

On the Loss Spike in Training Linear Models with Normalization Layers

Peifeng Gao¹, Wenyi Fang², Yang Zheng², Difan Zou¹

¹School of Computing and Data Science, The University of Hong Kong .

²RAMS Technologies Lab, Huawei Technologies Ltd.

Contributing authors: gaopeifeng@connect.hku.hk;
fangwenyi3@huawei.com; zhengyang31@huawei.com; dzou@cs.hku.hk;

Abstract

Loss spikes during neural network training are a widely observed phenomenon, yet their underlying mechanisms remain incompletely understood. While empirical studies report that such spikes predominantly emerge at intermediate training stages with relatively low loss values, existing theoretical analyses attribute spikes to large learning rates (exceeding the inverse of the loss function’s smoothness) and predict their occurrence during initial training phases—a discrepancy with practical observations. We conjecture that the key ingredient that leads to the delayed spike is the normalization layer in the neural network models, as its learning rate auto-tuning mechanism (Arora et al. 2018) can enable stable early training before precipitating delayed instability. To this end, we study the loss spike in arguably the simplest theoretical setting – linear models with batch normalization layer. We analyze two classical problems: logistic regression and linear regression, and develop a comprehensive analysis on the training dynamics of gradient descent. Specifically, in order to precisely characterize the properties of the loss spike, we (1) develop the conditions for its occurrence, and (2) characterize their duration and peak values. Crucially, our theoretical results reveal that as the training loss stabilizes at low values, the effective learning rate (implicitly modulated by normalization layers) experiences gradual amplification, consequently triggering a sudden loss spike when surpassing a critical threshold. Moreover, we also prove that the loss spike will not yield a gradient explosion, but will stabilize within a small number of iterations. Our findings offer new insights into the occurrence of loss spikes in neural network training and align more closely with empirical observations.

Keywords: Loss Spikes, Normalization Method, Linear Model, Large Learning Rate

1 Introduction

Pretrained Large Language Models (LLMs) have demonstrated extraordinary performance across a wide range of tasks. However, their pretraining process faces significant challenges, including instability and divergence (Zeng et al. 2022; Chowdhery et al. 2023; Molybog et al. 2023). During pretraining, loss spikes, sharp increases in the loss, can occur, forcing the training process to require additional iterations to recover the original loss level. Additionally, in some cases, the recovery may not be achieved, the loss spikes will potentially result in gradient explosion leading to training collapse and necessitating a restart from an earlier checkpoint. Given these immense additional computational cost of pretraining LLMs, it is crucial to investigate the mechanisms underlying such loss divergence.

The Edge of Stability (EOS) is a promising research direction for understanding training instability, offering valuable insights into the non-monotonic behaviors observed during training. In the EOS regime, training loss exhibits non-monotonic fluctuations while maintaining a decreasing trend over long timescales. This behavior seems very similar to the instability and divergence often seen during LLM pretraining. Existing EOS studies have primarily focused on specific, simplified functions, such as $(u, v) \mapsto (u^2v^2 - 1)^2$ (Zhu et al. 2022) and $[x_1, \dots, x_n] \mapsto (x_1 \times \dots \times x_n - 1)^2$ (Kreiser et al. 2023). Other works investigate EOS using more general assumptions, such as the existence of a forward-invariant subset near the minima of the objective function (Ahn et al. 2022) or additional restrictive properties of the model (Ma et al. 2022). However, these studies fails to explain the loss spike phenomenon observed in LLM pertaining. The specific scalar functions of (Zhu et al. 2022; Kreiser et al. 2023) analyzed are difficult to relate to practical models, and the required assumptions (Ahn et al. 2022; Ma et al. 2022) are often unverifiable in real-world scenarios.

A series of more related existing works lies in studying the training dynamics of gradient descent when using large learning rates, i.e., the ones exceeding the inverse of the smoothness. From a theoretical perspective, these works (Wu et al. 2024b; Andriushchenko et al. 2023; Wu et al. 2024a; Lu et al. 2023) showed that for some simple linear or neural network models, the stable (or even faster) convergence can still be achieved. However, we emphasize that the problem instances and theoretical explanations provided in these prior works do not adequately account for the loss spikes observed in practice. In practical scenarios, loss spikes may occur in the whole process of training, especially when the loss has already decreased to a relatively low level (as illustrated in Figure 1 of Takase et al. (2023) and Figure 1 of Molybog et al. (2023)). In contrast, existing studies primarily focus on convergence with large learning rates, often leading to loss spikes at the beginning of training (Wu et al. 2024b). More importantly, existing studies fail to identify the conditions that trigger loss spikes or characterize their shape. Consequently, there exists a lack of understanding of the spike mechanism to help us effectively predict or avoid spikes, which highlights significant gaps between existing research and the loss spike behaviors observed in practice.

To bridge this gap, we are motivated by the observation that empirical findings regarding loss spikes are predominantly discovered during the training of neural networks that employ normalization layers, such as weight normalization (Salimans and Kingma 2016), batch normalization (Ioffe 2015), and layer normalization (Ba

2016). Additionally, existing research has shown that these normalization methods can exhibit certain learning rate auto-tuning mechanisms (Morwani and Ramaswamy 2022; Arora et al. 2018; Hoffer et al. 2018), i.e., adaptively adjusting the effective learning rate during training. Therefore, we conjecture that normalization layers could be a key factor contributing to the loss spikes that occur after many iterations.

To this end, we study the loss spikes in arguably the simplest theoretical setting – linear models with batch normalization layer. We consider full-batch gradient descent and develop a comprehensive characterization on the training dynamics for both linear regression and logistic regression problems. We highlight main contributions of this work as follows:

- We demonstrate that the batch normalization layer can induce loss spikes by causing directional divergence, which refers to the deviation of the parameters from the direction in which they should have converged to reduce the loss along the gradient descent path. We then prove the conditions under which batch normalization leads to convergence or divergence, emphasizing that these conditions are problem-independent and applicable to other models utilizing normalization layers.
- We observe that the loss spike occurs when the effective learning rate surpasses a specific threshold and that the spike can always recover stability since a negative feedback mechanism is activated during *Rising Edge* of the loss spike, which counteracts the instability and restores equilibrium. Specifically, in linear regression, the effective learning rate acts as an adaptive parameter that controls the convergence and divergence of the training process (see Section 5.2). When the effective learning rate surpasses a specific threshold, the system enters a state called the *Rising Edge*, where a spike in loss is observed. In this state, there exists a negative feedback mechanism reducing the effective learning rate, allowing the system to stabilize and transition into the *Falling Edge* of the spike. An analogous mechanism is observed in logistic regression, where an adaptive process regulates the convergence and divergence behavior in a comparable way (see Section 5.3).
- Our theory explains the phenomenon that the loss spikes occur when the training loss has been reduced to a relatively low level. We provide a clear characterization on the interplay between the learning rate, conditions of the spike, and the training loss objective, which shows that the condition of the spike can be satisfied as the training objective becomes smaller. The developed technical tools can be potentially leveraged to characterize the spike phenomena in a broader class of scenarios.

