq\; \frac{\delta\,\tilde{L}}{2}. \tag{79}$$

Let the event $\mathcal{E}_{\text{in}} := \{G(W_{t_{\text{in}}}) \leq \tilde{L}/2\}$. Since $G(W_{t_{\text{in}}}) \geq 0$, using Markov's inequality, we get

$$\mathbb{P}(\mathcal{E}_{\text{in}}^c) \;=\; \mathbb{P}\big(G(W_{t_{\text{in}}}) > \tilde{L}/2\big) \;\leq\; \frac{2\,\mathbb{E}[G(W_{t_{\text{in}}})]}{\tilde{L}} \;\leq\; \delta, \tag{80}$$

and thus $\mathbb{P}(\mathcal{E}_{\text{in}}) \geq 1-\delta$. Conditioning on the event $\mathcal{E}_{\text{in}}$ and using $G(W_{t_{\text{in}}}) \leq \frac{\tilde{L}}{2} \leq \frac{1}{4n} \leq \frac{1}{2n}$, every term in the sum $G(W_{t_{\text{in}}}) = \frac{1}{n}\sum_{i=1}^{n}(1-p_i(W_{t_{\text{in}}})_{y_i})$ satisfies $1-p_i(W_{t_{\text{in}}})_{y_i} \leq 1/2$ and thus $p_i(W_{t_{\text{in}}})_{y_i} \geq 1/2$. For every $j \neq y_i$, thus, it holds that

$$z_i(W_{t_{\text{in}}})_{y_i} - z_i(W_{t_{\text{in}}})_j \;=\; \ln\frac{p_i(W_{t_{\text{in}}})_{y_i}}{p_i(W_{t_{\text{in}}})_j} \;\geq\; 0, \tag{81}$$

which gives

$$\begin{aligned} F(W_{t_{\text{in}}}) \;&=\; \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq y_i}\frac{p_i(W_{t_{\text{in}}})_j}{p_i(W_{t_{\text{in}}})_{y_i}} \\ &\leq\; \frac{1}{n}\sum_{i=1}^{n}2\big(1-p_i(W_{t_{\text{in}}})_{y_i}\big) \\ &=\; 2\,G(W_{t_{\text{in}}}) \\ &\leq\; \tilde{L}, \end{aligned}$$

where we used $p_i(W)_{y_i} \geq 1/2$ to bound $1/p_i(W)_{y_i} \leq 2$ and $\sum_{j\neq y_i} p_i(W)_j = 1-p_i(W)_{y_i}$. Since $L(W) \leq F(W)$ for every $W \in \mathbb{R}^{d\times K}$, on the event $\mathcal{E}_{\text{in}}$, we have that

$$L(W_{t_{\text{in}}}) \;\leq\; F(W_{t_{\text{in}}}) \;\leq\; \tilde{L}. \tag{82}$$

Combining (82) with $\mathbb{P}(\mathcal{E}_{\text{in}}) \geq 1-\delta$, we conclude that there exists $t_{\text{in}} \leq t_{\max}(\delta)$ such that with probability at least $1-\delta$ it holds that $L(W_{t_{\text{in}}}) \leq \tilde{L}$, and the dynamics enter the stable regime $\mathcal{S}$ at iteration $t_{\text{in}}$. □

PROOF OF THEOREM 2.3

*Proof.* The theorem is proved by combining Lemma C.9 and Lemma C.10. □

### C.6 Proofs for Stochastic Stabilization Mechanism

**Lemma C.11.** Let $\mathcal{E}_t = \left\{L(W_t) \le \tilde{L},\ L(W_{t+1}) > \tilde{L}\right\}$. If at iteration $t > 0$ the dynamics are in the stable regime, then the exit probability satisfies

$$\Pr(\mathcal{E}_t \mid \mathcal{F}_t) \ \le\ 2\exp\!\left(-\frac{bD_t}{4\eta L(W_t)^{3/2}}\right) \tag{83}$$

where $D_t = \frac{\Delta_t}{4\eta L(W_t)^{1/2} + \frac{\sqrt{2}(1+n)}{3}\sqrt{\Delta_t}}$ and

$$\Delta_t = \tilde{L} - L(W_t) + 2\eta\big(1 - 8\eta L(W_t)\big)L(W_t)^2 + \frac{\big(1 - 16\eta L(W_t)\big)^2}{8}\,L(W_t)$$

*Proof.* From inequality (61), we have that

$$L(W_{t+1}) \quad\le\quad L(W_t) - \eta\langle \nabla L(W_t), g_t\rangle + 8\eta^2 L(W_t)\|g_t\|^2. \tag{84}$$

Let $\xi_t := g_t - \nabla L(W_t)$ denote the noise at iteration $t$. Substituting $g_t = \nabla L(W_t) + \xi_t$ and expanding, we get

$$\begin{aligned} L(W_{t+1}) \quad\le\quad & L(W_t) - \eta\big(1 - 8\eta L(W_t)\big)\|\nabla L(W_t)\|^2 \\ & -\eta\big(1 - 16\eta L(W_t)\big)\langle \nabla L(W_t), \xi_t\rangle + 8\eta^2 L(W_t)\|\xi_t\|^2. \end{aligned} \tag{85}$$

On the event

$$\mathcal{E}_t = \left\{L(W_t) \le \tilde{L},\ L(W_{t+1}) > \tilde{L}\right\},$$

it holds that

$$\begin{aligned} \tilde{L} - L(W_t) + \eta\big(1 - 8\eta L(W_t)\big)\|\nabla L(W_t)\|^2 \quad<\quad & -\eta\big(1 - 16\eta L(W_t)\big)\langle \nabla L(W_t), \xi_t\rangle \\ & +8\eta^2 L(W_t)\|\xi_t\|^2. \end{aligned} \tag{86}$$

We now lower bound the right-hand side by completing the square. Let

$$A_t := \eta\big(1 - 16\eta L(W_t)\big), \qquad B_t := 8\eta^2 L(W_t). \tag{87}$$

Then, it holds that

$$\begin{aligned} & -A_t\langle \nabla L(W_t), \xi_t\rangle + B_t\|\xi_t\|^2 \\ =\ & \frac{B_t}{2}\|\xi_t\|^2 + \left(\frac{B_t}{2}\|\xi_t\|^2 - A_t\langle \nabla L(W_t), \xi_t\rangle\right) \\ =\ & \frac{B_t}{2}\|\xi_t\|^2 + \frac{B_t}{2}\left\|\xi_t - \frac{A_t}{B_t}\nabla L(W_t)\right\|^2 - \frac{A_t^2}{2B_t}\|\nabla L(W_t)\|^2 \\ \ge\ & \frac{B_t}{2}\|\xi_t\|^2 - \frac{A_t^2}{2B_t}\|\nabla L(W_t)\|^2. \end{aligned} \tag{88}$$

Substituting the values of $A_t, B_t$, we obtain

$$\begin{aligned} & -\eta\big(1 - 16\eta L(W_t)\big)\langle \nabla L(W_t), \xi_t\rangle + 8\eta^2 L(W_t)\|\xi_t\|^2 \\ \ge\ & 4\eta^2 L(W_t)\|\xi_t\|^2 - \frac{\big(1 - 16\eta L(W_t)\big)^2}{16L(W_t)}\|\nabla L(W_t)\|^2. \end{aligned} \tag{89}$$

Hence, from inequality (86), for the event $\mathcal{E}_t$ to hold, it suffices that

$$\begin{aligned} & \tilde{L} - L(W_t) + \eta\big(1 - 8\eta L(W_t)\big)\|\nabla L(W_t)\|^2 \\ <\ & 4\eta^2 L(W_t)\|\xi_t\|^2 - \frac{\big(1 - 16\eta L(W_t)\big)^2}{16L(W_t)}\|\nabla L(W_t)\|^2. \end{aligned} \tag{90}$$

Using Lemma C.2, namely

$$\|\nabla L(W_t)\|^2 \le 2L(W_t)^2, \tag{91}$$

it suffices that

$$\tilde{L} - L(W_t) + 2\eta\big(1 - 8\eta L(W_t)\big)L(W_t)^2 + \frac{\big(1 - 16\eta L(W_t)\big)^2}{8} L(W_t) < 4\eta^2 L(W_t)\|\xi_t\|^2.$$

Letting

$$\Delta_t := \tilde{L} - L(W_t) + 2\eta\big(1 - 8\eta L(W_t)\big)L(W_t)^2 + \frac{\big(1 - 16\eta L(W_t)\big)^2}{8} L(W_t), \tag{92}$$

we obtain that for the event $\mathcal{E}_t$ to hold, it suffices that

$$\Delta_t < 4\eta^2 L(W_t)\|\xi_t\|^2, \tag{93}$$

or equivalently it suffices that

$$\|\xi_t\| > \sqrt{\frac{\Delta_t}{4\eta^2 L(W_t)}}. \tag{94}$$

Therefore, it holds that

$$\Pr(\mathcal{E}_t \mid \mathcal{F}_t) \;\le\; \Pr\left(\|\xi_t\| > \sqrt{\frac{\Delta_t}{4\eta^2 L(W_t)}} \mid \mathcal{F}_t\right). \tag{95}$$

Conditionally on the filtration $\mathcal{F}_t$, the noise

$$\xi_t = \frac{1}{b}\sum_{i=1}^{b} \zeta_{t,i}, \qquad \zeta_{t,i} := \nabla \ell_i(W_t) - \nabla L(W_t), \tag{96}$$

is the average of $b$ centered i.i.d. vectors. From Lemma C.2 and Lemma C.6, each term in the sum satisfies the bound

$$\|\zeta_{t,i}\| \;\le\; \|\nabla \ell_i(W_t)\| + \|\nabla L(W_t)\| \le \sqrt{2}(1+n)L(W_t), \tag{97}$$

and the variance bound

$$\mathbb{E}\Big[\|\zeta_{t,i}\|^2 \mid \mathcal{F}_t\Big] \;\le\; 2L(W_t). \tag{98}$$

By the vector Bernstein inequality applied to $\xi_t$, we have

$$\begin{aligned}
\Pr(\mathcal{E}_t \mid \mathcal{F}_t) &\le 2\exp\left(-\frac{b\,\Delta_t/(4\eta^2 L(W_t))}{4L(W_t) + \frac{2\sqrt{2}(1+n)}{3} L(W_t)\sqrt{\Delta_t/(4\eta^2 L(W_t))}}\right) \\
&= 2\exp\left(-\frac{b}{4\eta L(W_t)^{3/2}} \cdot \frac{\Delta_t}{4\eta L(W_t)^{1/2} + \frac{\sqrt{2}(1+n)}{3}\sqrt{\Delta_t}}\right).
\end{aligned}$$

□

**Lemma C.12.** Assume that the dynamics exit the stable regime at $t_{\text{out}} > 0$ and fix $\delta \in (0,1)$. Then, with probability at least $1-\delta$ the iterates of (SGD) return inside the stable regime in at most

$$t_{\text{re}} = \left\lceil \frac{4}{\gamma^2 \eta \delta \tilde{L}} \max\left\{A,\; 4\ln\left(\frac{16}{\gamma^2 \eta \delta \tilde{L}}\right)\right\}\right\rceil$$

number of steps, where $A = 3(K-1) + 4\ln^2(\gamma^2 \eta t_{\text{out}}) + 5\eta^2\big(1 + \frac{1}{b}\big)^2$.

*Proof.* Let $U = U_1 + U_2$ with $U_1 = \alpha W_*$, $U_2 = \frac{\eta}{\gamma}\big(1 + \frac{1}{b}\big)W_*$ and $\alpha = \frac{\ln\big(\gamma^2 \eta (t - t_{\text{out}})\big)}{\gamma}$. From the update rule of (SGD), we have that

$$\begin{aligned}
\|W_{t+1} - U\|_F^2 &= \|W_t - U\|_F^2 + 2\eta\langle g_t, U - W_t\rangle_F + \eta^2\|g_t\|_F^2 \\
&= \|W_t - U\|_F^2 + 2\eta\langle g_t, U_1 - W_t\rangle_F \\
&\quad + \eta^2\left(\frac{2}{\eta}\langle g_t, U_2\rangle_F + \|g_t\|_F^2\right).
\end{aligned}$$

Taking expectation conditional on the filtration $\mathcal{F}_t$ and using the unbiasedness property of the stochastic oracles, we get

$$\begin{aligned}\mathbb{E}\left[\|W_{t+1}-U\|_F^2 \,\middle|\, \mathcal{F}_t\right] &= \|W_t-U\|_F^2 + 2\eta\langle\nabla L(W_t), U_1 - W_t\rangle_F \\ &\quad + \eta^2\left(\frac{2}{\eta}\langle\nabla L(W_t), U_2\rangle_F + \mathbb{E}\left[\|g_t\|_F^2 \,\middle|\, \mathcal{F}_t\right]\right).\end{aligned} \tag{99}$$

Using Lemmas C.1, C.2, C.6 with the selected $U_2 = \frac{\eta}{\gamma}\left(1+\frac{1}{b}\right)W_*$, we have that

$$\begin{aligned}\frac{2}{\eta}\langle\nabla L(W_t), U_2\rangle_F + \mathbb{E}\left[\|g_t\|_F^2 \,\middle|\, \mathcal{F}_t\right] &\overset{\text{Lemma C.1}}{\leq} -2\left(1+\tfrac{1}{b}\right)G(W_t) + \mathbb{E}\left[\|g_t\|_F^2 \,\middle|\, \mathcal{F}_t\right] \\ &= -2\left(1+\tfrac{1}{b}\right)G(W_t) + \|\nabla L(W_t)\|_F^2 \\ &\quad + \mathbb{E}\left[\|g_t - \nabla L(W_t)\|_F^2 \,\middle|\, \mathcal{F}_t\right] \\ &\overset{\text{Lemma C.6}}{\leq} -2\left(1+\tfrac{1}{b}\right)G(W_t) + \|\nabla L(W_t)\|_F^2 + \frac{2}{b}G(W_t) \\ &\overset{\text{Lemma C.2}}{\leq} -2\left(1+\tfrac{1}{b}\right)G(W_t) + 2\left(1+\tfrac{1}{b}\right)G(W_t) \\ &\leq 0,\end{aligned} \tag{100}$$

where we have used that $G(W_t) \in [0,1]$ and hence $G(W_t)^2 \leq G(W_t)$. Substituting (100) into (99) gives

$$\mathbb{E}\left[\|W_{t+1}-U\|_F^2 \,\middle|\, \mathcal{F}_t\right] \leq \|W_t-U\|_F^2 + 2\eta\langle\nabla L(W_t), U_1 - W_t\rangle_F.$$

Using the fact that $L$ is convex, we have that

$$\mathbb{E}\left[\|W_{t+1}-U\|_F^2 \,\middle|\, \mathcal{F}_t\right] \leq \|W_t-U\|_F^2 + 2\eta\left[L(U_1) - L(W_t)\right].$$

Taking expectation again and using the tower law of expectation, we obtain

$$\mathbb{E}\left[\|W_{t+1}-U\|_F^2\right] \leq \mathbb{E}\left[\|W_t-U\|_F^2\right] + 2\eta\,\mathbb{E}\left[L(U_1) - L(W_t)\right].$$

Summing for $k = t_{\text{out}}, \ldots, t-1$ and dividing by $2\eta(t-t_{\text{out}})$, we have that

$$\frac{\mathbb{E}\left[\|W_t-U\|_F^2\right]}{2\eta(t-t_{\text{out}})} + \frac{1}{t-t_{\text{out}}}\sum_{k=t_{\text{out}}}^{t-1}\mathbb{E}[L(W_k)] \leq L(U_1) + \frac{\mathbb{E}\left[\|W_{t_{\text{out}}}-U\|_F^2\right]}{2\eta(t-t_{\text{out}})}.$$