1.1 MLJ Contribution Information Sheet

- **What is the main claim of the paper?** Please refer to the last subsection for the main claim of the paper.
- **What is the evidence you provide to support your claim?** Please refer to the last subsection for the evidence of our claim.
- **What papers by other authors make the most closely related contributions, and how is your paper related to them?** Papers of other authors are not related to the contributions of this paper.

- Have you published parts of your paper before, for instance in a conference? No.

2 Related Work

Theoretical Studies on Normalization. Modern deep models extensively incorporate normalization layers, such as batch normalization (Ioffe 2015) and layer normalization (Ba 2016), to enhance generalization performance. Several studies have shown that batch normalization can dynamically influence the effective learning rate (Morwani and Ramaswamy 2022; Arora et al. 2018; Hoffer et al. 2018). In particular, Arora et al. (2018) demonstrates that batch-normalized parameters can converge to a stationary point even under arbitrarily large learning rates. We emphasize that our results and theirs fundamentally characterize the effects of batch normalization at different scales. While their study examines its long-term convergence behavior, our focus is on its short-term dynamics, specifically in the moments just before and after the occurrence of a loss spike.

Additionally, Cao et al. (2023) shares the same problem setting as ours. They investigate the implicit bias of batch normalization in logistic regression and demonstrate that gradient descent converges in the direction that minimizes the uniform margin. Their analysis follows a two-stage approach, where the first stage acts as a warm start, facilitating the smooth convergence of the second stage. We emphasize that our work focuses on the dynamic behavior in the first stage, during which smooth convergence is not guaranteed and loss spikes may occur.

Edge of Stability. Edge of Stability (EoS) (Cohen et al. 2021) is a phenomenon in deep learning optimization where gradient descent and its variants achieve stable convergence even when the learning rate exceeds the classical threshold $2/\lambda_{\max}$, where λ_{\max} denotes the largest eigenvalue of the loss Hessian. Contrary to traditional theory predicting divergence beyond this threshold, empirical studies indicate that optimization instead settles into a dynamic equilibrium, where loss oscillates periodically yet maintains overall convergence (Cohen et al. 2021).

To investigate its underlying mechanisms, most subsequent studies focus on specific models (Zhu et al. 2022; Chen and Bruna 2022; Ahn et al. 2023; Even et al. 2023) and functions (Ma et al. 2022; Ahn et al. 2022; Damian et al. 2022; Wang et al. 2022). However, significant gaps remain between these studies and practical deep models, as the functions examined often deviate substantially from models in the modern deep learning paradigm, or the assumptions imposed on these models are difficult to verify.

Recently, Wu et al. (2024b) investigated EoS in the context of logistic regression, demonstrating that logistic regression can converge to the max-margin solution under any constant learning rate. They observed that training losses oscillate when the learning rate is large, a behavior that appears similar to loss spikes. However, we emphasize that this phenomenon differs significantly from loss spikes. As shown in Figure 1 of (Wu et al. 2024b), loss oscillations occur primarily at the beginning of training, whereas loss spikes, as reported in (Molybog et al. 2023), typically emerge when the loss is relatively low. To bridge this gap, our study focuses on logistic regression with batch normalization.

3 Problem Setup

Since this work focuses on both linear regression and logistic regression problems, we introduce a unified set of notations. Let n denote the number of samples and d the feature dimension.

Dataset. The training dataset is represented as $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the input feature vector and $y_i \in \{1, -1\}$ is the corresponding label. Now, we define the following notations:

$$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}; \mathbf{y} = [y_1, \dots, y_n] \in \mathbb{R}^{n \times 1};$$

$$\tilde{\mathbf{X}} = \mathbf{X} \text{diag}(\mathbf{y}); \mathbf{\Sigma} = \frac{1}{n} \tilde{\mathbf{X}} \tilde{\mathbf{X}}^T, \mathbf{\mu} = \frac{1}{n} \tilde{\mathbf{X}} \mathbf{1}_n,$$

where $\mathbf{1}_n$ is n -dimension all one vector. Additionally, we introduce the inner product and norm in the $\mathbf{\Sigma}$ inner product space: for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, we define

$$\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{\Sigma}} = \mathbf{a}^T \mathbf{\Sigma} \mathbf{b}; \|\mathbf{a}\|_{\mathbf{\Sigma}}^2 = \mathbf{a}^T \mathbf{\Sigma} \mathbf{a}.$$

Linear Model and Risk. Then, we introduce the linear model with batch normalization in the same form as presented in [Cao et al. \(2023\)](#):

$$\text{logit}(\mathbf{x}_i; \mathbf{w}, \alpha) = \alpha \cdot \frac{\langle \mathbf{x}_i, \mathbf{w} \rangle}{\|\mathbf{w}\|_{\mathbf{\Sigma}}}, \mathbf{w} \in \mathbb{R}^d, \alpha \in \mathbb{R},$$

where \mathbf{w} is the linear model parameter and α is scaling factor of batch normalization. Different from [Cao et al. \(2023\)](#), this paper will frequently apply matrix representation in analysis. For logit of every sample, we can use the following representation:

$$\left[\text{logit}(\mathbf{x}_1; \mathbf{w}, \alpha), \dots, \text{logit}(\mathbf{x}_n; \mathbf{w}, \alpha) \right]^T = \alpha \cdot \frac{\mathbf{X}^T \mathbf{w}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}}$$

For a loss function $\ell(\cdot)$, we define $\ell \in \{\ell_{\log}, \ell_{squ}\}$, where $\ell_{\log}(\cdot) := \log(1 + \exp(-(\cdot)))$. Regarding ℓ_{squ} , since our labels are defined as $\{1, -1\}$, we set $\ell_{squ}(\cdot) := (1 - (\cdot))^2/2$ to establish a unified notation applicable to both linear regression and logistic regression. Additionally, we use boldface $\ell(\cdot)$ to denote the element-wise version of the loss function. Next, we introduce the empirical risk:

$$\mathcal{R}(\mathbf{w}, \alpha) = \frac{1}{n} \mathbf{1}_n^T \ell \left(\alpha \cdot \frac{\tilde{\mathbf{X}}^T \mathbf{w}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}} \right),$$