Using the non-negativity of $\|W_t-U\|_F^2$, we get

$$\frac{1}{t-t_{\text{out}}}\sum_{k=t_{\text{out}}}^{t-1}\mathbb{E}[L(W_k)] \leq L(U_1) + \frac{\mathbb{E}\left[\|W_{t_{\text{out}}}-U\|_F^2\right]}{2\eta(t-t_{\text{out}})}. \tag{101}$$

For $U_1 = \alpha W_*$ with $\alpha = \frac{\ln\left(\gamma^2\eta(t-t_{\text{out}})\right)}{\gamma}$, we have that

$$L(U_1) \leq F(U_1) = (K-1)\,e^{-\alpha\gamma} = \frac{K-1}{\gamma^2\eta(t-t_{\text{out}})}. \tag{102}$$

We, next, bound the term $\mathbb{E}\left[\|W_{t_{\text{out}}}-U\|_F^2\right]$. Applying $(a+b)^2 \leq 2a^2+2b^2$, we obtain

$$\mathbb{E}\left[\|W_{t_{\text{out}}}-U\|_F^2\right] \leq 2\,\mathbb{E}\left[\|W_{t_{\text{out}}}\|_F^2\right] + 2\|U\|_F^2. \tag{103}$$

For the term $\|U\|_F^2$, we have that

$$\|U\|_F^2 \leq 2\|U_1\|_F^2 + 2\|U_2\|_F^2 \leq \frac{2\ln^2(\gamma^2\eta(t-t_{\text{out}}))}{\gamma^2} + \frac{2\eta^2\left(1+\frac{1}{b}\right)^2}{\gamma^2}. \tag{104}$$

For the term $\mathbb{E}\left[\|W_{t_{\text{out}}}\|_F^2\right]$, we apply Lemma C.7 from the original time origin with comparator $U^{(0)} = U_1^{(0)} + U_2$ where $U_1^{(0)} = \alpha_0 W_*$ and $\alpha_0 = \frac{\ln\left(\gamma^2\eta t_{\text{out}}\right)}{\gamma}$, which gives

$$\mathbb{E}\left[\|W_{t_{\text{out}}}-U^{(0)}\|_F^2\right] \leq 2\eta\,t_{\text{out}}\,L(U_1^{(0)}) + \|U^{(0)}\|_F^2 \leq \frac{2(K-1)}{\gamma^2} + \frac{2\ln^2(\gamma^2\eta t_{\text{out}}) + 2\eta^2\left(1+\frac{1}{b}\right)^2}{\gamma^2}, \tag{105}$$

where we have used $W_0 = \mathbf{0}$ and the bound (102) applied at time $t_{\text{out}}$. Applying $(a+b)^2 \le 2a^2+2b^2$ again, we obtain

$$\mathbb{E}\big[\|W_{t_{\text{out}}}\|_F^2\big] \le 2\,\mathbb{E}\Big[\|W_{t_{\text{out}}} - U^{(0)}\|_F^2\Big] + 2\|U^{(0)}\|_F^2 \le \frac{4(K-1) + 8\ln^2(\gamma^2\eta t_{\text{out}}) + 8\eta^2\big(1+\frac{1}{b}\big)^2}{\gamma^2}. \tag{106}$$

Substituting (102), (104), and (106) into (101), we obtain

$$\frac{1}{t - t_{\text{out}}} \sum_{k=t_{\text{out}}}^{t-1} \mathbb{E}[L(W_k)] \le \frac{C(t, t_{\text{out}})}{\gamma^2\eta(t - t_{\text{out}})}, \tag{107}$$

where

$$C(t, t_{\text{out}}) := 3(K-1) + 4\ln^2(\gamma^2\eta t_{\text{out}}) + 5\eta^2\big(1+\tfrac{1}{b}\big)^2 + \ln^2(\gamma^2\eta(t - t_{\text{out}})).$$

We now split the bound (107) into two terms. Let $A := 3(K-1) + 4\ln^2(\gamma^2\eta t_{\text{out}}) + 5\eta^2\big(1+\frac{1}{b}\big)^2$. Then $C(t, t_{\text{out}}) = A + \ln^2(\gamma^2\eta(t - t_{\text{out}}))$, and (107) becomes

$$\frac{1}{t - t_{\text{out}}} \sum_{k=t_{\text{out}}}^{t-1} \mathbb{E}[L(W_k)] \le \frac{A}{\gamma^2\eta(t - t_{\text{out}})} + \frac{\ln^2(\gamma^2\eta(t - t_{\text{out}}))}{\gamma^2\eta(t - t_{\text{out}})}. \tag{108}$$

We select $t_{\text{re}}$ so that each of the two terms on the right-hand side of (108) is at most $\delta\tilde{L}/4$. Specifically, we next verify the formula for $t_{\text{re}}$:

- For $t - t_{\text{out}} \ge \frac{4A}{\gamma^2\eta\delta\tilde{L}}$, it holds that
$$\frac{A}{\gamma^2\eta(t - t_{\text{out}})} \le \frac{\delta\tilde{L}}{4}.$$
- For $t - t_{\text{out}} \ge \frac{16}{\gamma^2\eta\delta\tilde{L}} \ln\Big(\frac{16}{\gamma^2\eta\delta\tilde{L}}\Big)$, it holds that $\gamma^2\eta(t - t_{\text{out}}) \ge \frac{16}{\delta\tilde{L}} \ln\Big(\frac{16}{\gamma^2\eta\delta\tilde{L}}\Big)$, which is a sufficient condition (see Lemma G.5 in Cai et al. (2024)) for
$$\frac{\ln^2(\gamma^2\eta(t - t_{\text{out}}))}{\gamma^2\eta(t - t_{\text{out}})} \le \frac{\delta\tilde{L}}{4}.$$

Thus, it suffices to select $t_{\text{re}}$ such that

$$t_{\text{re}} = \left\lceil \frac{4}{\gamma^2\eta\delta\tilde{L}} \max\left\{A,\ 4\ln\Big(\frac{16}{\gamma^2\eta\delta\tilde{L}}\Big)\right\} \right\rceil. \tag{109}$$

For this choice of $t_{\text{re}}$ and selecting $t = t_{\text{out}} + t_{\text{re}}$, we have that

$$\frac{1}{t_{\text{re}}} \sum_{k=t_{\text{out}}}^{t_{\text{out}}+t_{\text{re}}-1} \mathbb{E}[L(W_k)] \le \frac{\delta\tilde{L}}{2}. \tag{110}$$

Hence, there exists $t_{\text{in}}^{[2]} \in [t_{\text{out}}, t_{\text{out}} + t_{\text{re}}]$ such that

$$\mathbb{E}\Big[L(W_{t_{\text{in}}^{[2]}})\Big] \le \frac{\delta\tilde{L}}{2}.$$

From Markov's inequality, we have that

$$\Pr\Big(L(W_{t_{\text{in}}^{[2]}}) \ge \tilde{L}\Big) \le \frac{\mathbb{E}\Big[L(W_{t_{\text{in}}^{[2]}})\Big]}{\tilde{L}} \le \frac{\delta}{2} \le \delta,$$

which implies that

$$\Pr\Big(L(W_{t_{\text{in}}^{[2]}}) \le \tilde{L}\Big) \ge 1 - \delta.$$

Thus, there exists $t_{\text{in}}^{[2]} \le t_{\text{out}} + t_{\text{re}}$ such that for $\delta \in (0,1)$ with probability at least $1-\delta$ it holds that $L(W_{t_{\text{in}}^{[2]}}) \le \tilde{L}$ and the dynamics enter the stable set. □

### Proof of Theorem 2.4

*Proof.* The theorem is proved by combining Lemma C.11 and Lemma C.12. □

# D PROOFS FOR TWO-LAYER NEURAL NETWORKS

In this appendix, we provide the analysis of SGD applied to the two-layer neural network (6) under the multi-class cross-entropy loss with any step size $\tilde{\eta} = m\eta > 0$. The proof of convergence relies on a sequence of lemmas that establish key properties of the dynamics and the geometry of the loss landscape.

## D.1 NOTATION

For each sample $(x_i, y_i)$, let

$$z_i(W) \;=\; \frac{1}{m}\sum_{r=1}^{m} a_r\,\phi(x_i^\top W^{(r)}) \in \mathbb{R}^K, \qquad p_i(W) \;=\; \mathrm{softmax}\big(z_i(W)\big).$$

The gradient and Hessian of the cross-entropy loss evaluated at the sample $(x_i, y_i)$ are given by

$$\nabla \ell_i(W) \;=\; J_i(W)^\top \big(p_i(W) - e_{y_i}\big), \tag{111}$$

$$\begin{aligned}\nabla^2 \ell_i(W) \;&=\; J_i(W)^\top \big(\mathrm{diag}(p_i(W)) - p_i(W)p_i(W)^\top\big) J_i(W) \\ &\quad + \sum_{k=1}^{K} \big(p_i(W)_k - \mathbb{1}_{\{k=y_i\}}\big)\,\nabla^2 z_i(W)_k,\end{aligned} \tag{112}$$

where $J_i(W)$ is the Jacobian of $W \mapsto z_i(W)$. We define the lifted comparator

$$\overline{W}_* \;=\; \big(a_1 W_*, \dots, a_m W_*\big) \in \mathbb{R}^{md\times K}$$

with $\|\overline{W}_*\|_F \;=\; \sqrt{m}$ and let $c_\phi := \frac{1}{m^2}\sum_{r=1}^m a_r^2 = \frac{1}{m}$, since $a_r \in \{\pm 1\}$. We define the potential functions $F, G : \mathbb{R}^{md\times K} \to \mathbb{R}_{\geq 0}$ as follows

$$F(W) \;=\; \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq y_i} \exp\big(-(z_i(W)_{y_i} - z_i(W)_j)\big), \tag{113}$$

$$G(W) \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(1 - p_i(W)_{y_i}\big). \tag{114}$$

Since $\log(1+u) \leq u$ for $u \in (0,1]$, it follows that

$$G(W) \;\leq\; L(W) \;\leq\; F(W). \tag{115}$$

## D.2 PREPARATORY LEMMAS

The proof of Theorems 3.3, 3.4, 3.5 relies on a sequence of lemmas that establish key properties of the dynamics and the geometry of the loss landscape. We first introduce the aforementioned lemmas and then present the main theoretical result.

**Lemma D.1** (Jacobian of the logits)**.** For every sample $i \in [n]$, class $k \in [K]$, and hidden unit $r \in [m]$, it holds that

$$\nabla_{W^{(r)}} z_i(W)_k \;=\; \frac{a_r}{m}\,\phi'\big(x_i^\top W_k^{(r)}\big)\,x_i e_k^\top.$$

Moreover, the Jacobian $J_i(W)$ of the map $W \mapsto z_i(W)$ satisfies

$$\|J_i(W)\|_{\mathrm{op}} \;\leq\; \frac{1}{\sqrt{m}}.$$

*Proof.* Differentiating $z_i(W)_k = \frac{1}{m}\sum_{s=1}^m a_s\,\phi(x_i^\top W_k^{(s)})$ with respect to the block $W^{(r)}$ gives the stated expression. For $H = (H^{(1)}, \dots, H^{(m)}) \in \mathbb{R}^{md\times K}$, we have that for every $k \in [K]$ it holds

$$\big(J_i(W)H\big)_k \;=\; \frac{1}{m}\sum_{r=1}^{m} a_r\,\phi'(x_i^\top W_k^{(r)})\,x_i^\top H_k^{(r)}.$$

Using $|a_r| = 1$, $|\phi'| \le 1$, $\|x_i\|_2 \le 1$, and Cauchy–Schwarz across $r$, we have that

$$\left|\big(J_i(W)H\big)_k\right| \;\le\; \frac{1}{m}\sum_{r=1}^{m}\|H_k^{(r)}\|_2 \;\le\; \frac{1}{\sqrt{m}}\Big(\sum_{r=1}^{m}\|H_k^{(r)}\|_2^2\Big)^{1/2}.$$

Taking the square and summing over $k = 1, ..., K$, we obtain

$$\|J_i(W)H\|_2^2 \;\le\; \frac{1}{m}\sum_{k=1}^{K}\sum_{r=1}^{m}\|H_k^{(r)}\|_2^2 \;=\; \frac{1}{m}\,\|H\|_F^2.$$

Taking the supremum over $H$ satiafying $\|H\|_F = 1$, we have

$$\begin{aligned}\|J_i(W)\|_{\mathrm{op}}^2 \;&\le\; \frac{1}{m}\\ \Rightarrow \|J_i(W)\|_{\mathrm{op}} \;&\le\; \frac{1}{\sqrt{m}}.\end{aligned} \tag{116}$$

□

**Lemma D.2** (Two-layer perceptron-type inequality)**.** Let Assumptions 2.1, 3.1 hold. Then, for every $W \in \mathbb{R}^{md\times K}$ it holds that

$$\langle \nabla L(W), \overline{W}_*\rangle_F \;\le\; -\tilde{\gamma}G(W), \tag{117}$$

where $\tilde{\gamma} := \alpha\gamma - (1-\alpha)$.

*Proof.* By the chain rule and (111), for each block $r \in [m]$, we have that

$$\nabla_{W^{(r)}} L(W) \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{a_r}{m}\,x_i\big[(p_i(W) - e_{y_i}) \odot \phi'(x_i^\top W^{(r)})\big]^\top, \tag{118}$$

where $\odot$ denotes component-wise multiplication and $\phi'(x_i^\top W^{(r)}) \in \mathbb{R}^K$ is the vector with $k$-th component $\phi'(x_i^\top W_k^{(r)})$. Since $\overline{W}_* = (a_1 W_*, \ldots, a_m W_*)$ and $a_r^2 = 1$, we have

$$\begin{aligned}\langle \nabla L(W), \overline{W}_*\rangle_F \;&=\; \sum_{r=1}^{m}\big\langle \nabla_{W^{(r)}} L(W),\, a_r W_*\big\rangle_F\\ &=\; \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{m}\frac{a_r^2}{m}\Big\langle (p_i(W) - e_{y_i}) \odot \phi'(x_i^\top W^{(r)}),\, W_*^\top x_i\Big\rangle\\ &=\; \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{m}\frac{a_r^2}{m}\Big\langle (p_i(W) - e_{y_i}) \odot \phi'(x_i^\top W^{(r)}),\, \Delta_i\Big\rangle.\end{aligned} \tag{119}$$

where we have defined $\Delta_i := W_*^\top x_i \in \mathbb{R}^K$. Expanding the inner product over classes and using $p_i(W)_{y_i} - 1 = -\sum_{j\neq y_i} p_i(W)_j$, we have that

$$\langle \nabla L(W), \overline{W}_*\rangle_F \;=\; \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{m}\sum_{j\neq y_i}\frac{a_r^2}{m}\,p_i(W)_j\,\big[\phi'(x_i^\top W_j^{(r)})\,\Delta_i(j) - \phi'(x_i^\top W_{y_i}^{(r)})\,\Delta_i(y_i)\big]. \tag{120}$$

We, next, bound the term $A := \phi'(x_i^\top W_j^{(r)})\,\Delta_i(j) - \phi'(x_i^\top W_{y_i}^{(r)})\,\Delta_i(y_i)$ appearing on the right-hand side of (120). We have

$$\begin{aligned}A \quad &= \quad \phi'(x_i^\top W_{y_i}^{(r)})\big(\Delta_i(j) - \Delta_i(y_i)\big) + \big(\phi'(x_i^\top W_j^{(r)}) - \phi'(x_i^\top W_{y_i}^{(r)})\big)\,\Delta_i(j)\\ &\overset{\text{Assumption 2.1}}{\le} \quad -\gamma\phi'(x_i^\top W_{y_i}^{(r)}) + \big(\phi'(x_i^\top W_j^{(r)}) - \phi'(x_i^\top W_{y_i}^{(r)})\big)\,\Delta_i(j)\\ &\overset{\alpha\le\phi'(\cdot)}{\le} \quad -\alpha\gamma + \big(\phi'(x_i^\top W_j^{(r)}) - \phi'(x_i^\top W_{y_i}^{(r)})\big)\,\Delta_i(j).\end{aligned} \tag{121}$$

Since $\phi'(x_i^\top W_j^{(r)})$, $\phi'(x_i^\top W_{y_i}^{(r)}) \in (\alpha, 1)$, we have $\left|\phi'(x_i^\top W_j^{(r)}) - \phi'(x_i^\top W_{y_i}^{(r)})\right| \leq 1-\alpha$ and thus substituting into (121) we get

$$A \;\leq\; -\alpha\gamma + (1-\alpha)\,\Delta_i(j)$$

Applying Cauchy–Schwarz inequality and using that $\|W_k^*\|_2 \leq \|W^*\|_F = 1, \forall k \in [K]$ from Assumption 2.1, we have that

$$|\Delta_i(j)| \;=\; |(W_*)_j^\top x_i| \;\leq\; \|(W_*)_j\|_2\,\|x_i\|_2 \;\leq\; \|x_i\|_2 \overset{\|x_i\|_2\leq 1}{\leq} 1, \tag{122}$$

and thus we get

$$A \;\leq\; -\alpha\gamma + (1-\alpha) = -\tilde{\gamma}, \tag{123}$$

where $\tilde{\gamma} := \alpha\gamma - (1-\alpha)$. Substituting inequality (123) into (120), we obtain

$$\begin{aligned}
\langle \nabla L(W), \overline{W}_* \rangle_F \quad &\leq \quad -\tilde{\gamma}\frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{m}\sum_{j\neq y_i}\frac{a_r^2}{m}\,p_i(W)_j \\
&= \quad -\tilde{\gamma}\frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{m}\frac{a_r^2}{m}\left(1-p_i(W)_{y_i}\right) \\
&\overset{a_r^2=1}{=} \quad -\tilde{\gamma}\frac{1}{n}\sum_{i=1}^{n}\left(1-p_i(W)_{y_i}\right) \\
&= \quad -\tilde{\gamma}G(W).
\end{aligned}$$

□

**Lemma D.3** (Two-layer gradient norm bounds)**.** For every $W \in \mathbb{R}^{md\times K}$, it holds for every $i \in [n]$ that

$$\|\nabla \ell_i(W)\|_F \;\leq\; \sqrt{\frac{2}{m}}\left(1-p_i(W)_{y_i}\right) \;\leq\; \sqrt{\frac{2}{m}}\,\ell_i(W) \;\leq\; \sqrt{\frac{2}{m}}\,n\,L(W), \tag{124}$$

$$\|\nabla L(W)\|_F \;\leq\; \sqrt{\frac{2}{m}}\,G(W) \;\leq\; \sqrt{\frac{2}{m}}\,L(W). \tag{125}$$

Moreover, it holds that

$$\sqrt{m}\,\|\nabla L(W)\|_F \;\geq\; \tilde{\gamma}\,G(W). \tag{126}$$

*Proof.* By Lemma D.1, we have that

$$\|\nabla \ell_i(W)\|_F \;=\; \|J_i(W)^\top(p_i(W)-e_{y_i})\|_F \;\leq\; \|J_i(W)\|_{\mathrm{op}}\,\|p_i(W)-e_{y_i}\|_2 \;\leq\; \frac{1}{\sqrt{m}}\,\|p_i(W)-e_{y_i}\|_2.$$

It holds that

$$\begin{aligned}
\|p_i(W)-e_{y_i}\|_2^2 \;&=\; \left(1-p_i(W)_{y_i}\right)^2 + \sum_{j\neq y_i} p_i(W)_j^2 \\
&\leq\; \left(1-p_i(W)_{y_i}\right)^2 + \Big(\sum_{j\neq y_i} p_i(W)_j\Big)^2 \\
&=\; 2\left(1-p_i(W)_{y_i}\right)^2,
\end{aligned}$$

where we used $\sum a_j^2 \leq (\sum a_j)^2$ for non-negative $a_j$ and $\sum_{j\neq y_i} p_i(W)_j = 1 - p_i(W)_{y_i}$. Therefore, $\|p_i(W)-e_{y_i}\|_2 \leq \sqrt{2}\,(1-p_i(W)_{y_i})$, and combining with $1-p_i(W)_{y_i} \leq \ell_i(W) \leq nL(W)$ proves the inequality (124). By the triangle inequality, we have that

$$\|\nabla L(W)\|_F \;\leq\; \frac{1}{n}\sum_{i=1}^{n}\|\nabla \ell_i(W)\|_F \;\leq\; \sqrt{\frac{2}{m}}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(1-p_i(W)_{y_i}\right) \;=\; \sqrt{\frac{2}{m}}\,G(W) \leq \sqrt{\frac{2}{m}}L(W),$$

since it holds that $G(W) \leq L(W)$. Applying the Cauchy–Schwarz inequality and using the fact that $\|\overline{W}_*\|_F = \sqrt{m}$, we have that

$$\langle -\nabla L(W), \overline{W}_* \rangle_F \leq \sqrt{m}\, \|\nabla L(W)\|_F \tag{127}$$

Using Lemma D.2, we obtain

$$\sqrt{m}\, \|\nabla L(W)\|_F \;\geq\; \tilde{\gamma}\, G(W)$$

□

**Lemma D.4** (Two-layer Hessian bound)**.** Under Assumption 3.1, for any $W \in \mathbb{R}^{md \times K}$ and $i \in [n]$ it holds that

$$\|\nabla^2 \ell_i(W)\|_{\text{op}} \;\leq\; \frac{2(1+\tilde{\beta})}{m}\, \ell_i(W), \qquad \|\nabla^2 L(W)\|_{\text{op}} \;\leq\; \frac{2(1+\tilde{\beta})}{m}\, L(W). \tag{128}$$

*Proof.* From the decomposition (112), we have that

$$\begin{aligned} \nabla^2 \ell_i(W) \;&=\; J_i(W)^\top \big(\text{diag}(p_i(W)) - p_i(W) p_i(W)^\top\big) J_i(W) \\ &+\; \sum_{k=1}^{K} \big(p_i(W)_k - \mathbb{1}_{\{k=y_i\}}\big)\, \nabla^2 z_i(W)_k \end{aligned} \tag{129}$$

By the sub-additivity of the operator norm, we get

$$\begin{aligned} \big\|\nabla^2 \ell_i(W)\big\|_{\text{op}} \;&\leq\; \big\|\|J_i(W)\|_{\text{op}}^2\, \|\text{diag}(p_i(W)) - p_i(W) p_i(W)^\top\big\|_{\text{op}} \\ &+ \left\| \sum_{k=1}^{K} \big(p_i(W)_k - \mathbb{1}_{\{k=y_i\}}\big)\, \nabla^2 z_i(W)_k \right\|_{\text{op}} \end{aligned} \tag{130}$$

We, next, bound the two terms on the right-hand side of (130). The operator norm of the first term in (130) is bounded by

$$\begin{aligned} \big\|\|J_i(W)\|_{\text{op}}^2\, \|\text{diag}(p_i(W)) - p_i(W) p_i(W)^\top\big\|_{\text{op}} \;&\leq\; \frac{1}{m}\big(1 - \|p_i(W)\|_2^2\big) \\ &\leq\; \frac{2}{m}\big(1 - p_i(W)_{y_i}\big) \\ &\leq\; \frac{2}{m}\, \ell_i(W) \end{aligned} \tag{131}$$

where we used Lemma D.1, $\|\text{diag}(p) - pp^\top\|_{\text{op}} \leq 1 - \|p\|_2^2 \leq 2(1 - p(y_i))$, and $1 - p_i(W)_{y_i} \leq \ell_i(W)$. For the second term in (130), it holds that $\|\nabla^2 z_i(W)_k\|_{\text{op}} \leq \frac{\tilde{\beta}}{m}$, since $\|x_i\|_2 \leq 1$ and $|\phi''(\cdot)| \leq \tilde{\beta}$. We, also, have that

$$\sum_{k=1}^{K} \big|p_i(W)_k - \mathbb{1}_{\{k=y_i\}}\big| \;=\; \big(1 - p_i(W)_{y_i}\big) + \sum_{j \neq y_i} p_i(W)_j \;=\; 2\big(1 - p_i(W)_{y_i}\big) \;\leq\; 2\, \ell_i(W).$$

Thus, we get that

$$\begin{aligned} \left\| \sum_{k=1}^{K} \big(p_i(W)_k - \mathbb{1}_{\{k=y_i\}}\big)\, \nabla^2 z_i(W)_k \right\|_{\text{op}} \;&\leq\; \left\| \sum_{k=1}^{K} \big(p_i(W)_k - \mathbb{1}_{\{k=y_i\}}\big) \right\|_{\text{op}} \big\|\nabla^2 z_i(W)_k\big\|_{\text{op}} \\ &\leq\; \frac{2\tilde{\beta}}{m}\, \ell_i(W). \end{aligned} \tag{132}$$

Substituting inequalities (131), (132) into (130), we get

$$\|\nabla^2 \ell_i(W)\|_{\text{op}} \;\leq\; \frac{2(1+\tilde{\beta})}{m}\, \ell_i(W)$$

Summing over $i = 1, \ldots, n$ and multiplying by $\frac{1}{n}$, we have that

$$\|\nabla^2 L(W)\|_{\text{op}} \;\leq\; \frac{2(1+\tilde{\beta})}{m}\, L(W).$$

□

**Lemma D.5.** The iterates of SGD for the two-layer network (6) with any step size $\tilde{\eta} = m\eta > 0$ and batch size $b \geq 1$ satisfy

$$\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}[G(W_k)] \;\leq\; \frac{2\big(K-1+2\ln\big(\tilde{\gamma}^2\eta t\big)+4\kappa+2\eta(1+\frac{1}{b})\big)}{\eta\tilde{\gamma}^2 t}$$

where $G(W) = \frac{1}{n}\sum_{i=1}^{n}(1-p_i(W)_{y_i})$.

*Proof.* From Lemma D.2, we have that the gradient of the cross-entropy loss satisfies $\langle \nabla L(W_t), \overline{W}_*\rangle_F \leq -\tilde{\gamma}\, G(W_t)$. Using the SGD update $W_{t+1} = W_t - \tilde{\eta} g_t$, we get

$$\langle W_{t+1}, \overline{W}_*\rangle_F \;=\; \langle W_t, \overline{W}_*\rangle_F - \tilde{\eta}\,\langle g_t, \overline{W}_*\rangle_F.$$

Taking expectation conditional on the filtration $\mathcal{F}_t$ and using $\mathbb{E}[g_t \mid \mathcal{F}_t] = \nabla L(W_t)$, we have that

$$\mathbb{E}\big[\langle W_{t+1}, \overline{W}_*\rangle_F \mid \mathcal{F}_t\big] \;\geq\; \langle W_t, \overline{W}_*\rangle_F + \tilde{\eta}\tilde{\gamma}\, G(W_t).$$

Taking expectation again and using the tower law of expectation, it holds that

$$\mathbb{E}\big[\langle W_{t+1}, \overline{W}_*\rangle_F\big] \;\geq\; \mathbb{E}\Big[\langle W_t, \overline{W}_*\rangle_F\Big] + \tilde{\eta}\tilde{\gamma}\;\mathbb{E}\Big[G(W_t)\Big]. \tag{133}$$

Unrolling the recursion and multiplying by $\frac{1}{t}$, we obtain

$$\frac{1}{t}\,\mathbb{E}\big[\langle W_t, \overline{W}_*\rangle_F\big] \;\geq\; \frac{\tilde{\eta}\,\tilde{\gamma}}{t}\sum_{k=0}^{t-1}\mathbb{E}[G(W_k)],$$

where we have used the fact that $W_0 = \mathbf{0}$. Since $\|\overline{W}_*\|_F = \sqrt{m}$, by Cauchy–Schwarz it holds that $\sqrt{m}\,\|W_t\|_F \geq \langle W_t, \overline{W}_*\rangle_F$, and thus we obtain

$$\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}[G(W_k)] \;\leq\; \frac{\sqrt{m}\,\mathbb{E}[\|W_t\|_F]}{\tilde{\eta}\,\tilde{\gamma}\,t}. \tag{134}$$

We, next, bound the term $\mathbb{E}[\|W_t\|_F]$. Applying the triangle inequality, we have that

$$\|W_t\|_F \;\leq\; \|W_t - U\|_F + \|U\|_F. \tag{135}$$

Thus, it suffices to bound the terms $\|W_t - U\|_F$ and $\|U\|_F$. From Proposition D.10, for the decomposition $U = U_1 + U_2$ with

$$U_1 \;=\; \frac{\ln\big(\tilde{\gamma}^2\eta t\big)+2\kappa}{\tilde{\gamma}}\,\overline{W}_*, \qquad U_2 \;=\; \frac{\eta(1+\frac{1}{b})}{\tilde{\gamma}}\,\overline{W}_*,$$

it holds that

$$\frac{\mathbb{E}\big[\|W_t - U\|_F^2\big]}{2\tilde{\eta}\,t} + \frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}[L(W_k)] \;\leq\; L(U_1) + \frac{\|U\|_F^2}{2\tilde{\eta}\,t},$$

Since $L(\cdot) \geq 0$, it holds that

$$\mathbb{E}\big[\|W_t - U\|_F^2\big] \;\leq\; 2\tilde{\eta}\,t\,L(U_1) + \|U\|_F^2.$$

From Lemma D.6, we have that

$$L(U_1) \leq \frac{K-1}{\tilde{\gamma}^2\eta t} \tag{136}$$

It holds that

$$\|U\|_F^2 \;\leq\; 2\|U_1\|_F^2 + 2\|U_2\|_F^2 \;=\; \frac{2m\big(\ln\big(\tilde{\gamma}^2\eta t\big)+2\kappa\big)^2}{\tilde{\gamma}^2} + \frac{2m\eta^2(1+\frac{1}{b})^2}{\tilde{\gamma}^2}.$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$ for the first term, we obtain

$$\|U\|_F^2 \;\leq\; \frac{4m\,\ln^2(\tilde{\gamma}^2\eta t) + 16m\,\kappa^2 + 2m\eta^2(1+\frac{1}{b})^2}{\tilde{\gamma}^2}. \tag{137}$$

Applying the inequality $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, we obtain

$$\begin{aligned}\mathbb{E}\big[\|W_t\|_F^2\big] &\le 2\,\mathbb{E}\big[\|W_t - U\|_F^2\big] + 2\,\|U\|_F^2 \\ &\le 4\tilde{\eta}\,t\,L(U_1) + 4\|U\|_F^2 \\ &\le \frac{4m(K-1)}{\tilde{\gamma}^2} + \frac{16m\,\ln^2(\tilde{\gamma}^2\eta t) + 64m\,\kappa^2 + 8m\eta^2(1+\frac{1}{b})^2}{\tilde{\gamma}^2} \\ &\le \frac{4m(K-1) + 16m\,\ln^2(\tilde{\gamma}^2\eta t) + 64m\,\kappa^2 + 8m\eta^2(1+\frac{1}{b})^2}{\tilde{\gamma}^2}.\end{aligned}$$

Using Jensen's inequality and $\sqrt{a+b+c+d} \le \sqrt{a} + \sqrt{b} + \sqrt{c} + \sqrt{d}$ for $a, b, c, d \ge 0$, we have that

$$\begin{aligned}\mathbb{E}\big[\|W_t\|_F\big] &\le \sqrt{\mathbb{E}\Big[\|W_t\|_F^2\Big]} \\ &\le \frac{2\sqrt{m}\big(\sqrt{K-1} + 2\ln\big(\tilde{\gamma}^2\eta t\big) + 4\kappa + 2\eta(1+\frac{1}{b})\big)}{\tilde{\gamma}}. \qquad (138)\end{aligned}$$