Gradient Descent. This paper considers gradient descent, then we introduce the gradient:

$$\nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}, \alpha) = \frac{\alpha}{n \|\mathbf{w}\|_{\Sigma}} \left(\mathbf{I} - \frac{\Sigma \mathbf{w} \mathbf{w}^T}{\|\mathbf{w}\|_{\Sigma}^2} \right) \tilde{\mathbf{X}} \ell' \left(\frac{\alpha \tilde{\mathbf{X}}^T \mathbf{w}}{\|\mathbf{w}\|_{\Sigma}} \right);$$

$$\frac{\partial \mathcal{R}}{\partial \alpha}(\mathbf{w}, \alpha) = \frac{1}{n} \left(\frac{\tilde{\mathbf{X}}^T \mathbf{w}}{\|\mathbf{w}\|_{\Sigma}} \right)^T \ell' \left(\frac{\alpha \tilde{\mathbf{X}}^T \mathbf{w}}{\|\mathbf{w}\|_{\Sigma}} \right),$$

where we use bold ell ℓ' to denote the vector element-wise ℓ derivative function. We study a sequence of parameter (\mathbf{w}_t, α_t) produced by the following gradient descent with learning rate η, η_{α} :

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta \nabla_{\mathbf{w}} \mathcal{R}_t; \quad \alpha_{t+1} \leftarrow \alpha_t - \eta_{\alpha} \frac{\partial \mathcal{R}_t}{\partial \alpha}, \quad (1)$$

starting from initial parameter (\mathbf{w}_0, α_0) , where we use \mathcal{R}_t to denote $\mathcal{R}(\mathbf{w}_t, \alpha_t)$ to simplify notations. Note that gradient of \mathbf{w} is always in $\text{span}(\mathbf{X}) := \text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, which means we can only focus on the component of \mathbf{w} in $\text{span}(\mathbf{X})$. Therefore, we specify that the \mathbf{w}_0 are in \mathcal{X} . Besides, without loss of generality, we specify $\alpha_0 > 0$.

4 Main Results

We identify that the loss spike arises due to *the sharp directional divergence*. The “direction” means the direction that parameter should converge to in gradient descent to reduce the loss. We denote $\hat{\mathbf{w}}$ as the reference direction. For linear regression, this is the direction of the least squares solution, and for logistic regression, it corresponds to the SVM solution (Gunasekar et al. 2018). Roughly speaking, a decreasing loss value indicates that \mathbf{w}_t is converging to $\hat{\mathbf{w}}$, provided that the scaling factor α_t does not change dramatically. Conversely, a sharp divergence of \mathbf{w}_t from $\hat{\mathbf{w}}$ may also lead to a sudden increase in losses, that is, loss spikes. Before we present our results, we introduce notations ρ_t and ρ_t^{\perp} to describe how close the direction of \mathbf{w}_t is to the reference direction $\hat{\mathbf{w}}$:

$$\rho_t := \rho(\mathbf{w}_t) := \frac{\langle \hat{\mathbf{w}}, \mathbf{w}_t \rangle}{\|\mathbf{w}_t\|}; \quad \rho_t^{\perp} := \rho^{\perp}(\mathbf{w}_t) := \left\| \hat{\mathbf{w}} - \rho_t \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|} \right\|.$$

The left of Figure 1 clearly show that $\rho(\mathbf{w})$ is the length of the component of $\hat{\mathbf{w}}$ in the direction of \mathbf{w} and $\rho^{\perp}(\mathbf{w})$ is the length of the orthogonal component of $\hat{\mathbf{w}}$ with respect to \mathbf{w} . By Pythagorean theorem, their square summation is always equal to $\|\hat{\mathbf{w}}\|^2$, that is $\forall \mathbf{w}, (\rho^{\perp}(\mathbf{w}))^2 + \rho(\mathbf{w})^2 = \|\hat{\mathbf{w}}\|^2$. We note that ρ_t^{\perp}/ρ_t is, in fact, the tangent of $\angle(\mathbf{w}_t, \hat{\mathbf{w}})$, an important metric used to measure the closeness between \mathbf{w}_t and $\hat{\mathbf{w}}$ in subsequent results. The convergence of ρ_t^{\perp}/ρ_t to 0 indicates that \mathbf{w}_t

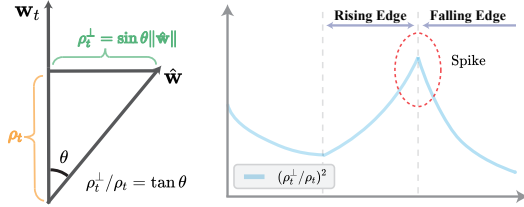


Fig. 1: Left: Use ρ_t^{\perp} and ρ_t^{\perp}/ρ_t to measure the directional deviation between $\hat{\mathbf{w}}$ and \mathbf{w}_t . Right: Here is a spike of ρ_t^{\perp}/ρ_t . We define *Rising Edge* and *Falling Edge* states based on the monotonicity trend of ρ_t^{\perp}/ρ_t .

gradually becomes collinear with $\hat{\mathbf{w}}$, leading to a reduction in loss. In this section, we first formally prove that the sharp increase in ρ_t^\perp and ρ_t^\perp/ρ_t can lead to loss spike in both linear and logistic regression settings, then provide training dynamics to see how ρ_t^\perp and ρ_t^\perp/ρ_t diverges during gradient descent. For the sake of explanation, we define two system states according to the increase and decrease of ρ_t^\perp/ρ_t , as shown in right of Figure 1:

- *Rising Edge*: refer to the time span when ρ_t^\perp/ρ_t is increasing;
- *Falling Edge*: refer to the time span when ρ_t^\perp/ρ_t is decreasing.

In our work, we aim to provide a detailed characterization over the shape of the spike, including the condition that triggers the spike, the durations of the rising edge and falling edge.