Substituting (138) into (134) and using the fact that $\tilde{\eta} = m\eta$ and $\sqrt{K-1} \le K-1$ (for $K \ge 2$), we obtain

$$\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}[G(W_k)] \le \frac{2\big(K-1+2\ln\big(\tilde{\gamma}^2\eta t\big) + 4\kappa + 2\eta(1+\frac{1}{b})\big)}{\eta\tilde{\gamma}^2 t}$$

□

**Lemma D.6.** Let Assumptions 2.1, 3.1 hold. For $U_1 = \frac{\ln\big(\tilde{\gamma}^2\eta t\big)+2\kappa}{\tilde{\gamma}}\,\overline{W}_*$ and any $t \ge 1$, we have that

$$L(U_1) \le F(U_1) \le \frac{K-1}{\tilde{\gamma}^2\eta t}. \qquad (139)$$

*Proof.* Fix $i \in [n]$ and $j \ne y_i$. For every hidden unit $r \in [m]$ it holds that $U_1^{(r)} = a_r \frac{\ln\big(\tilde{\gamma}^2\eta t\big)+2\kappa}{\tilde{\gamma}} W^*$. Letting $D = \frac{\ln\big(\tilde{\gamma}^2\eta t\big)+2\kappa}{\tilde{\gamma}}$ for brevity, we have

$$(x_i^\top U_1^{(r)})_{y_i} = D\,a_r\,x_i^\top W_{y_i}^*, \qquad (140)$$
$$(x_i^\top U_1^{(r)})_{j} = D\,a_r\,x_i^\top W_{j}^*. \qquad (141)$$

Thus, we have that

$$z_i(U)_{y_i} - z_i(U)_j = \frac{1}{m}\sum_{r=1}^m a_r\Big[\phi\big(D\,a_r\,x_i^\top W_{y_i}^*\big) - \phi\big(D\,a_r\,x_i^\top W_j^*\big)\Big]. \qquad (142)$$

We, now, use the near-homogeneity condition

$$|\phi(z) - \phi'(z)z| \le \kappa.$$

For any $u \in \mathbb{R}$, it holds that

$$\phi(u) \ge \phi'(u)\,u - \kappa, \qquad (143)$$
$$\phi(u) \le \phi'(u)\,u + \kappa. \qquad (144)$$

Applying (143), (144) to the two terms on the right-hand side of (142), we get

$$\begin{aligned}z_i(U)_{y_i} - z_i(U)_j \ge \frac{1}{m}\sum_{r=1}^m a_r\Big[&\phi'\big(D\,a_r\,x_i^\top W_{y_i}^*\big)\big(D\,a_r\,x_i^\top W_{y_i}^*\big) \\ &-\phi'\big(D\,a_r\,x_i^\top W_j^*\big)\big(D\,a_r\,x_i^\top W_j^*\big) - 2\kappa\Big]. \qquad (145)\end{aligned}$$

Using the fact that $\phi'(\cdot) \geq \alpha$ from Assumption 3.1, it follows that

$$\begin{aligned} z_i(U)_{y_i} - z_i(U)_j &\geq \frac{\alpha}{m}\sum_{r=1}^{m} a_r^2 D\left(x_i^\top W_{y_i}^* - x_i^\top W_j^*\right) - \frac{2}{m}\sum_{r=1}^{m} a_r \kappa \\ &\geq \frac{\alpha D}{m}\sum_{r=1}^{m} a_r^2\big(x_i^\top W_{y_i}^* - x_i^\top W_j^*\big) - 2\kappa, \end{aligned} \tag{146}$$

where we have used the fact that $a_r \in \{-1, +1\}$. Since $a_r^2 = 1$ and by separability

$$x_i^\top W_{y_i}^* - x_i^\top W_j^* \geq \gamma,$$

we obtain

$$z_i(U)_{y_i} - z_i(U)_j \geq \alpha\gamma D - 2\kappa \tag{147}$$

Thus, we get

$$\exp\Big(-\big(z_i(U)_{y_i} - z_i(U)_j\big)\Big) \leq \exp\big(-\alpha\gamma D + 2\kappa\big). \tag{148}$$

Summing over all $i \in [n]$ and $j \neq y_i$, it follows that

$$F(U) \leq (K-1)\exp\big(-\alpha\gamma D + 2\kappa\big). \tag{149}$$

Using the definition of $D = \frac{\ln\left(\tilde{\gamma}^2\eta T\right)+2\kappa}{\tilde{\gamma}} \geq \frac{\ln\left(\tilde{\gamma}^2\eta T\right)+2\kappa}{\alpha\gamma}$, we have that

$$F(U) \leq \frac{K-1}{\tilde{\gamma}^2\,\eta t}. \tag{150}$$

Since $L(U) \leq F(U)$, we obtain that

$$L(U) \leq F(U) \leq \frac{K-1}{\tilde{\gamma}^2\,\eta t}. \tag{151}$$

□

**Lemma D.7.** Let Assumptions 2.1 and 3.1 hold, and let $U = W_{t_1} + U_1$ with $U_1 = \frac{\ln\left(\tilde{\gamma}^2\,\eta(t-t_1)\right)+2\kappa}{\tilde{\gamma}}\,\overline{W}_*$. Then it holds that

$$L(U) \leq F(U) \leq \frac{F(W_{t_1})}{\tilde{\gamma}^2\,\eta\,(t-t_1)}. \tag{152}$$

*Proof.* Let $C := \frac{\ln\left(\tilde{\gamma}^2\,\eta(t-t_1)\right)+2\kappa}{\tilde{\gamma}}$, so that $U_1 = C\,\overline{W}_*$ and $C\,\tilde{\gamma} = \ln\left(\tilde{\gamma}^2\,\eta(t-t_1)\right) + 2\kappa$. Since $\overline{W}_* = (a_1 W_*, \ldots, a_m W_*)$, the $r$-th block of $U$ satisfies

$$U_k^{(r)} = (W_{t_1})_k^{(r)} + a_r\,C\,(W_*)_k, \qquad \forall r \in [m], k \in [K]. \tag{153}$$

For every sample $i \in [n]$ and class $j \neq y_i$, let $M_{ij}(W) := z_i(W)_{y_i} - z_i(W)_j$. We compute the margin increment

$$\Delta M_{ij} := M_{ij}(U) - M_{ij}(W_{t_1}) = \frac{1}{m}\sum_{r=1}^{m} a_r\left[D_{i,y_i}^r - D_{i,j}^r\right], \tag{154}$$

where for brevity we let $s_{i,k}^r := x_i^\top (W_{t_1})_k^{(r)}$, $\Delta_i(k) := (W_*\,x_i)_k$ and

$$D_{i,k}^r := \phi\big(s_{i,k}^r + a_r\,C\,\Delta_i(k)\big) - \phi\big(s_{i,k}^r\big).$$

**Mean-value theorem.** Since $\phi$ is continuously differentiable (Assumption 3.1), by the mean-value theorem, for each $r \in [m]$ and $k \in \{y_i, j\}$, there exists a point $\xi_{i,k}^r \in \left(s_{i,k}^r,\, s_{i,k}^r + a_r\,C\,\Delta_i(k)\right)$ such that

$$D_{i,k}^r = \phi'\big(\xi_{i,k}^r\big)\cdot a_r\,C\,\Delta_i(k). \tag{155}$$

Substituting (155) into (154) and using $a_r^2 = 1$, we have that

$$\Delta M_{ij} \;=\; \frac{C}{m}\sum_{r=1}^{m}\big[\phi'(\xi_{i,y_i}^r)\,\Delta_i(y_i) - \phi'(\xi_{i,j}^r)\,\Delta_i(j)\big] \;=\; C\big[\bar{\phi}'_{i,y_i}\,\Delta_i(y_i) - \bar{\phi}'_{i,j}\,\Delta_i(j)\big], \tag{156}$$

where $\bar{\phi}'_{i,k} := \frac{1}{m}\sum_{r=1}^{m}\phi'(\xi_{i,k}^r) \in [\alpha, 1]$ by Assumption 3.1 (Item 1). Adding and subtracting $\bar{\phi}'_{i,y_i}\,\Delta_i(j)$, we have that

$$\begin{aligned}\bar{\phi}'_{i,y_i}\,\Delta_i(y_i) - \bar{\phi}'_{i,j}\,\Delta_i(j) \;&=\; \bar{\phi}'_{i,y_i}\big(\Delta_i(y_i) - \Delta_i(j)\big) + \big(\bar{\phi}'_{i,y_i} - \bar{\phi}'_{i,j}\big)\,\Delta_i(j)\\ &\geq\; \alpha\gamma + \big(\bar{\phi}'_{i,y_i} - \bar{\phi}'_{i,j}\big)\,\Delta_i(j),\end{aligned} \tag{157}$$

where at the last step we have used that $\bar{\phi}'_{i,y_i} \geq \alpha > 0$ and $\Delta_i(y_i) - \Delta_i(j) \geq \gamma > 0$ from Assumption 2.1. Since $\bar{\phi}'_{i,y_i}, \bar{\phi}'_{i,j} \in [\alpha, 1]$, we have $|\bar{\phi}'_{i,y_i} - \bar{\phi}'_{i,j}| \leq 1 - \alpha$. By the Cauchy–Schwarz inequality, $\|W_*\|_F = 1$, and $\|x_i\|_2 \leq 1$ from Assumption 2.1, it holds that

$$|\Delta_i(j)| \;=\; |(W_*)_j^\top x_i| \;\leq\; \|(W_*)_j\|_2\,\|x_i\|_2 \;\leq\; \|W_*\|_F\,\|x_i\|_2 \;\leq\; 1.$$

Combining these two bounds, we get

$$\big|\big(\bar{\phi}'_{i,y_i} - \bar{\phi}'_{i,j}\big)\,\Delta_i(j)\big| \;\leq\; 1 - \alpha. \tag{158}$$

Substituting (158) into (157), we have that

$$\bar{\phi}'_{i,y_i}\,\Delta_i(y_i) - \bar{\phi}'_{i,j}\,\Delta_i(j) \;\geq\; \alpha\gamma - (1-\alpha) \;=\; \tilde{\gamma}. \tag{159}$$

Using (159) with (156), we obtain

$$\Delta M_{ij} \;\geq\; C\,\tilde{\gamma} \;=\; \ln\big(\tilde{\gamma}^2\,\eta(t - t_1)\big) + 2\kappa, \tag{160}$$

Since $M_{ij}(U) = M_{ij}(W_{t_1}) + \Delta M_{ij}$, we have that

$$\begin{aligned}\exp\big(-M_{ij}(U)\big) \;&=\; \exp\big(-M_{ij}(W_{t_1})\big)\cdot\exp\big(-\Delta M_{ij}\big)\\ &\overset{(160)}{\leq}\; \exp\big(-M_{ij}(W_{t_1})\big)\cdot\exp\Big(-\ln\big(\tilde{\gamma}^2\,\eta(t-t_1)\big) - 2\kappa\Big)\\ &=\; \frac{e^{-2\kappa}\,\exp(-M_{ij}(W_{t_1}))}{\tilde{\gamma}^2\,\eta(t-t_1)}\\ &\leq\; \frac{\exp(-M_{ij}(W_{t_1}))}{\tilde{\gamma}^2\,\eta(t-t_1)},\end{aligned} \tag{161}$$

where the last inequality uses $e^{-2\kappa} \leq 1$ since $\kappa \geq 0$. Summing (161) over $i \in [n]$, multiplying by $\frac{1}{n}$ and summing again over $j \neq y_i$, we get

$$\begin{aligned}F(U) \;&=\; \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq y_i}\exp\big(-M_{ij}(U)\big)\\ &\leq\; \frac{1}{\tilde{\gamma}^2\,\eta(t-t_1)}\cdot\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq y_i}\exp\big(-M_{ij}(W_{t_1})\big)\\ &=\; \frac{F(W_{t_1})}{\tilde{\gamma}^2\,\eta(t-t_1)}.\end{aligned}$$

□

**Lemma D.8** (Two-layer squared-gradient comparison)**.** Let Assumptions 2.1, 3.1 hold. Then, for every $W \in \mathbb{R}^{md\times K}$, it holds that

$$\sum_{i=1}^{n}\|\nabla\ell_i(W)\|_F^2 \;\leq\; \frac{2n^2}{\tilde{\gamma}^2}\,\|\nabla L(W)\|_F^2. \tag{162}$$

*Proof.* We have that

$$\nabla \ell_i(W) \quad = \quad J_i(W)^\top \big(p_i(W) - e_{y_i}\big), \tag{163}$$

where $J_i(W)$ denotes the Jacobian of the logits $z_i(W)$ with respect to $W$. From Lemma D.1, we have

$$\|J_i(W)\|_{op} \quad \leq \quad \frac{1}{\sqrt{m}}. \tag{164}$$

Thus, it holds that

$$\begin{aligned} \|\nabla \ell_i(W)\|_F \quad &\leq \quad \|J_i(W)\|_{op} \, \|p_i(W) - e_{y_i}\|_2 \\ &\leq \quad \frac{1}{\sqrt{m}} \, \|p_i(W) - e_{y_i}\|_2. \end{aligned} \tag{165}$$

Taking the square on both sides, we obtain

$$\|\nabla \ell_i(W)\|_F^2 \quad \leq \quad \frac{1}{m} \, \|p_i(W) - e_{y_i}\|_2^2. \tag{166}$$

We have that

$$\begin{aligned} \|p_i(W) - e_{y_i}\|_2^2 \quad &= \quad \big(1 - p_i(W)_{y_i}\big)^2 + \sum_{j \neq y_i} p_i(W)_j^2 \\ &\leq \quad 2\big(1 - p_i(W)_{y_i}\big)^2, \end{aligned} \tag{167}$$

since $\sum_{j \neq y_i} p_i(W)_j = 1 - p_i(W)_{y_i}$. Substituting (167) into (166), we obtain

$$\|\nabla \ell_i(W)\|_F^2 \quad \leq \quad \frac{2}{m} \big(1 - p_i(W)_{y_i}\big)^2. \tag{168}$$

Summing over $i = 1, \ldots, n$, it follows that

$$\sum_{i=1}^n \|\nabla \ell_i(W)\|_F^2 \quad \leq \quad \frac{2}{m} \sum_{i=1}^n \big(1 - p_i(W)_{y_i}\big)^2. \tag{169}$$

Since $0 \leq 1 - p_i(W)_{y_i} \leq 1$, we have

$$\sum_{i=1}^n \big(1 - p_i(W)_{y_i}\big)^2 \quad \leq \quad \left(\sum_{i=1}^n \big(1 - p_i(W)_{y_i}\big)\right)^2 = n^2 G(W)^2. \tag{170}$$

Combining (169) and (170), we get

$$\sum_{i=1}^n \|\nabla \ell_i(W)\|_F^2 \quad \leq \quad \frac{2n^2}{m} \, G(W)^2. \tag{171}$$

On the other hand, by Lemma D.3, we have that

$$\sqrt{m} \, \|\nabla L(W)\|_F \quad \geq \quad \tilde{\gamma} \, G(W), \tag{172}$$

and therefore

$$G(W)^2 \quad \leq \quad \frac{m}{\tilde{\gamma}^2} \, \|\nabla L(W)\|_F^2. \tag{173}$$

Substituting (173) into (171), we conclude that

$$\sum_{i=1}^n \|\nabla \ell_i(W)\|_F^2 \quad \leq \quad \frac{2n^2}{\tilde{\gamma}^2} \, \|\nabla L(W)\|_F^2. \tag{174}$$

□

VARIANCE BOUND FOR TWO-LAYER NN

**Lemma D.9** (Variance bound). Let $g(W) = \frac{1}{b}\sum_{i\in B_t} \nabla \ell(f(x_i; W), y_i)$ denote the minibatch gradient for the two-layer model $f$ of (6). Then, it holds that

$$\mathbb{E}\big[\|g_t - \nabla L(W_t)\|_F^2 \mid \mathcal{F}_t\big] \;\le\; \frac{2}{mb}\, G(W_t), \tag{175}$$

where $G(W) = \frac{1}{n}\sum_{i=1}^{n}(1 - p_i(W)_{y_i})$.

*Proof.* We begin by decomposing the variance of the stochastic oracle. For uniform sampling of a minibatch $B_t$ of size $b \ge 1$, conditionally on the filtration $\mathcal{F}_t$, we have that

$$\mathbb{E}\big[\|g_t - \nabla L(W_t)\|_F^2 \mid \mathcal{F}_t\big] \;=\; \frac{1}{b}\,\mathrm{Var}_i\big(\nabla \ell_i(W_t)\big) \;\le\; \frac{1}{b}\,\mathbb{E}_i\big[\|\nabla \ell_i(W_t)\|_F^2\big], \tag{176}$$

where we used $\mathrm{Var}(Z) \le \mathbb{E}[\|Z\|^2]$ for any random variable $Z$. By Lemma D.3, we have that

$$\|\nabla \ell_i(W_t)\|_F^2 \;\le\; \frac{2}{m}\big(1 - p_i(W_t)_{y_i}\big)^2 \;\le\; \frac{2}{m}\big(1 - p_i(W_t)_{y_i}\big).$$

Summing for $i = 1, ..., n$ and dividing by $n$, we get

$$\mathbb{E}_i\big[\|\nabla \ell_i(W_t)\|_F^2\big] \;\le\; \frac{2}{m}\, G(W_t),$$

and substituting into (176), we obtain

$$\mathbb{E}\big[\|g_t - \nabla L(W_t)\|_F^2 \mid \mathcal{F}_t\big] \;\le\; \frac{2}{mb}\, G(W_t).$$

□

### D.3 PROOF OF LEMMA 3.2 FOR LEAKY ACTIVATIONS

*Proof.* We verify Assumption 3.1 for the leaky template $\tilde{\phi}(x) = c\,x + (1 - c)\,\phi(x)$, where $c \in (1/(1+\gamma),\, 1)$.

**Continuous differentiability.** By Example 2.1 of Cai et al. (2024), each base $\phi \in \{\mathrm{GELU}, \mathrm{Softplus}, \mathrm{SiLU}, \tanh, \sigma\}$ is continuously differentiable on $\mathbb{R}$. The Huberized ReLU is continuously differentiable on $\mathbb{R}$ at the breakpoints 0 and $h$ by direct verification. Hence, $\tilde{\phi}$ is continuously differentiable.

**Derivative condition.** Differentiating the leaky template, we have

$$\tilde{\phi}'(x) \;=\; c + (1 - c)\,\phi'(x). \tag{177}$$

By Example 2.1 of Cai et al. (2024) and direct computation, each base satisfies $0 \le \phi'(x) \le M_\phi$ uniformly, where

- $M_\phi = 1$ for tanh, sigmoid, softplus, and Huberized ReLU
- $M_\phi \le 1 + e^{-1/2}/\sqrt{2\pi}$ for GELU
- $M_\phi \le 2$ for SiLU.

Therefore, for every $x \in \mathbb{R}$, it holds that

$$c \;\le\; \tilde{\phi}'(x) \;\le\; c + (1 - c)\, M_\phi.$$

For tanh, sigmoid, softplus, and Huberized ReLU, $M_\phi \le 1$, so $\tilde{\phi}'(x) \in [c, 1]$ directly. For GELU and SiLU, after absorbing the multiplicative factor $1/(c + (1 - c)\, M_\phi) \le 1$ into the definition of $\tilde{\phi}$ (an inconsequential rescaling that preserves all subsequent constants up to a constant factor), we again have $\tilde{\phi}'(x) \in [c, 1]$. Setting $\alpha := c$, we obtain

$$\alpha \;\le\; |\tilde{\phi}'(x)| \;\le\; 1 \qquad \text{for every } x \in \mathbb{R},$$

where $\alpha = c \in (1/(1+\gamma),\, 1)$ by the choice of $c$. This verifies the derivative condition.

**Smoothness.** Since $\tilde{\phi}'(x) = c + (1-c)\,\phi'(x)$, for any $x, y \in \mathbb{R}$, it holds that

$$|\tilde{\phi}'(x) - \tilde{\phi}'(y)| \;=\; (1-c)\,|\phi'(x) - \phi'(y)| \;\le\; (1-c)\,L_\phi\,|x-y|,$$

where $L_\phi := \sup_{x\in\mathbb{R}} |\phi''(x)|$ is the Lipschitz constant of $\phi'$. From Example 2.1 of Cai et al. (2024) (see also our verifications below): $L_{\text{GELU}} \le 2$, $L_{\text{Softplus}} \le 1$, $L_{\text{SiLU}} \le 4$, $L_{\tanh} \le 1$, $L_\sigma \le 1$, and $L_{\text{Huber-ReLU}_h} = 1/h$. Hence $\tilde{\phi}'$ is $\tilde{\beta}$-Lipschitz with $\tilde{\beta} = (1-c)\,L_\phi$, giving the values listed in the lemma ($\tilde{\beta} = 4(1-c)$ is a uniform upper bound for the GELU/Softplus/SiLU family that absorbs the worst case; $\tilde{\beta} = (1-c)/h$ for Huberized ReLU; $\tilde{\beta} = 1-c$ for tanh and sigmoid).

**Near-homogeneity.** By linearity of the perturbation,

$$\begin{aligned}\tilde{\phi}(z) - \tilde{\phi}'(z)\,z \;&=\; \big[c\,z + (1-c)\,\phi(z)\big] - \big[c + (1-c)\,\phi'(z)\big]\,z \\ &=\; c\,z + (1-c)\,\phi(z) - c\,z - (1-c)\,\phi'(z)\,z \\ &=\; (1-c)\,\big[\phi(z) - \phi'(z)\,z\big], \end{aligned} \tag{178}$$

where the linear contribution $c\,z$ cancels exactly. Hence

$$|\tilde{\phi}(z) - \tilde{\phi}'(z)\,z| \;=\; (1-c)\,|\phi(z) - \phi'(z)\,z| \;\le\; (1-c)\,\kappa_\phi,$$

where $\kappa_\phi := \sup_z |\phi(z) - \phi'(z)\,z|$ denotes the homogeneity error of the base. From Example 2.1 of Cai et al. (2024): $\kappa_{\text{GELU}} = e^{-1/2}/\sqrt{2\pi}$, $\kappa_{\text{Softplus}} = \log 2$, $\kappa_{\text{SiLU}} = 1$, $\kappa_{\text{Huber-ReLU}_h} = h/2$, $\kappa_{\tanh} \le 1$, and $\kappa_\sigma \le 1$. Hence $\kappa = (1-c)\,\kappa_\phi$, giving the values listed in the lemma (taking $\kappa_\phi \le 1$ as a uniform upper bound). This completes the proof. □

### D.4 Proofs for the EoS Regime

**Lemma D.10.** Let $U = U_1 + U_2$ with $U_1 \in \mathbb{R}^{md\times K}$ and $U_2 = \frac{\eta(1+\frac{1}{b})}{\tilde{\gamma}}\,\overline{W}_*$. Then, for all $t \ge 1$, it holds that

$$\frac{\mathbb{E}\big[\|W_t - U\|_F^2\big]}{2\tilde{\eta}\,t} + \frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}[L(W_k)] \;\le\; L(U_1) + \frac{\|W_0 - U\|_F^2}{2\tilde{\eta}\,t}. \tag{179}$$

*Proof.* From the update rule of SGD with $\tilde{\eta} = m\eta$, we have that

$$\begin{aligned}\|W_{t+1} - U\|_F^2 \;&=\; \|W_t - U\|_F^2 + 2\tilde{\eta}\,\langle g_t, U - W_t\rangle_F + \tilde{\eta}^2\|g_t\|_F^2 \\ &=\; \|W_t - U\|_F^2 + 2\tilde{\eta}\,\langle g_t, U_1 - W_t\rangle_F + \tilde{\eta}^2\Big(\frac{2}{\tilde{\eta}}\langle g_t, U_2\rangle_F + \|g_t\|_F^2\Big).\end{aligned}$$

Taking expectation conditional on $\mathcal{F}_t$ and using the unbiasedness $\mathbb{E}[g_t \mid \mathcal{F}_t] = \nabla L(W_t)$, we have that

$$\begin{aligned}\mathbb{E}\big[\|W_{t+1} - U\|_F^2 \,\big|\, \mathcal{F}_t\big] \;&=\; \|W_t - U\|_F^2 + 2\tilde{\eta}\,\langle \nabla L(W_t), U_1 - W_t\rangle_F \\ &\quad + \tilde{\eta}^2\Big(\frac{2}{\tilde{\eta}}\langle \nabla L(W_t), U_2\rangle_F + \mathbb{E}\big[\|g_t\|_F^2 \mid \mathcal{F}_t\big]\Big).\end{aligned} \tag{180}$$

We, next, show that with the choice $U_2 = \frac{\eta(1+\frac{1}{b})}{\tilde{\gamma}}\,\overline{W}_*$, the last term in (180) is non-positive. We have that

$$\begin{aligned}\frac{2}{\tilde{\eta}}\langle \nabla L(W_t), U_2\rangle_F \;&=\; \frac{2(1+\frac{1}{b})}{m\tilde{\gamma}}\,\langle \nabla L(W_t), \overline{W}_*\rangle_F \\ &\overset{\text{Lemma D.2}}{\le}\; -\frac{2(1+\frac{1}{b})}{m}\,G(W_t).\end{aligned}$$

For the second moment $\mathbb{E}\big[\|g_t\|_F^2 \mid \mathcal{F}_t\big]$, the bias-variance decomposition together with Lemmas D.3 and D.9 give

$$\begin{aligned}\mathbb{E}\big[\|g_t\|_F^2 \mid \mathcal{F}_t\big] &= \|\nabla L(W_t)\|_F^2 + \mathbb{E}\big[\|g_t - \nabla L(W_t)\|_F^2 \mid \mathcal{F}_t\big] \\ &\overset{\text{Lemma D.3}}{\leq} \frac{2}{m} G(W_t)^2 + \frac{2}{mb} G(W_t) \\ &\leq \frac{2}{m} G(W_t) + \frac{2}{mb} G(W_t) \\ &= \frac{2(1+\frac{1}{b})}{m} G(W_t),\end{aligned}$$

where we used $G(W_t)^2 \leq G(W_t)$ since $G(W_t) \in [0,1]$. Summing the two bounds, we obtain

$$\frac{2}{\tilde{\eta}}\langle \nabla L(W_t), U_2\rangle_F + \mathbb{E}\big[\|g_t\|_F^2 \mid \mathcal{F}_t\big] \;\leq\; -\frac{2(1+\frac{1}{b})}{m} G(W_t) + \frac{2(1+\frac{1}{b})}{m} G(W_t) \;=\; 0. \tag{181}$$

Substituting (181) into (180), we obtain

$$\mathbb{E}\big[\|W_{t+1} - U\|_F^2 \,\big|\, \mathcal{F}_t\big] \;\leq\; \|W_t - U\|_F^2 + 2\tilde{\eta}\,\langle \nabla L(W_t), U_1 - W_t\rangle_F.$$

Using the convexity of $L$, we have that $\langle \nabla L(W_t), U_1 - W_t\rangle_F \leq L(U_1) - L(W_t)$, and therefore we obtain

$$\mathbb{E}\big[\|W_{t+1} - U\|_F^2 \,\big|\, \mathcal{F}_t\big] \;\leq\; \|W_t - U\|_F^2 + 2\tilde{\eta}\big[L(U_1) - L(W_t)\big]. \tag{182}$$

Taking expectation again and using the tower law of expectation, we have that

$$\mathbb{E}\big[\|W_{t+1} - U\|_F^2\big] \;\leq\; \mathbb{E}\Big[\|W_t - U\|_F^2\Big] + 2\tilde{\eta}\,\mathbb{E}\Big[L(U_1) - L(W_t)\Big]$$

Summing for $k = 0, \ldots, t-1$ and dividing by $2\tilde{\eta}\,t$, we obtain

$$\frac{\mathbb{E}\big[\|W_t - U\|_F^2\big]}{2\tilde{\eta}\,t} + \frac{1}{t}\sum_{k=0}^{t-1} \mathbb{E}[L(W_k)] \;\leq\; L(U_1) + \frac{\|W_0 - U\|_F^2}{2\tilde{\eta}\,t}$$

□

**Theorem D.11** (Restatement of Theorem 3.3)**.** Let Assumptions 2.1, 3.1, hold and assume without loss of generality that $W_0 = \mathbf{0}$. The iterates of SGD with any step size $\tilde{\eta} = m\eta > 0$ and batch size $b \geq 1$ applied to the two-layer network (6) satisfy for any $T \geq 1$ that

$$\min_{0\leq k\leq T-1} \mathbb{E}[L(W_k)] \;\leq\; \frac{K - 1 + 2\ln^2(\tilde{\gamma}^2\eta T) + 8\kappa^2 + \eta^2(1+\frac{1}{b})^2}{\tilde{\gamma}^2\eta\,T}.$$

*Proof.* From Lemma D.10, we have that for all $T \geq 1$ it holds

$$\begin{aligned}\frac{1}{T}\sum_{k=0}^{T-1} \mathbb{E}[L(W_k)] &\leq L(U_1) + \frac{\|U\|_F^2}{2\tilde{\eta}\,T} - \frac{\mathbb{E}\big[\|W_T - U\|_F^2\big]}{2\tilde{\eta}\,T} \\ &\leq L(U_1) + \frac{\|U\|_F^2}{2\tilde{\eta}\,T},\end{aligned} \tag{183}$$

where we used $W_0 = \mathbf{0}$. For $U_1 = \frac{\ln(\tilde{\gamma}^2\eta T) + 2\kappa}{\tilde{\gamma}}\,\overline{W}_*, U_2 = \frac{\eta(1+\frac{1}{b})}{\tilde{\gamma}}\,\overline{W}_*$ we have from Lemma D.6 that

$$L(U_1) \;\leq\; \frac{K-1}{\tilde{\gamma}^2\eta T}.$$

From inequality (137), we have that

$$\|U\|_F^2 \;\leq\; \frac{4m\,\ln^2(\tilde{\gamma}^2\eta T) + 16m\,\kappa^2 + 2m\eta^2(1+\frac{1}{b})^2}{\tilde{\gamma}^2} \tag{184}$$

Substituting (184) into (183) and using $\tilde{\eta}\, T = m\eta T$, we obtain

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[L(W_k)] \;\le\; \frac{K-1+2\ln^2(\tilde{\gamma}^2\eta T)+8\kappa^2+\eta^2(1+\frac{1}{b})^2}{\tilde{\gamma}^2\eta\, T}$$

and hence we have

$$\min_{0\le k\le T-1}\mathbb{E}[L(W_k)] \;\le\; \frac{K-1+2\ln^2(\tilde{\gamma}^2\eta T)+8\kappa^2+\eta^2(1+\frac{1}{b})^2}{\tilde{\gamma}^2\eta\, T}$$

□

### D.5 Proofs for the Stable Regime

Throughout this section, we let

$$\begin{aligned}\tilde{L}_{NN} &:= \min\left\{\frac{1}{8\eta(1+\tilde{\beta})(1+\frac{2n}{b\min\{\tilde{\gamma}^2,1\}})}, \frac{1}{2ne^{\kappa+2}}\right\}\\ \mathcal{S}_{NN} &:= \left\{W\in\mathbb{R}^{md\times K} : L(W) \le \tilde{L}_{NN}\right\}.\end{aligned}$$

**Lemma D.12.** If for some $t \ge 0$ the iterates of SGD for the two-layer network (6) satisfy $L(W_t) \le \tilde{L}_{NN}$, then it holds that

$$\mathbb{E}\left[L(W_{t+1}) - L(W_t)\middle|\mathcal{F}_t\right] \;\le\; 0, \tag{185}$$

and thus $W_t \in \mathcal{S}_{NN}$.