4.1 Linear Regression

In this part, we analyze the dynamics of linear regression and explore the reasons behind loss spikes. This serves as a warm-up to provide insights for analyzing the more complex case of logistic regression. To simplify our analysis, we consider the whitened data, i.e., the data is preprocessed via linear operations such that the empirical covariance matrix satisfies $\Sigma = \mathbf{I}$. Then, we deliver the following lemma that decomposes the risk into the directional divergence terms.

Lemma 1 (Decomposition of Mean Square Loss) *Let ℓ be ℓ_{squ} and let $\hat{\mathbf{w}}$ be the least squares solution, that is, $\hat{\mathbf{w}} := \Sigma^{-1}\boldsymbol{\mu}$. Suppose $\Sigma = \mathbf{I}$. Then, the following holds:*

$$\inf_{\mathbf{w}, \alpha} \mathcal{R}(\mathbf{w}, \alpha) = 1 - \|\hat{\mathbf{w}}\|^2; \quad \mathcal{R}_t = (\alpha_t - \rho_t)^2 + (\rho_t^\perp)^2 + 1 - \|\hat{\mathbf{w}}\|^2.$$

This lemma demonstrates that a sharp rise in $(\rho_t^\perp)^2$ can trigger a loss spike. Consequently, we can investigate the occurrence of such spikes by analyzing the growth of ρ_t^\perp (or ρ_t^\perp/ρ_t). Next, we focus on the small-loss regime, where the training loss (or ρ_t^\perp/ρ_t) remains relatively low, and present the following theorem, which characterizes the conditions that trigger the spike of ρ_t^\perp/ρ_t and time of spike occurrence.

Theorem 2 (Condition of Spike) *Let $\ell = \ell_{squ}$ be the square loss and $\hat{\mathbf{w}} = \Sigma^{-1}\boldsymbol{\mu}$. Suppose $\Sigma = \mathbf{I}$. Consider the gradient descent (1) for $t > t_0$ with $\eta_\alpha \in (0, 1)$, where t_0 is such that $\rho_{t_0}^\perp/\rho_{t_0} \leq 1/\sqrt{3}$ and $0 < \alpha_{t_0} < \rho_{t_0}$. The following results hold:*

1. **Condition of No Spike.** *If $\frac{\eta}{\|\mathbf{w}_{t_0}\|^2} < \frac{2}{\|\hat{\mathbf{w}}\|^2}$, there shall be no spikes for any $t \geq t_0$.*
2. **Condition of Spike.** *There exists a constant C such that if η satisfies $\frac{8}{\|\hat{\mathbf{w}}\|^2} < \frac{\eta}{\|\mathbf{w}_{t_0}\|^2} \leq C$, then a spike will occur within at most $\Delta T_0 = \Theta\left(\ln\left(\frac{\eta_\alpha}{\eta\|\hat{\mathbf{w}}\|^2(\rho_{t_0}^\perp/\rho_{t_0})^2}\right)/\eta_\alpha\right)$ iterations. Formally speaking, there exists $t_1 \in (t_0, t_0 + \Delta T_0]$ such that*

$$\frac{\rho_{t+1}^\perp}{\rho_{t+1}} \leq \frac{\rho_t^\perp}{\rho_t} \quad \forall t \in [t_0, t_1) \quad \text{and} \quad \frac{\rho_{t_1+1}^\perp}{\rho_{t_1+1}} \geq \frac{\rho_{t_1}^\perp}{\rho_{t_1}}.$$

The above theorem characterizes the occurrence of the spike. Specifically, we show that

- the spike occurs when the effective learning rate $\eta/\|\mathbf{w}_{t_0}\|^2$ is larger than a certain threshold, otherwise the training loss will converge smoothly and the spike will not occur. This implies that a spike is more likely to be triggered if the learning rate η is large and the norm of the weight \mathbf{w}_t is small.
- when the effective learning rate reaches a certain value, a spike will occur after at most ΔT_1 iterations, which scales as $-\log(\rho_{t_0}^\perp/\rho_{t_0})$, i.e., this period will be longer if the training loss is smaller.

Given the spike conditions detailed in Theorem 2, we then shift our focus to characterizing the properties of the spike itself. Specifically, we provide the following theorem which characterizes the entire shape of the spike for the iterations $t \geq t_1$.

Theorem 3 (Shape of the Loss Spike) *Following the second setting in Theorem 2 holds, then starting from the spike at time t_1 , the Rising Edge will last for at most*

$$\Delta T_1 = \left\lceil \frac{1}{4} \frac{\|\hat{\mathbf{w}}\|^4}{\alpha_{t_1}^2} \left(1/(\rho_{t_1}^\perp)^2 - 1/\rho_{t_1}^2\right)^2 \right\rceil + \left\lceil \frac{1}{4} \frac{\|\hat{\mathbf{w}}\|^2}{\rho_{t_1}^2} \left(\rho_{t_1}/\rho_{t_1}^\perp - \rho_{t_1}^\perp/\rho_{t_1}\right)^2 \right\rceil$$

iterations and turn to the Falling Edge. Specifically, there exists a $t_2 \in (t_1, t_1 + \Delta T_1]$ such that

$$\frac{\rho_{t+1}^\perp}{\rho_{t+1}} \geq \frac{\rho_t^\perp}{\rho_t} \quad \forall t \in [t_1, t_2) \quad \text{and} \quad \frac{\rho_{t_2+1}^\perp}{\rho_{t_2+1}} \leq \frac{\rho_{t_2}^\perp}{\rho_{t_2}},$$

Moreover, define a time $\phi \in [t_1, t_2]$ as the first iteration when α_t catches up with ρ_t , i.e., the time such that $\alpha_t \leq \rho_t \quad \forall t \in [t_1, \phi]$ and $\alpha_t \geq \rho_t \quad \forall t \in (\phi, t_2)$. Then, it holds that:

$$\forall t \in [t_1, \phi], \quad (\rho_t^\perp/\rho_t)^2 \leq 1 - \frac{2\rho_{t_1}^\perp \alpha_{t_1}}{\|\hat{\mathbf{w}}\|^2} \sqrt{t - t_1}; \quad \forall t > \phi, \quad (\rho_t^\perp/\rho_t)^2 \leq 1 - \frac{2\rho_{t_1}^\perp}{\|\hat{\mathbf{w}}\|} \sqrt{t - \phi}.$$

Theorem 3 establishes that the loss spike's rising phase terminates after a finite number of iterations, transitioning into a falling phase. This confirms that the growth of ρ_t^\perp remains bounded and does not diverge. Specifically, we prove that the loss spike stabilizes within at most ΔT_1 iterations, where ΔT_1 is a function of ρ_{t_1} and $\rho_{t_1}^\perp$. Moreover, the peak value of ρ_t^\perp/ρ_t is guaranteed not to exceed 1.