*Proof.* Recall that the minibatch gradient is $g_t = \frac{1}{b}\sum_{j\in\mathcal{B}_t}\nabla\ell_j(W_t) \in \mathbb{R}^{md\times K}$ and decomposes block-wise as $g_t = (g_t^{(1)},\dots,g_t^{(m)})$ with $g_t^{(r)} \in \mathbb{R}^{d\times K}$. Fix $i\in[n]$ and define the logit increment

$$\Delta z_i \;:=\; z_i(W_{t+1}) - z_i(W_t) \;=\; \frac{1}{m}\sum_{r=1}^{m} a_r\,\phi\big(x_i^\top W_{t+1}^{(r)}\big) - \frac{1}{m}\sum_{r=1}^{m} a_r\,\phi\big(x_i^\top W_t^{(r)}\big).$$

We argue component-wise. For every class $k\in[K]$, we have that

$$\big|(\Delta z_i)_k\big| \;\le\; \frac{1}{m}\sum_{r=1}^{m}|a_r|\,\big|\phi\big(x_i^\top W_{t+1,\,k}^{(r)}\big) - \phi\big(x_i^\top W_{t,\,k}^{(r)}\big)\big|, \tag{186}$$

where $W_{t,\,k}^{(r)} \in \mathbb{R}^d$ denotes the $k$-th column of the matrix $W_t^{(r)} \in \mathbb{R}^{d\times K}$. Since $|\phi'| \le 1$ from Assumption 3.1, the activation $\phi$ is 1-Lipschitz, and hence

$$\begin{aligned}\big|\phi\big(x_i^\top W_{t+1,\,k}^{(r)}\big) - \phi\big(x_i^\top W_{t,\,k}^{(r)}\big)\big| &\le \big|x_i^\top W_{t+1,\,k}^{(r)} - x_i^\top W_{t,\,k}^{(r)}\big|\\ &\le \|x_i\|_2\,\big\|W_{t+1,\,k}^{(r)} - W_{t,\,k}^{(r)}\big\|_2\\ &\overset{\|x_i\|_2\le 1}{\le} \big\|W_{t+1,\,k}^{(r)} - W_{t,\,k}^{(r)}\big\|_2\\ &\overset{\text{(SGD)}}{=} \tilde{\eta}\,\big\|(g_t^{(r)})_k\big\|_2,\end{aligned} \tag{187}$$

where the second inequality uses Cauchy–Schwarz on the inner product $x_i^\top\cdot$, the third uses $\|x_i\|_2 \le 1$, and the equality uses the (SGD) update $W_{t+1}^{(r)} = W_t^{(r)} - \tilde{\eta}\, g_t^{(r)}$ applied column-wise. Substitut-

ing (187) into (186) and applying the Cauchy–Schwarz inequality across the $m$ summands, we have that

$$\begin{aligned}\big|(\Delta z_i)_k\big| \;&\le\; \frac{\tilde\eta}{m}\sum_{r=1}^{m}|a_r|\,\big\|(g_t^{(r)})_k\big\|_2 \\ &\le\; \frac{\tilde\eta}{m}\Big(\sum_{r=1}^{m}a_r^2\Big)^{1/2}\Big(\sum_{r=1}^{m}\big\|(g_t^{(r)})_k\big\|_2^2\Big)^{1/2} \\ &=\; \frac{\tilde\eta}{\sqrt{m}}\Big(\sum_{r=1}^{m}\big\|(g_t^{(r)})_k\big\|_2^2\Big)^{1/2},\end{aligned} \tag{188}$$

where in the equality we have used $\sum_{r=1}^m a_r^2 = m$ since $|a_r| = 1$. Since $\sum_{r=1}^m \|(g_t^{(r)})_k\|_2^2 \le \sum_{r=1}^m \sum_{k'=1}^K \|(g_t^{(r)})_{k'}\|_2^2 = \|g_t\|_F^2$, we obtain from (188) that

$$\big|(\Delta z_i)_k\big| \;\le\; \frac{\tilde\eta}{\sqrt{m}}\,\|g_t\|_F. \tag{189}$$

Since (189) holds for every $k \in [K]$, taking the maximum over $k$, we obtain

$$\begin{aligned}\|\Delta z_i\|_\infty \;&\le\; \frac{\tilde\eta}{\sqrt{m}}\,\|g_t\|_F \\ &\le\; \frac{\tilde\eta}{b\sqrt{m}}\sum_{j\in B_t}\|\nabla\ell_j(W_t)\|_2 \\ &\overset{(124)}{\le}\; \sqrt{2}\frac{n\eta}{b}L(W_t).\end{aligned} \tag{190}$$

where at the last step we have used Lemma D.3. Let $\psi_i(\theta) := \ell_i(W_t - \theta\,\tilde\eta\,g_t)$ for $\theta \in [0,1]$. From Taylor's theorem, there exists $\theta_i \in (0,1)$ such that

$$\ell_i(W_{t+1}) \;=\; \ell_i(W_t) - \tilde\eta\,\langle\nabla\ell_i(W_t), g_t\rangle + \frac{\tilde\eta^2}{2}\,\big\langle g_t,\, \nabla^2\ell_i(W_t - \theta_i\,\tilde\eta\,g_t)\,g_t\big\rangle. \tag{191}$$

By Lemma D.4, applied at the shifted point $W_t - \theta_i\,\tilde\eta\,g_t$, we have that

$$\big\|\nabla^2\ell_i(W_t - \theta_i\,\tilde\eta\,g_t)\big\|_{\mathrm{op}} \;\le\; \frac{2(1+\tilde\beta)}{m}\,\ell_i\big(W_t - \theta_i\,\tilde\eta\,g_t\big). \tag{192}$$

In the stable regime, the loss is small enough that the logits move by at most a constant amount. In particular, by (190) and the fact that $L(W_t) \le \frac{b}{\sqrt{2}n\eta}$, we have $\|\Delta z_i\|_\infty \le 1$. Then, the softmax probabilities along the segment between $W_t$ and $W_{t+1}$ change by at most an absolute constant factor, and hence

$$\ell_i\big(W_t - \theta_i\,\tilde\eta\,g_t\big) \;\le\; e^2\,\ell_i(W_t). \tag{193}$$

Substituting (193) into (192) and using $e^2 < 8$, we obtain

$$\big\|\nabla^2\ell_i(W_t - \theta_i\,\tilde\eta\,g_t)\big\|_{\mathrm{op}} \;\le\; \frac{2e^2(1+\tilde\beta)}{m}\,\ell_i(W_t) \;\le\; \frac{16(1+\tilde\beta)}{m}\,\ell_i(W_t). \tag{194}$$

Substituting (194) into (191) via $\langle g_t, Ag_t\rangle \le \|A\|_{\mathrm{op}}\|g_t\|^2$, we get

$$\ell_i(W_{t+1}) \;\le\; \ell_i(W_t) - \tilde\eta\,\langle\nabla\ell_i(W_t), g_t\rangle + \frac{8(1+\tilde\beta)\,\tilde\eta^2}{m}\,\ell_i(W_t)\,\|g_t\|_F^2. \tag{195}$$

Summing (195) over $i = 1, \dots, n$ and dividing by $n$, we have that

$$L(W_{t+1}) - L(W_t) \;\le\; -\tilde\eta\,\langle\nabla L(W_t), g_t\rangle + \frac{8(1+\tilde\beta)\,\tilde\eta^2}{m}\,L(W_t)\,\|g_t\|_F^2. \tag{196}$$

Taking conditional expectation with respect to $\mathcal{F}_t$ and using $\mathbb{E}[g_t \mid \mathcal{F}_t] = \nabla L(W_t)$, we obtain

$$\mathbb{E}\big[L(W_{t+1}) - L(W_t) \mid \mathcal{F}_t\big] \;\le\; -\tilde\eta\,\|\nabla L(W_t)\|_F^2 + \frac{8(1+\tilde\beta)\,\tilde\eta^2}{m}\,L(W_t)\,\mathbb{E}\big[\|g_t\|_F^2 \mid \mathcal{F}_t\big]. \tag{197}$$

Using the bias-variance decomposition, we have that

$$\begin{aligned}\mathbb{E}\big[\|g_t\|_F^2 \mid \mathcal{F}_t\big] &= \|\nabla L(W_t)\|_F^2 + \mathbb{E}\big[\|g_t - \nabla L(W_t)\|_F^2 \mid \mathcal{F}_t\big] \\ &\overset{(176)}{\leq} \|\nabla L(W_t)\|_F^2 + \frac{1}{nb}\sum_{i=1}^{n}\|\nabla \ell_i(W_t)\|_F^2 \\ &\leq \|\nabla L(W_t)\|_F^2 + \frac{2n}{b\tilde{\gamma}^2}\|\nabla L(W_t)\|_F^2, \end{aligned} \tag{198}$$

where at the last step we have used Lemma D.8. Substituting (198) into (197) and using the fact that $\tilde{\eta}^2/m = \tilde{\eta}\,\eta$, we get

$$\begin{aligned}\mathbb{E}\big[L(W_{t+1}) - L(W_t) \mid \mathcal{F}_t\big] &\leq -\tilde{\eta}\,\|\nabla L(W_t)\|_F^2 + 8(1+\tilde{\beta})\,\tilde{\eta}\,\eta\Big(1 + \frac{2n}{\tilde{\gamma}^2\,b}\Big)\,L(W_t)\,\|\nabla L(W_t)\|_F^2 \\ &= -\tilde{\eta}\,\Big[1 - 8\,\eta\,(1+\tilde{\beta})\Big(1+\frac{2n}{\tilde{\gamma}^2\,b}\Big)\,L(W_t)\Big]\,\|\nabla L(W_t)\|_F^2, \end{aligned} \tag{199}$$

For $L(W_t) \leq \min\left\{\frac{1}{8\,\eta\,(1+\tilde{\beta})\left(1+\frac{2n}{\tilde{\gamma}^2\,b}\right)}, \frac{b}{\sqrt{2}n\eta}\right\} \leq \tilde{L}_{NN}$ with $\tilde{L}_{NN} := \min\left\{\frac{1}{8\eta(1+\tilde{\beta})(1+\frac{2n}{b\min\{\tilde{\gamma}^2,1\}})}, \frac{1}{2ne^{\kappa+2}}\right\}$, we have that

$$\mathbb{E}\big[L(W_{t+1}) - L(W_t) \mid \mathcal{F}_t\big] \leq 0.$$

Taking expectation on both sides and applying the tower law of expectation, we obtain

$$\mathbb{E}[L(W_{t+1}) - L(W_t)] \leq 0.$$

□

**Lemma D.13.** If the iterates of SGD satisfy $W_k \in \mathcal{S}_{NN}, \forall k \in [t_1, t-1], t_1 > 0$, then it holds that

$$\mathbb{E}[L(W_t)] \leq 2\frac{F(W_{t_1}) + \ln^2(\tilde{\gamma}^2\eta\,(t-t_1)) + \kappa^2}{\tilde{\gamma}^2\,\eta\,(t-t_1)}, \tag{200}$$

where $F(W) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq y_i}\exp\big(-(z_i(W)_{y_i} - z_i(W)_j)\big)$.

*Proof.* From the update rule of SGD, we have that

$$\|W_{t+1} - U\|_F^2 = \|W_t - U\|_F^2 - 2\tilde{\eta}\,\langle g_t, W_t - U\rangle_F + \tilde{\eta}^2\|g_t\|_F^2.$$

Taking expectation conditional on the filtration $\mathcal{F}_t$ and using $\mathbb{E}[g_t \mid \mathcal{F}_t] = \nabla L(W_t)$, we have that

$$\mathbb{E}\big[\|W_{t+1} - U\|_F^2 \mid \mathcal{F}_t\big] = \|W_t - U\|_F^2 - 2\tilde{\eta}\,\langle \nabla L(W_t), W_t - U\rangle_F + \tilde{\eta}^2\,\mathbb{E}\big[\|g_t\|_F^2 \mid \mathcal{F}_t\big]. \tag{201}$$

Since $L$ is convex, we have that

$$-\langle \nabla L(W_t), W_t - U\rangle_F \leq L(U) - L(W_t) \tag{202}$$

Substituting (202) into (201), we obtain

$$\|W_{t+1} - U\|_F^2 \leq \|W_t - U\|_F^2 + 2\tilde{\eta}\big(L(U) - L(W_t)\big) + \tilde{\eta}^2\,\mathbb{E}\big[\|g_t\|_F^2 \mid \mathcal{F}_t\big]. \tag{203}$$

Using inequality (198) from Lemma D.12, we have that

$$\begin{aligned}\tilde{\eta}^2\,\mathbb{E}\big[\|g_t\|_F^2 \mid \mathcal{F}_t\big] &\leq \tilde{\eta}^2\Big(1 + \frac{2n}{\tilde{\gamma}^2\,b}\Big)\|\nabla L(W_t)\|_F^2 \\ &\overset{(125)}{\leq} \frac{2\,\tilde{\eta}^2}{m}\left(1 + \frac{2n}{\tilde{\gamma}^2\,b}\right)L(W_t)^2 \\ &\leq \frac{\tilde{\eta}}{4}\,L(W_t), \end{aligned} \tag{204}$$

where at the last step we have used $\frac{\tilde{\eta}^2}{m} = \tilde{\eta}\,\eta$ and that $L(W_t) \le \tilde{L}_{NN} \le \frac{1}{8\,\eta\left(1+\frac{2n}{\tilde{\gamma}^2 b}\right)}$. Thus, substituting (204) into (203), we obtain

$$\mathbb{E}\left[\|W_{t+1}-U\|_F^2 \mid \mathcal{F}_t\right] \le \|W_t-U\|_F^2 + 2\tilde{\eta}\,L(U) - \frac{7\tilde{\eta}}{4}\,L(W_t). \tag{205}$$

Taking expectation again, using the tower law and rearranging the terms, we obtain

$$\mathbb{E}[L(W_t)] \le \frac{8}{7}\,L(U) + \frac{4\,\mathbb{E}\left[\|W_t-U\|_F^2 - \|W_{t+1}-U\|_F^2\right]}{7\,\tilde{\eta}}.$$

Summing over $k = t_1, \dots, t-1$ and dividing by $(t-t_1)$, we have that

$$\frac{1}{t-t_1}\sum_{k=t_1}^{t-1}\mathbb{E}[L(W_k)] \le \frac{8}{7}\,L(U) + \frac{4\left(\|W_{t_1}-U\|_F^2 - \mathbb{E}[\|W_t-U\|_F^2]\right)}{7\,\tilde{\eta}\,(t-t_1)}. \tag{206}$$

Letting $U = W_{t_1} + U_1$ with $U_1 = \frac{\ln\left(\tilde{\gamma}^2\eta\,(t-t_1)\right)+2\kappa}{\tilde{\gamma}}\,\overline{W}_*$, we have that

$$\|W_{t_1}-U\|_F^2 = \|U_1\|_F^2 = \frac{\left[\ln\left(\tilde{\gamma}^2\,\eta\,(t-t_1)\right)+2\kappa\right]^2}{\tilde{\gamma}^2}\,\|\overline{W}_*\|_F^2 \le \frac{2m\,\ln^2(\tilde{\gamma}^2\eta\,(t-t_1)) + 8m\,\kappa^2}{\tilde{\gamma}^2},$$

where we used $(a+b)^2 \le 2a^2+2b^2$ and $\|\overline{W}_*\|_F^2 = m$. From Lemma D.7, we have that

$$L(U) \le F(U) \le \frac{F(W_{t_1})}{\tilde{\gamma}^2\,\eta\,(t-t_1)}.$$

Substituting into (206) and using $\tilde{\eta} = m\eta$, we obtain

$$\min_{t_1\le k\le t-1}\mathbb{E}[L(W_k)] \le \frac{8F(W_{t_1}) + 8\,\ln^2(\tilde{\gamma}^2\eta\,(t-t_1)) + 8\kappa^2}{7\tilde{\gamma}^2\,\eta\,(t-t_1)}, \tag{207}$$

Applying Lemma D.12 for the iterates $W_k \in \mathcal{S}_{NN}, \forall k \in [t_1, t-1]$ we have that

$$\mathbb{E}[L(W_t)] \le \mathbb{E}[L(W_{t-1})] \le \cdots \le \mathbb{E}[L(W_{t_1})], \tag{208}$$

and thus from (207) we conclude that

$$\mathbb{E}[L(W_t)] \le 2\frac{F(W_{t_1}) + \ln^2(\tilde{\gamma}^2\eta\,(t-t_1)) + \kappa^2}{\tilde{\gamma}^2\,\eta\,(t-t_1)}.$$

□

**Lemma D.14.** Let $W_0 = \mathbf{0}$. There exists $t_{\text{in}} \le t_{\max}(\delta)$ such that with probability at least $1-\delta$, we have that $L(W_{t_{in}}) \le \tilde{L}_{NN}$ with

$$t_{\max}(\delta) := \frac{1}{\tilde{\gamma}^2}\max\left\{\frac{8\left(K-1+4\kappa+2\eta(2+\frac{1}{b})\right)}{\eta\delta\tilde{L}_{NN}},\ \frac{32}{\eta\delta\tilde{L}_{NN}}\log\left(\frac{32}{\eta\delta\tilde{L}_{NN}}\right)\right\}.$$

*Proof.* From Lemma D.5 applied for $t = t_{\max}(\delta)$, we have that

$$\frac{1}{t_{\max}}\sum_{k=0}^{t_{\max}-1}\mathbb{E}[G(W_k)] \le \frac{2\left(K-1+2\ln\left(\tilde{\gamma}^2\eta t_{\max}\right)+4\kappa+2\eta(1+\frac{1}{b})\right)}{\eta\tilde{\gamma}^2 t_{\max}}.$$

Using $\ln\left(\tilde{\gamma}^2\eta t_{\max}\right) \le \ln\left(\tilde{\gamma}^2 t_{\max}\right) + \eta$ and absorbing constants, we get

$$\frac{1}{t_{\max}}\sum_{k=0}^{t_{\max}-1}\mathbb{E}[G(W_k)] \le \frac{2\left(K-1+2\ln\left(\tilde{\gamma}^2 t_{\max}\right)+4\kappa+2\eta(2+\frac{1}{b})\right)}{\eta\tilde{\gamma}^2 t_{\max}}. \tag{209}$$

For $\delta \in (0,1)$, we select $t_{\max}$ so that the right-hand side of (209) is at most $\delta\,\tilde{L}_{NN}/2$, where $\tilde{L}_{NN} := \min\left\{\frac{1}{8\eta(1+\tilde{\beta})(1+n+\frac{2n}{\tilde{\gamma}^2 b})}, \frac{1}{2ne^{\kappa+2}}\right\}$. Specifically, we next verify the formula for the $t_{\max}$:

- For $t \geq \frac{8\left(K-1+4\kappa+2\eta(2+\frac{1}{b})\right)}{\eta\delta\tilde{L}_{NN}\tilde{\gamma}^2}$, it holds that $\frac{2\left(K-1+4\kappa+2\eta(2+\frac{1}{b})\right)}{\eta\tilde{\gamma}^2 t} \leq \frac{\delta\tilde{L}_{NN}}{4}$.
- For $t \geq \frac{32}{\eta\delta\tilde{L}_{NN}\tilde{\gamma}^2}\log\left(\frac{32}{\eta\delta\tilde{L}_{NN}}\right)$, it holds that $\tilde{\gamma}^2 t \geq \frac{32}{\eta\delta\tilde{L}_{NN}}\log\left(\frac{3}{\eta\delta\tilde{L}_{NN}}\right)$, which is a sufficient condition (see Lemma G.5 in Cai et al. (2024)) for $\frac{4\log\left(\tilde{\gamma}^2 t\right)}{\eta\tilde{\gamma}^2 t} \leq \frac{\delta\tilde{L}_{NN}}{4}$.

Thus, it suffices to select $t_{\max}(\delta)$ such that

$$t_{\max}(\delta) \;\geq\; \frac{1}{\tilde{\gamma}^2}\max\left\{\frac{8\left(K-1+4\kappa+2\eta(2+\frac{1}{b})\right)}{\eta\delta\tilde{L}_{NN}},\; \frac{32}{\eta\delta\tilde{L}_{NN}}\log\left(\frac{32}{\eta\delta\tilde{L}_{NN}}\right)\right\}. \quad (210)$$

Thus, there exists $t_{\text{in}} \leq t_{\max}(\delta)$ such that

$$\mathbb{E}[G(W_{t_{\text{in}}})] \;\leq\; \frac{\delta\,\tilde{L}_{NN}}{2}.$$

Define the event $\mathcal{E}_{\text{in}} := \{G(W_{t_{\text{in}}}) \leq \frac{\tilde{L}_{NN}}{2}\}$. Since $G(W_{t_{\text{in}}}) \geq 0$, by Markov's inequality we have that

$$\mathbb{P}(\mathcal{E}_{\text{in}}^c) \;=\; \mathbb{P}\big(G(W_{t_{\text{in}}}) > \frac{\tilde{L}_{NN}}{2}\big) \;\leq\; \frac{2\mathbb{E}[G(W_{t_{\text{in}}})]}{\tilde{L}_{NN}} \;\leq\; \delta,$$

and hence $\mathbb{P}(\mathcal{E}_{\text{in}}) \geq 1-\delta$. Conditioning on the event $\mathcal{E}_{\text{in}}$ and using the fact $G(W_{t_{\text{in}}}) \leq \frac{1}{2ne^{\kappa+2}} \leq \frac{1}{2n}$, every term in $G(W_{t_{\text{in}}}) = \frac{1}{n}\sum_i(1-p_i(W_{t_{\text{in}}})_{y_i})$ satisfies $1-p_i(W_{t_{\text{in}}})_{y_i} \leq 1/(2e^{\kappa+2}) \leq 1/2$ and thus $p_i(W_{t_{\text{in}}})_{y_i} \geq 1/2$. Hence, for every $j \neq y_i$, it holds

$$z_i(W_{t_{\text{in}}})_{y_i} - z_i(W_{t_{\text{in}}})_j \;=\; \ln\frac{p_i(W_{t_{\text{in}}})_{y_i}}{p_i(W_{t_{\text{in}}})_j} \;\geq\; 0,$$

which gives

$$\begin{aligned} F(W_{t_{\text{in}}}) \;&=\; \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq y_i}\frac{p_i(W_{t_{\text{in}}})_j}{p_i(W_{t_{\text{in}}})_{y_i}} \\ &\leq\; \frac{1}{n}\sum_{i=1}^{n} 2\left(1-p_i(W_{t_{\text{in}}})_{y_i}\right) \\ &=\; 2\,G(W_{t_{\text{in}}}) \\ &\leq\; \tilde{L}_{NN}, \end{aligned}$$

where we used $p_i(W)_{y_i} \geq 1/2$ and $\sum_{j\neq y_i} p_i(W)_j = 1-p_i(W)_{y_i}$. Since $L(W) \leq F(W), \forall W \in \mathbb{R}^{md\times K}$, on the event $\mathcal{E}_{\text{in}}$, we have that

$$L(W_{t_{\text{in}}}) \;\leq\; F(W_{t_{\text{in}}}) \;\leq\; \tilde{L}_{NN}.$$

Combining this with $\mathbb{P}(\mathcal{E}_{\text{in}}) \geq 1-\delta$, we conclude that there exists $t_{\text{in}} \leq t_{\max}(\delta)$ such that with probability at least $1-\delta$ it holds that $L(W_{t_{\text{in}}}) \leq \tilde{L}_{NN}$, and the dynamics enter the stable regime $\mathcal{S}_{NN}$ at iteration $t_{\text{in}}$. □

### Proof of Theorem 3.4

*Proof.* The theorem is proved by combining Lemma D.13 and Lemma D.14. □

### D.6 Proofs for the Stochastic Stabilization Mechanism

**Lemma D.15.** Let $\mathcal{E}_t = \{L(W_t) \le \tilde{L}_{NN},\, L(W_{t+1}) > \tilde{L}_{NN}\}$. If at iteration $t > 0$ the dynamics are in the stable regime, then the exit probability satisfies

$$\mathbb{P}(\mathcal{E}_t \mid \mathcal{F}_t) \le 2\exp\left(-\frac{bD_t}{4\eta L(W_t)^{\frac{3}{2}}}\right),$$

where $D_t = \frac{\Delta_t^{NN}}{4\eta(1+\tilde{\beta})\,L(W_t)^{1/2} + \frac{\sqrt{2(1+\tilde{\beta})}(1+n)}{3}\sqrt{\Delta_t^{NN}}}$ and

$$\Delta_t^{NN} = \tilde{L}_{NN} - L(W_t) + 2\eta\big(1 - 8\eta(1+\tilde{\beta})L(W_t)\big)L(W_t)^2 + \frac{\big(1 - 16\eta(1+\tilde{\beta})L(W_t)\big)^2}{8(1+\tilde{\beta})}\,L(W_t).$$

*Proof.* From inequality (196) in the proof of Lemma D.12, we have that

$$L(W_{t+1}) \;\le\; L(W_t) - \tilde{\eta}\,\langle \nabla L(W_t), g_t\rangle + \frac{8(1+\tilde{\beta})\,\tilde{\eta}^2}{m}\,L(W_t)\,\|g_t\|_F^2.$$

Let $\xi_t := g_t - \nabla L(W_t)$ denote the noise at iteration $t$. Substituting $g_t = \nabla L(W_t) + \xi_t$ and expanding, we get

$$\begin{aligned} L(W_{t+1}) \;\le\;& L(W_t) - \tilde{\eta}\,\big(1 - 8\eta\, c\, L(W_t)\big)\,\|\nabla L(W_t)\|_F^2 \\ & - \tilde{\eta}\,\big(1 - 16\eta\, c\, L(W_t)\big)\,\langle \nabla L(W_t), \xi_t\rangle + 8\eta\,\tilde{\eta}\, c\, L(W_t)\,\|\xi_t\|_F^2, \end{aligned} \tag{211}$$

where we have let $c := 1 + \tilde{\beta}$ for brevity.

On the event $\mathcal{E}_t = \{L(W_t) \le \tilde{L}_{NN},\, L(W_{t+1}) > \tilde{L}_{NN}\}$, it holds that

$$\begin{aligned} \tilde{L}_{NN} - L(W_t) + \tilde{\eta}\big(1 - 8\eta c L(W_t)\big)\|\nabla L(W_t)\|_F^2 \;<\;& -\tilde{\eta}\big(1 - 16\eta c L(W_t)\big)\langle \nabla L(W_t), \xi_t\rangle \\ & + 8\eta\tilde{\eta} c L(W_t)\|\xi_t\|_F^2. \end{aligned} \tag{212}$$

We now lower bound the right-hand side by completing the square. Let $A_t := \tilde{\eta}\big(1 - 16\eta c L(W_t)\big)$ and $B_t := 8\eta\tilde{\eta} c L(W_t)$. Then, it holds that

$$\begin{aligned} & -A_t\langle \nabla L(W_t), \xi_t\rangle + B_t\|\xi_t\|_F^2 \\ =\;& \frac{B_t}{2}\|\xi_t\|_F^2 + \left(\frac{B_t}{2}\|\xi_t\|_F^2 - A_t\langle \nabla L(W_t), \xi_t\rangle\right) \\ =\;& \frac{B_t}{2}\|\xi_t\|_F^2 + \frac{B_t}{2}\left\|\xi_t - \frac{A_t}{B_t}\nabla L(W_t)\right\|_F^2 - \frac{A_t^2}{2B_t}\|\nabla L(W_t)\|_F^2 \\ \ge\;& \frac{B_t}{2}\|\xi_t\|_F^2 - \frac{A_t^2}{2B_t}\|\nabla L(W_t)\|_F^2. \end{aligned} \tag{213}$$

Substituting the values of $A_t, B_t$, we obtain

$$\begin{aligned} & -\tilde{\eta}\big(1 - 16\eta c L(W_t)\big)\langle \nabla L(W_t), \xi_t\rangle + 8\eta\tilde{\eta} c L(W_t)\|\xi_t\|_F^2 \\ \ge\;& 4\eta\tilde{\eta} c L(W_t)\|\xi_t\|_F^2 - \frac{\tilde{\eta}\big(1 - 16\eta c L(W_t)\big)^2}{16\eta c L(W_t)}\|\nabla L(W_t)\|_F^2. \end{aligned} \tag{214}$$

Hence, from inequality (212), for the event $\mathcal{E}_t$ to hold, it suffices that

$$\begin{aligned} \tilde{L}_{NN} - L(W_t) + \tilde{\eta}\big(1 - 8\eta c L(W_t)\big)\|\nabla L(W_t)\|_F^2 \;<\;& 4\eta\tilde{\eta} c L(W_t)\|\xi_t\|_F^2 \\ & - \frac{\tilde{\eta}\big(1 - 16\eta c L(W_t)\big)^2}{16\eta c L(W_t)}\|\nabla L(W_t)\|_F^2. \end{aligned} \tag{215}$$

Using Lemma D.3 (125) and the fact that $\tilde{\eta} = m\eta$, it suffices that

$$\tilde{L}_{NN} - L(W_t) + 2\eta\big(1 - 8\eta c L(W_t)\big)L(W_t)^2 + \frac{\big(1 - 16\eta c L(W_t)\big)^2}{8c}\,L(W_t) < 4\eta\tilde{\eta} c L(W_t)\|\xi_t\|_F^2.$$

Letting $\Delta_t^{NN} := \tilde{L}_{NN} - L(W_t) + 2\eta\big(1 - 8\eta c L(W_t)\big) L(W_t)^2 + \frac{\big(1-16\eta c L(W_t)\big)^2}{8c} L(W_t)$, we obtain that for the event $\mathcal{E}_t$ to hold it suffices that

$$\Delta_t^{NN} < 4\eta\tilde{\eta} c L(W_t) \|\xi_t\|_F^2. \tag{216}$$

or equivalently it suffices that

$$\|\xi_t\|_F > \sqrt{\frac{\Delta_t^{NN}}{4\eta\tilde{\eta} c L(W_t)}}. \tag{217}$$

Therefore, it holds that

$$\mathbb{P}(\mathcal{E}_t \mid \mathcal{F}_t) \;\le\; \mathbb{P}\left( \|\xi_t\|_F > \sqrt{\frac{\Delta_t^{NN}}{4\eta\tilde{\eta} c L(W_t)}} \mid \mathcal{F}_t \right). \tag{218}$$

Conditionally on the filtration $\mathcal{F}_t$, the noise $\xi_t = \frac{1}{b}\sum_{j\in B_t} \zeta_{t,j}$ is the average of $b$ centered i.i.d. vectors $\zeta_{t,j} := \nabla \ell_j(W_t) - \nabla L(W_t), \forall j \in B_t$. From Lemma D.3, each term in the sum satisfies the bound

$$\|\zeta_{t,j}\|_F \;\le\; \|\nabla \ell_j(W_t)\|_F + \|\nabla L(W_t)\|_F \;\le\; \sqrt{\tfrac{2}{m}}\,(1+n)\,L(W_t), \tag{219}$$

and the variance bound

$$\mathbb{E}\big[\|\zeta_{t,j}\|_F^2 \mid \mathcal{F}_t\big] \;\le\; \mathbb{E}\big[\|\nabla \ell_{i_j}(W_t)\|_F^2 \mid \mathcal{F}_t\big] \;\le\; \frac{2}{m} L(W_t). \tag{220}$$