As we will discuss in Section 5.2, this stabilization occurs because directional divergence triggers a negative feedback adjustment induced by batch normalization. This adjustment causes the effective learning rate to rapidly decline during the *Rising Edge* of the spike, which further prevents uncontrolled growth. Figure 2 visually illustrates this mechanism, showing how batch normalization helps maintain control over the training dynamics even in the presence of loss spikes. This finding aligns with the study by Arora et al. (2018), which highlights the learning rate auto-tuning effect of batch normalization. We note that prior work by Wu et al. (2024b) did not consider the normalization layer, resulting in loss spikes being observed only in the early stages of training when a large learning rate is employed.

4.2 Logistic Regression

In this part, we study loss spikes in the logistic regression problem. In our analysis, we follow [Cao et al. \(2023\)](#) by considering the model to be overparameterized, i.e., the number of features is greater than the number of training examples.

Assumption 1 (Overparameterization) *Assume $n < d$ and $\text{rank}\{\mathbf{X}\} = n$.*

We now introduce the Support Vector Machine (SVM) solution, which serves as the reference convergence direction of \mathbf{w}_t in the gradient descent dynamics.

Definition 1 (Support Vector Machine solution) *Let $\hat{\mathbf{w}}$ be the SVM solution and define margin as $\gamma := 1/\|\hat{\mathbf{w}}\|$:*

$$\hat{\mathbf{w}} := \arg \min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{w}\|_2, \text{ s.t. } y_i \cdot \langle \mathbf{x}_i, \mathbf{w} \rangle \geq 1, i \in [n].$$

Next, the following lemma upper bounds the risk by direction and length of parameter.

Lemma 4 (Upper Bound of Logistic Loss) *Let ℓ be ℓ_{\log} and $\hat{\mathbf{w}}$ be the SVM solution as presented in Definition 1. Suppose Assumption 1 holds. Consider the gradient descent (1), for all $t \geq 0$, if $\rho_t > 0$ and $\alpha_t > 0$, it holds that*

$$\mathcal{R}_t \leq \ell(\alpha_t) + \alpha_t \left| \ell' \left(\left[1 - C_0 \gamma \cdot \rho_t^\perp \right] \alpha_t \right) \right| \cdot C_0 \gamma \cdot \rho_t^\perp,$$

where C_0 is a data-depnt constant.

Lemma 4 demonstrates that a smaller ρ_t^\perp corresponds to a smaller loss. In addition, if the dataset is not degenerated, that is, $\hat{\mathbf{w}}$ can be represented as the strictly positive linear combination of support vector (See Assumption 2 in Appendix), we can lower bound the risk by ρ_t^\perp , as demonstrated below. It is worth mentioning that this assumption is mild. It holds almost surely for all linearly separable dataset that is sampled from continuous distributions (see Appendix B of [Soudry et al. \(2018\)](#)).

Lemma 5 (Lower Bound of Logistic Loss) *Let ℓ be ℓ_{\log} and $\hat{\mathbf{w}}$ be the SVM solution as presented in Definition 1. Suppose Assumption 1 and 2 hold. Consider the gradient descent (1), for all $t \geq 0$, if $\alpha_t > 0$, it holds that*

$$\mathcal{R}_t \geq \frac{1}{n} \ell \left(\frac{\alpha_t}{\sqrt{\lambda_{\min}}} \left(\rho_t \gamma^2 - \rho_t^\perp b \gamma \right) \right),$$

where $b > 0$ is the margin offset (see Definition 2 in Appendix).

By Lemma 5, a sharp increase in ρ_t^\perp may result in a loss spike. We then investigate the loss spike in logistic regression by studying ρ_t^\perp . In contrast to the analysis for linear regression, which characterizes the training behavior before and after a specific spike, the following theorem describes the overall trend of ρ_t^\perp over the time period $[0, T_0)$.

Theorem 6 Let ℓ be ℓ_{\log} and $\hat{\mathbf{w}}$ be the SVM solution as defined in Definition 1. Suppose Assumption 1 holds and $\lambda_{\max} \geq 1$. Consider the gradient descent (1). If $\alpha_0 \leq \frac{1}{3} \log(\lambda_{\max})$, $\Theta(\gamma) \leq \frac{\eta}{\|\mathbf{w}_0\|^2} \leq \Theta(\gamma^{-1})$ and $\eta_\alpha \leq \Theta\left(\frac{\|\mathbf{w}_0\|^2}{\gamma\eta}\right)$, then let $\Phi = \Theta\left(\frac{\eta\gamma}{\|\mathbf{w}_0\|^2}\right)$ and $T_0 = \Theta\left(\left(1 + \frac{1}{\tan_{\min}^2}\right) \frac{\eta\gamma}{\|\mathbf{w}_0\|^2}\right)$, it holds that during $t \in [0, T_0)$

1. **Monotonic Decrease of ρ_t^\perp :** ρ_t^\perp keeps decreasing as long as $(\rho_t^\perp/\rho_t)^2 \geq 4/(\sqrt{\Phi^2 + 4} - \Phi)^2 - 1$ and $\rho_t > 0$;
2. **Occurrence of Spike:** there exists a t_0 such that $(\rho_{t_0}^\perp/\rho_{t_0})^2 \leq \tan_{\min}^2 = \frac{\gamma^2 \lambda_{\min}}{8\lambda_{\max}^2}$, and a spike occurs at t_0 ;
3. **Peak of Rising Edge:** when it is in Rising Edge of a spike, the peak of $(\rho_t^\perp/\rho_t)^2$ is at most $C_6 \cdot \frac{\eta^2}{\|\mathbf{w}_0\|^4} - 1$ for some absolute constant C_6 .

From Theorem 6, some key observations can be summarized as follows:

- The loss spike occurs when the training loss falls below a threshold determined by the learning rate and initialization parameters. Specifically, with a smaller learning rate η and larger initialization norm $\|\mathbf{w}_0\|$, we will have a larger Φ , which in turn reduces the quantity $4/(\sqrt{\Phi^2 + 4} - \Phi)^2 - 1$. Consequently, this pushes the spike to occur a smaller training loss.
- Moreover, the occurrence of a loss spike depends on both the margin and the condition number of the data covariance matrix. As indicated by the second part of Theorem 6, when the margin γ decreases and the condition number $\lambda_{\max}/\lambda_{\min}$ increases, the spike tends to occur in regimes with smaller training loss. This suggests that training becomes less stable under such conditions.

The theoretical results for logistic regression (Theorem 6) and linear regression (Theorem 3) share several key similarities. Both theorems demonstrate that loss spikes can occur when the training loss is relatively small, revealing the learning rate auto-tuning effect in batch-normalized models. Besides, there are also distinctions between these results. The linear regression analysis primary focuses on characterizing the precise shape of individual spikes, while the result of logistic regression describes the overall behavior of loss spikes over an extended time period. Moreover, as the third part of Theorem 6 shown, the spike peak in logistic regression depends on $\eta/\|\mathbf{w}_t\|^2$, while for linear regression, the peak values are bounded by a constant-order quantity.

5 Proof Overview

We first discuss the mechanism by which batch normalization causes directional convergence in gradient descent at a high level. Then, we consider specific cases of linear regression and logistic regression to examine how the directional divergence of the parameter \mathbf{w}_t can lead to loss spikes.

5.1 Directional Convergence Induced by BN

For now, we do not specify the training objective and task, but instead focus on the dynamics of gradient descent in normalized models. We present the following lemma that describes how normalized models converge to and diverge from the reference direction during gradient descent.

Lemma 7 (Directional Convergence and Divergence) *Suppose there exists a reference direction $\hat{\mathbf{w}}$. Consider gradient descent (1) on an objective function $\mathcal{R}(\mathbf{w}, \alpha)$, where \mathbf{w} is parameterized by normalization and α represents the scaling factor of the normalization. We have the following direction convergence condition: if there exists a $t \geq 0$ such that*

$$\rho_t > 0 \text{ and } \frac{\eta \rho_t}{\|\mathbf{w}_t\|} \|\nabla_{\mathbf{w}} \mathcal{R}_t\|^2 \leq -2 \langle \hat{\mathbf{w}}, \nabla_{\mathbf{w}} \mathcal{R}_t \rangle,$$

it holds that $(\rho_{t+1}^\perp)^2 \leq (\rho_t^\perp)^2$, and the following direction divergence condition: if there exists a $t \geq 0$ such that

$$0 < \rho_t^\perp / \rho_t \leq 1, \quad \alpha_t > 0 \text{ and } \frac{\eta}{\|\mathbf{w}_t\|} \|\nabla_{\mathbf{w}} \mathcal{R}_t\| \geq \frac{2\rho_t \rho_t^\perp}{\rho_t^2 - (\rho_t^\perp)^2},$$

it holds that $(\rho_{t+1}^\perp)^2 \geq (\rho_t^\perp)^2$.

Remark 1 Recall the definition of ρ_t^\perp , the smaller ρ_t^\perp is, the closer \mathbf{w}_t 's direction is to $\hat{\mathbf{w}}$'s. Therefore, this lemma describes the conditions under which \mathbf{w}_t converges to and diverges from the reference direction $\hat{\mathbf{w}}$ during gradient descent in a normalized model. It is worth noting that this lemma applies to any normalization model in gradient descent and is not just restricted to batch normalization. Moreover, it does not depend on the specific training objective and the reference direction.

Lemma 7 is one of our main contributions. It characterizes the instability induced by normalization layers. In Sections 5.2 and 5.3, this lemma serves as a foundation for explaining how directional divergence occurs and leads to loss spikes, specifically when the loss function is at a relatively low level.

5.2 Linear Regression

In the gradient descent of linear regression, an elegant relationship between the gradient of the square loss and ρ_t and ρ_t^\perp can be established through careful reformulation.

$$\nabla_{\mathbf{w}} \mathcal{R}_t = -\frac{\alpha_t}{\|\mathbf{w}_t\|} \left(\mathbf{I} - \frac{\mathbf{w} \mathbf{w}^T}{\|\mathbf{w}\|^2} \right) \hat{\mathbf{w}}; \quad \frac{\partial \mathcal{R}_t}{\partial \alpha} = -(\rho_t - \alpha_t).$$

We note that $\nabla_{\mathbf{w}} \mathcal{R}_t$ is collinear with the projection of $\hat{\mathbf{w}}$ onto $\text{span}^\perp\{\mathbf{w}_t\}$, a remarkable property that is highly beneficial for our analysis. By combining this with Lemma 7, we can precisely characterize the training dynamics of linear regression.

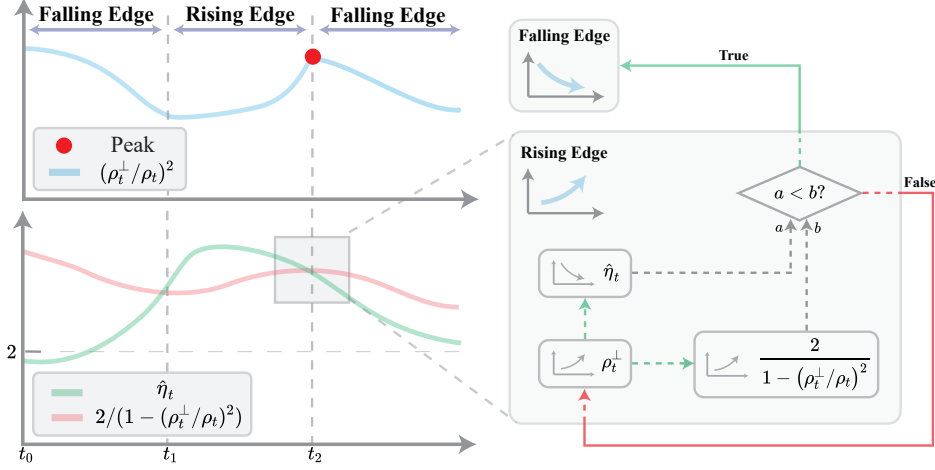


Fig. 2 Loss spike process of batch normalized linear regression. The relative magnitude of $2/(1 - (\rho_t^\perp/\rho_t)^2)$ and $\hat{\eta}_t$ determines the system state. When it is *Rising Edge*, a negative feedback loop is activated due to the increase in ρ_t^\perp , ensuring that the training dynamics inevitably returns to *Falling Edge*.