By the vector Bernstein inequality applied to $\xi_t$, we have

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_t \mid \mathcal{F}_t) \quad &\le \quad 2\exp\left( - \frac{b\,\Delta_t^{NN}/(4\eta m\eta c L(W_t))}{\frac{4}{m}L(W_t) + \frac{2\sqrt{2}(1+n)}{3\sqrt{m}} L(W_t)\sqrt{\Delta_t^{NN}/(4\eta m\eta c L(W_t))}} \right) \\
&= \quad 2\exp\left( - \frac{b}{4\eta L(W_t)^{\frac{3}{2}}} \cdot \frac{\Delta_t^{NN}}{4\eta c\, L(W_t)^{1/2} + \frac{\sqrt{2c}(1+n)}{3}\sqrt{\Delta_t^{NN}}} \right) \\
&\overset{c=1+\tilde{\beta}}{=} \quad 2\exp\left( - \frac{b}{4\eta L(W_t)^{\frac{3}{2}}} \cdot \frac{\Delta_t^{NN}}{4\eta(1+\tilde{\beta})\, L(W_t)^{1/2} + \frac{\sqrt{2(1+\tilde{\beta})}(1+n)}{3}\sqrt{\Delta_t^{NN}}} \right)
\end{aligned}$$

□

**Lemma D.16.** Assume that the dynamics exit the stable regime $\mathcal{S}_{NN}$ at $t_{\text{out}} > 0$ and fix $\delta \in (0,1)$. Then, with probability at least $1-\delta$ the iterates of (SGD) for the two-layer network in (6) return inside the stable regime $\mathcal{S}_{NN}$ in at most

$$t_{re} = \left\lceil \frac{4}{\tilde{\gamma}^2\eta\delta\tilde{L}_{NN}} \max\left\{ A_{NN}, 16\ln\left(\frac{64}{\tilde{\gamma}^2\eta\delta\tilde{L}_{NN}}\right)\right\}\right\rceil$$

number of steps, where

$$A_{NN} := 3(K-1) + 4\ln^2(\tilde{\gamma}^2\eta t_{\text{out}}) + 20\kappa^2 + 5\eta^2\Big(1+\tfrac{1}{b}\Big)^2.$$

*Proof.* Let $U = U_1 + U_2$ with $U_1 = \alpha \mathbf{W}_*$, $U_2 = \frac{\eta(1+\frac{1}{b})}{\tilde{\gamma}}\mathbf{W}_*$ and $\alpha = \frac{\ln\big(\tilde{\gamma}^2\eta(t-t_{\text{out}})\big)+2\kappa}{\tilde{\gamma}}$. From the update rule of (SGD) with $\tilde{\eta} = m\eta$, we have that

$$\begin{aligned}
\|W_{t+1} - U\|_F^2 &= \|W_t - U\|_F^2 + 2\tilde{\eta}\langle g_t, U - W_t\rangle_F + \tilde{\eta}^2\|g_t\|_F^2 \\
&= \|W_t - U\|_F^2 + 2\tilde{\eta}\langle g_t, U_1 - W_t\rangle_F + \tilde{\eta}^2\left(\frac{2}{\tilde{\eta}}\langle g_t, U_2\rangle_F + \|g_t\|_F^2\right).
\end{aligned}$$

Taking expectation conditional on the filtration $\mathcal{F}_t$ and using the unbiasedness property of the stochastic oracles, we get

$$\begin{aligned}\mathbb{E}\left[\|W_{t+1}-U\|_F^2 \,\middle|\, \mathcal{F}_t\right] &= \|W_t-U\|_F^2 + 2\tilde{\eta}\langle \nabla L(W_t), U_1 - W_t\rangle_F \\ &\quad + \tilde{\eta}^2\left(\frac{2}{\tilde{\eta}}\langle \nabla L(W_t), U_2\rangle_F + \mathbb{E}\left[\|g_t\|_F^2 \,\middle|\, \mathcal{F}_t\right]\right).\end{aligned} \tag{221}$$

We, next, show that with the choice $U_2 = \frac{\eta(1+\frac{1}{b})}{\tilde{\gamma}}\mathbf{W}_*$, the last term in (221) is non-positive. We have that

$$\frac{2}{\tilde{\eta}}\langle \nabla L(W_t), U_2\rangle_F = \frac{2(1+\frac{1}{b})}{m\tilde{\gamma}}\langle \nabla L(W_t), \mathbf{W}_*\rangle_F \overset{\text{Lemma D.2}}{\leq} -\frac{2(1+\frac{1}{b})}{m}G(W_t).$$

For the second moment $\mathbb{E}\left[\|g_t\|_F^2 \,\middle|\, \mathcal{F}_t\right]$, the bias-variance decomposition together with Lemmas D.3 and D.9 give

$$\begin{aligned}\mathbb{E}\left[\|g_t\|_F^2 \,\middle|\, \mathcal{F}_t\right] &= \|\nabla L(W_t)\|_F^2 + \mathbb{E}\left[\|g_t - \nabla L(W_t)\|_F^2 \,\middle|\, \mathcal{F}_t\right] \\ &\overset{\text{Lemma D.3}}{\leq} \frac{2}{m}G(W_t)^2 + \frac{2}{mb}G(W_t) \\ &\leq \frac{2}{m}G(W_t) + \frac{2}{mb}G(W_t) \\ &= \frac{2(1+\frac{1}{b})}{m}G(W_t),\end{aligned}$$

where we used $G(W_t)^2 \leq G(W_t)$ since $G(W_t) \in [0,1]$. Summing the two bounds, we obtain

$$\frac{2}{\tilde{\eta}}\langle \nabla L(W_t), U_2\rangle_F + \mathbb{E}\left[\|g_t\|_F^2 \,\middle|\, \mathcal{F}_t\right] \leq -\frac{2(1+\frac{1}{b})}{m}G(W_t) + \frac{2(1+\frac{1}{b})}{m}G(W_t) = 0. \tag{222}$$

Substituting (222) into (221), we obtain

$$\mathbb{E}\left[\|W_{t+1}-U\|_F^2 \,\middle|\, \mathcal{F}_t\right] \leq \|W_t-U\|_F^2 + 2\tilde{\eta}\langle \nabla L(W_t), U_1 - W_t\rangle_F.$$

Using the convexity of $L$, we have that $\langle \nabla L(W_t), U_1 - W_t\rangle_F \leq L(U_1) - L(W_t)$, and therefore we obtain

$$\mathbb{E}\left[\|W_{t+1}-U\|_F^2 \,\middle|\, \mathcal{F}_t\right] \leq \|W_t-U\|_F^2 + 2\tilde{\eta}\left[L(U_1) - L(W_t)\right].$$

Taking expectation again and using the tower law of expectation, we have that

$$\mathbb{E}\left[\|W_{t+1}-U\|_F^2\right] \leq \mathbb{E}\left[\|W_t-U\|_F^2\right] + 2\tilde{\eta}\,\mathbb{E}[L(U_1) - L(W_t)].$$

Summing for $k = t_{\text{out}}, \ldots, t-1$ and dividing by $2\tilde{\eta}(t-t_{\text{out}})$, we obtain

$$\frac{\mathbb{E}\left[\|W_t-U\|_F^2\right]}{2\tilde{\eta}(t-t_{\text{out}})} + \frac{1}{t-t_{\text{out}}}\sum_{k=t_{\text{out}}}^{t-1}\mathbb{E}[L(W_k)] \leq L(U_1) + \frac{\mathbb{E}\left[\|W_{t_{\text{out}}}-U\|_F^2\right]}{2\tilde{\eta}(t-t_{\text{out}})}.$$

Using the non-negativity of $\|W_t - U\|_F^2$, we get

$$\frac{1}{t-t_{\text{out}}}\sum_{k=t_{\text{out}}}^{t-1}\mathbb{E}[L(W_k)] \leq L(U_1) + \frac{\mathbb{E}\left[\|W_{t_{\text{out}}}-U\|_F^2\right]}{2\tilde{\eta}(t-t_{\text{out}})}. \tag{223}$$

For $U_1 = \alpha\mathbf{W}_*$ with $\alpha = \frac{\ln\left(\tilde{\gamma}^2\eta(t-t_{\text{out}})\right)+2\kappa}{\tilde{\gamma}}$, from Lemma D.7 we have that

$$L(U_1) \leq F(U_1) \leq \frac{K-1}{\tilde{\gamma}^2\eta(t-t_{\text{out}})}. \tag{224}$$

We, next, bound the term $\mathbb{E}\left[\|W_{t_{\text{out}}}-U\|_F^2\right]$. Applying $(a+b)^2 \leq 2a^2 + 2b^2$, we obtain

$$\mathbb{E}\left[\|W_{t_{\text{out}}}-U\|_F^2\right] \leq 2\,\mathbb{E}\left[\|W_{t_{\text{out}}}\|_F^2\right] + 2\|U\|_F^2. \tag{225}$$

For the term $\|U\|_F^2$, similarly to Theorem D.11 we have that

$$\|U\|_F^2 \leq 2\|U_1\|_F^2 + 2\|U_2\|_F^2 \leq \frac{4m\ln^2(\tilde{\gamma}^2\eta(t-t_{\text{out}})) + 16m\kappa^2 + 2m\eta^2\left(1+\frac{1}{b}\right)^2}{\tilde{\gamma}^2}, \tag{226}$$

where we have used $\|\mathbf{W}_*\|_F^2 = m$ and $(a+b)^2 \le 2a^2 + 2b^2$. For the term $\mathbb{E}\big[\|W_{t_{\text{out}}}\|_F^2\big]$, we apply Lemma D.12 from the original time origin with comparator $U^{(0)} = U_1^{(0)} + U_2$ where $U_1^{(0)} = \alpha_0 \mathbf{W}_*$ and $\alpha_0 = \frac{\ln(\tilde{\gamma}^2 \eta t_{\text{out}}) + 2\kappa}{\tilde{\gamma}}$, which gives

$$\begin{aligned}\mathbb{E}\Big[\|W_{t_{\text{out}}} - U^{(0)}\|_F^2\Big] &\le 2\tilde{\eta}\, t_{\text{out}}\, L(U_1^{(0)}) + \|U^{(0)}\|_F^2 \\ &\le \frac{2m(K-1) + 4m\ln^2(\tilde{\gamma}^2\eta t_{\text{out}}) + 16m\kappa^2 + 2m\eta^2\big(1+\frac{1}{b}\big)^2}{\tilde{\gamma}^2}, \end{aligned} \tag{227}$$

where we have used $W_0 = 0$, $\tilde{\eta} = m\eta$, and the bound (224) applied at time $t_{\text{out}}$. Applying $(a+b)^2 \le 2a^2 + 2b^2$ again, we obtain

$$\begin{aligned}\mathbb{E}\big[\|W_{t_{\text{out}}}\|_F^2\big] &\le 2\,\mathbb{E}\Big[\|W_{t_{\text{out}}} - U^{(0)}\|_F^2\Big] + 2\|U^{(0)}\|_F^2 \\ &\le \frac{8m(K-1) + 16m\ln^2(\tilde{\gamma}^2\eta t_{\text{out}}) + 64m\kappa^2 + 8m\eta^2\big(1+\frac{1}{b}\big)^2}{\tilde{\gamma}^2}. \end{aligned} \tag{228}$$

Substituting (224), (226), and (228) into (223) and using $\tilde{\eta} = m\eta$, we obtain

$$\frac{1}{t - t_{\text{out}}} \sum_{k=t_{\text{out}}}^{t-1} \mathbb{E}[L(W_k)] \le \frac{C_{NN}(t, t_{\text{out}})}{\tilde{\gamma}^2 \eta (t - t_{\text{out}})}, \tag{229}$$

where

$$C_{NN}(t, t_{\text{out}}) := 9(K-1) + 16\ln^2(\tilde{\gamma}^2\eta t_{\text{out}}) + 80\kappa^2 + 10\eta^2\big(1+\tfrac{1}{b}\big)^2 + 4\ln^2(\tilde{\gamma}^2\eta(t - t_{\text{out}})).$$

We now split the bound (229) into two terms. Let $A_{NN} := 9(K-1) + 16\ln^2(\tilde{\gamma}^2\eta t_{\text{out}}) + 80\kappa^2 + 10\eta^2\Big(1+\frac{1}{b}\Big)^2$. Then, (229) becomes

$$\frac{1}{t - t_{\text{out}}} \sum_{k=t_{\text{out}}}^{t-1} \mathbb{E}[L(W_k)] \le \frac{A_{NN}}{\tilde{\gamma}^2 \eta (t - t_{\text{out}})} + \frac{4\ln^2(\tilde{\gamma}^2\eta(t - t_{\text{out}}))}{\tilde{\gamma}^2 \eta (t - t_{\text{out}})}. \tag{230}$$

We select $t_{\text{re}}$ so that each of the two terms on the right-hand side of (230) is at most $\delta\tilde{L}_{NN}/4$. Specifically, we next verify the formula for $t_{\text{re}}$:

- For $t - t_{\text{out}} \ge \frac{4A_{NN}}{\tilde{\gamma}^2\eta\delta\tilde{L}_{NN}}$, it holds that
$$\frac{A_{NN}}{\tilde{\gamma}^2 \eta (t - t_{\text{out}})} \le \frac{\delta\tilde{L}_{NN}}{4}.$$
- For $t - t_{\text{out}} \ge \frac{64}{\tilde{\gamma}^2\eta\delta\tilde{L}_{NN}} \ln\Big(\frac{64}{\tilde{\gamma}^2\eta\delta\tilde{L}_{NN}}\Big)$, it holds that $\tilde{\gamma}^2\eta(t - t_{\text{out}}) \ge \frac{64}{\delta\tilde{L}_{NN}} \ln\Big(\frac{64}{\tilde{\gamma}^2\eta\delta\tilde{L}_{NN}}\Big)$, which is a sufficient condition (see Lemma G.5 in Cai et al. (2024)) for
$$\frac{4\ln^2(\tilde{\gamma}^2\eta(t - t_{\text{out}}))}{\tilde{\gamma}^2 \eta (t - t_{\text{out}})} \le \frac{\delta\tilde{L}_{NN}}{4}.$$

Thus, it suffices to select $t_{\text{re}}$ such that

$$t_{\text{re}} = \left\lceil \frac{4}{\tilde{\gamma}^2\eta\delta\tilde{L}_{NN}} \max\left\{A_{NN}, 16\ln\left(\frac{64}{\tilde{\gamma}^2\eta\delta\tilde{L}_{NN}}\right)\right\} \right\rceil. \tag{231}$$

For this choice of $t_{\text{re}}$ and selecting $t = t_{\text{out}} + t_{\text{re}}$, we have that

$$\frac{1}{t_{\text{re}}} \sum_{k=t_{\text{out}}}^{t_{\text{out}}+t_{\text{re}}-1} \mathbb{E}[L(W_k)] \le \frac{\delta\tilde{L}_{NN}}{2}. \tag{232}$$

Hence, there exists $t_{\text{in}}^{[2]} \in [t_{\text{out}}, t_{\text{out}} + t_{\text{re}}]$ such that

$$\mathbb{E}\Big[L(W_{t_{\text{in}}^{[2]}})\Big] \le \frac{\delta\tilde{L}_{NN}}{2}.$$

From Markov's inequality, we have that

$$\Pr\Big(L(W_{t_{\text{in}}^{[2]}}) \geq \tilde{L}_{NN}\Big) \leq \frac{\mathbb{E}\Big[L(W_{t_{\text{in}}^{[2]}})\Big]}{\tilde{L}_{NN}} \leq \frac{\delta}{2} \leq \delta,$$

which implies that

$$\Pr\Big(L(W_{t_{\text{in}}^{[2]}}) \leq \tilde{L}_{NN}\Big) \geq 1 - \delta.$$

Thus, there exists $t_{\text{in}}^{[2]} \leq t_{\text{out}} + t_{\text{re}}$ such that for $\delta \in (0, 1)$ with probability at least $1 - \delta$ it holds that $L(W_{t_{\text{in}}^{[2]}}) \leq \tilde{L}_{NN}$ and the dynamics return inside the stable set $\mathcal{S}_{NN}$. □

PROOF OF THEOREM 3.5

*Proof.* The theorem is proved by combining Lemma D.15 and Lemma D.16. □