Lemma 8 (The Dynamics of BN Linear Regression) *Let $\ell = \ell_{squ}$, $\hat{\mathbf{w}} = \Sigma^{-1}\mu$ and $\Sigma = \mathbf{I}$. Consider the gradient descent (1), it holds that*

$$\begin{aligned} (1). \frac{\rho_{t+1}^\perp}{\rho_{t+1}} &= \frac{|\hat{\eta}_t - 1|}{1 + \hat{\eta}_t (\rho_t^\perp/\rho_t)^2} \frac{\rho_t^\perp}{\rho_t}; \\ (2). \alpha_{t+1} &= \alpha_t + \eta_\alpha (\rho_t - \alpha_t); \\ (3). \|\mathbf{w}_{t+1}\|^2 &= \|\mathbf{w}_t\|^2 + \frac{\eta^2 \alpha_t^2}{\|\mathbf{w}_t\|^2} (\rho_t^\perp)^2, \end{aligned}$$

where $\hat{\eta}_t$ is effective learning rate, defined as $\hat{\eta}_t := \eta \alpha_t \rho_t / \|\mathbf{w}_t\|^2$.

Remark 2 Lemma 8 provides an almost minimal and simplest dynamical system model for analyzing batch normalization. It not only retains the mechanisms introduced by BN that are not yet fully understood while discarding other unimportant details, but also offers a simplified model that helps understanding the training instability caused by BN in more complex problems and tasks.

The key to understanding loss spikes and training instability in batch normalization model is to determine the system state transitions between *Falling Edge* and *Rising Edge*. Lemma 8 provides precise characterization of the trend of ρ_t^\perp/ρ_t . An immediate observation is: if $\hat{\eta}_t > 2/(1 - (\rho_t^\perp/\rho_t)^2)$, then ρ_t^\perp/ρ_t decreases; otherwise, ρ_t^\perp/ρ_t increases. Inspired by that, we discuss how the dynamics transitions between two states.

The Transition from *Falling Edge* to *Rising Edge*. During the *Falling Edge*, since $\hat{\mathbf{w}}^2 = \rho_t^2 + (\rho_t^\perp)^2$, we observe that ρ_t^\perp is decreasing while ρ_t is increasing. (2) of Lemma 8 indicates that α_t keeps tracking ρ_t at a linear rate. When \mathbf{w}_t becomes

nearly collinear with $\hat{\mathbf{w}}$, $\|\mathbf{w}_t\|$ does not almost increase by (3) of Lemma 8. Overall, $\hat{\eta}_t$ increases while $2/(1 - (\rho_t^\perp/\rho_t)^2)$ decreases. Over time, if $\hat{\eta}_t$ exceeds $2/(1 - (\rho_t^\perp/\rho_t)^2)$, the system transitions to the *Rising Edge* and the loss spike occurs.

Next, we demonstrate that *Rising Edge* always terminates, as a negative feedback loop is activated to regulate $\hat{\eta}_t$, ensuring convergence.

The Transition from *Rising Edge* to *Falling Edge*. During the *Rising Edge*, we observe that ρ_t^\perp is increasing while ρ_t is decreasing. Since α_t continues to track ρ_t , α_t is also decreasing. Moreover, another factor controlling the descent of $\hat{\eta}_t$ is the increase in $\|\mathbf{w}_t\|$. Specifically, by condition (3) of Lemma 8, if we set $t = 0$ as the start of the *Rising Edge*, a rough calculation yields:

$$\|\mathbf{w}_t\|^2 \geq c \cdot \left(\int_{\tau=0}^t (\rho_\tau^\perp)^2 d\tau \right)^{1/2} \geq c \cdot \rho_0^\perp \cdot \sqrt{t} \quad \text{before } \alpha_t \text{ is too small}$$

for some constant c . Therefore, $\|\mathbf{w}_t\|^2$ keeps growing as long as the *Rising Edge* persists and α_t has not converged to 0. If α_t has already been very close to 0, then $\hat{\eta}_t$ is also small. Consequently, there will always exist a point where $\hat{\eta}_t$ becomes smaller than $2/(1 - (\rho_t^\perp/\rho_t)^2)$ (refer to Lemma 14 in Appendix for formal result), even without considering its growth, whose lower bound is 2. At this point, the *Rising Edge* ends, the loss spike meets its *Falling Edge* and begins to decline.

Readers may refer to Figure 2 for a more intuitive understanding. The idea of formal proofs of Theorems 2 and 3 are rooted in the above discussions about the mutual transitions between the *Falling Edge* and *Rising Edge* states.

5.3 Logistic Regression

Since \mathbf{w}_t is only updated in $\text{span}(\mathbf{X})$, we introduce the maximal and minimal eigenvalues in $\text{span}(\mathbf{X})$ of for further analysis:

$$\lambda_{\max} := \sup_{\mathbf{w} \in \mathcal{X}} \frac{\|\mathbf{w}\|_{\Sigma}^2}{\|\mathbf{w}\|^2}, \quad \lambda_{\min} := \inf_{\mathbf{w} \in \mathcal{X}} \frac{\|\mathbf{w}\|_{\Sigma}^2}{\|\mathbf{w}\|^2},$$

where $\mathcal{X} = \text{span}(\mathbf{X}) \setminus \{0\}$. Due to the non-linearity of logistic loss, the analysis of logistic regression becomes significantly more complex. To related \mathcal{R}_t with ρ_t^\perp , we then need to bound logit by direction-related quantities.

Lemma 9 *Let ℓ be ℓ_{\log} and $\hat{\mathbf{w}}$ be the SVM solution as presented in Definition 1. Suppose Assumption 1 holds. Consider the gradient descent (1), for any $t \geq 0$, if $\rho(\mathbf{w}) > 0$, it holds that $\forall i \in [n]$*

$$\begin{aligned} (1). \quad & \left| \|\mathbf{w}_t\|_{\Sigma} - \gamma^2 \|\mathbf{w}_t\| \cdot \rho_t \right| \leq 2\sqrt{2}\lambda_{\max} \cdot \gamma \frac{\|\mathbf{w}_t\|_{\Sigma}^2}{\|\mathbf{w}_t\|_{\Sigma}} \cdot \rho_t^\perp; \\ (2). \quad & \left| y_i \langle \mathbf{w}_t, \mathbf{x}_i \rangle - \gamma^2 \|\mathbf{w}_t\| \cdot \rho_t \right| \leq \sqrt{\lambda_{\max} \gamma} \|\mathbf{w}_t\| \cdot \rho_t^\perp. \end{aligned}$$

By Lemma 9, one can upper bound the difference between $\ell(\alpha_t)$ and $\ell(\alpha_t \langle \mathbf{w}_t, \mathbf{x}_i y_i \rangle / \|\mathbf{w}_t\|_\Sigma)$, which constitutes the primary step in proving Lemma 4. Next, to describe the transitions of *Rising Edge* and *Falling Edge*, as established in Lemma 7, we need to bound $\|\nabla \mathcal{R}_t\|$ and $\langle \nabla \mathcal{R}_t, \hat{\mathbf{w}} \rangle$. Unfortunately, logistic regression does not exhibit the same elegant properties as linear regression. However, these terms can still be bounded by ρ_t^\perp .

Lemma 10 *Let ℓ be ℓ_{\log} and $\hat{\mathbf{w}}$ be the SVM solution as defined in Definition 1. Suppose Assumption 1 holds. Consider the gradient descent (1), for any $t \geq 0$, it holds that $-\alpha_t \cdot \langle \hat{\mathbf{w}}, \nabla_{\mathbf{w}} \mathcal{R}_t \rangle \geq 0$; and if $\alpha_t > 0$, we have*

$$-\langle \hat{\mathbf{w}}, \nabla_{\mathbf{w}} \mathcal{R}_t \rangle \geq \frac{\lambda_{\min}}{8} \frac{\alpha_t e^{-\alpha_t}}{\|\mathbf{w}_t\|_\Sigma} (\rho_t^\perp)^2.$$

Lemma 11 *Let ℓ be ℓ_{\log} and $\hat{\mathbf{w}}$ be the SVM solution as defined in Definition 1. Suppose Assumption 1 holds. Consider the gradient descent (1), for any $t \geq 0$, if $\alpha_t > 0$, it holds that*

$$\frac{\lambda_{\min}}{4} \frac{\alpha_t e^{-\alpha_t}}{\|\mathbf{w}_t\|_\Sigma} \rho_t^\perp \leq \|\nabla_{\mathbf{w}} \mathcal{R}_t\| \leq \sqrt{\lambda_{\max}} \frac{\alpha_t}{\|\mathbf{w}_t\|_\Sigma}; \quad \|\nabla_{\mathbf{w}} \mathcal{R}_t\| \leq \lambda_{\max} \frac{\alpha_t}{\|\mathbf{w}_t\|_\Sigma} (\alpha_t + 1) \rho_t^\perp.$$

By Lemma 11 and 10, we can characterize the transitions between *Falling Edge* and *Rising Edge* states.

The Transition from *Falling Edge* to *Rising Edge*. Combining the lower bound of $\|\nabla_{\mathbf{w}} \mathcal{R}_t\|$ with the divergence condition in Lemma 7, we can obtain the sufficient condition for the divergence of ρ_t^\perp :

$$\frac{\lambda_{\min}}{4} \cdot \frac{\alpha_t}{e^{\alpha_t}} \cdot \frac{\eta \rho_t}{\|\mathbf{w}_t\|_\Sigma \|\mathbf{w}_t\|} \geq \frac{2}{1 - (\rho_t^\perp / \rho_t)^2} \quad (2)$$

By (1) of Lemma 9, we can characterize why a spike can happen when the margin of dataset γ and ρ_t^\perp is very small. Specifically, by (1) of Lemma 9, we find $\|\mathbf{w}_t\|_\Sigma \rightarrow \gamma^2 \rho_t \|\mathbf{w}_t\|$ when $\rho_t^\perp \rightarrow 0$. In this case, (2) becomes

$$\frac{\lambda_{\min}}{4} \cdot \frac{\alpha_t}{e^{\alpha_t}} \cdot \frac{\eta}{\gamma^2 \|\mathbf{w}_t\|^2} \geq \frac{2}{1 - (\rho_t^\perp / \rho_t)^2}.$$

We observe that the left-hand side of the above inequality is of $1/\gamma^2$ order when \mathbf{w}_t is close $\hat{\mathbf{w}}$ and loss is very low. Therefore, with a small margin γ , the above condition is easily satisfied. This means datasets with a small margin are more prone to experiencing loss spikes.

The Transition from *Rising Edge* to *Falling Edge*. Similarly, plugging the bounds for $\|\nabla_{\mathbf{w}} \mathcal{R}_t\|$ and $\langle \nabla_{\mathbf{w}} \mathcal{R}_t, \hat{\mathbf{w}} \rangle$ into the convergence condition of Lemma 7, we can obtain:

$$\frac{4\lambda_{\max}^{3/2}}{\lambda_{\min}^2} \cdot \frac{\alpha_t}{e^{\alpha_t}} \cdot \frac{\eta \rho_t}{\|\mathbf{w}_t\|^2} \leq (\rho_t^\perp)^2$$

As both $\|\mathbf{w}_t\|$ and ρ_t^\perp keeps increasing, the spike will rapidly return to convergence. Moreover, this condition provides a lower bound, ensuring that ρ_t^\perp will steadily decrease as long as it remains above this value. The preceding discussion remains incomplete. We still need to prove that ρ_t^\perp can become sufficiently small to trigger the divergence condition in the gradient descent.

Lemma 12 *Let ℓ be ℓ_{\log} and $\hat{\mathbf{w}}$ be the SVM solution as defined in Definition 1. Suppose Assumption 1 holds and $\lambda_{\max} \geq 1$. Consider the gradient descent (1), for any $\tan_{\min} > 0$, if $\alpha_0 \leq \frac{1}{3} \log(\lambda_{\max})$, $\frac{\eta}{\|\mathbf{w}_0\|^2} \geq \Theta\left(\left(1 + \frac{1}{\tan_{\min}^2}\right) \cdot \gamma\right)$ and $\eta_\alpha \leq \Theta\left(\frac{\tan_{\min}^2}{1 + \tan_{\min}^2} \cdot \frac{\|\mathbf{w}_0\|^2}{\eta\gamma}\right)$, then there exists $t_0 < T_0$ such that $(\rho_{t_0}^\perp / \rho_{t_0})^2 \leq \tan_{\min}^2$, and for any $t < T_0$, we have $\frac{1}{2}\alpha_0 \leq \alpha_t \leq \frac{3}{2}\alpha_0$, where $T_0 = \Theta\left(\left(1 + \frac{1}{\tan_{\min}^2}\right) \cdot \frac{\eta\gamma}{\|\mathbf{w}_0\|^2}\right)$.*

By Lemma 12, we can expect ρ_t^\perp to be arbitrarily small at some time, provided setting a large enough η . Combining Lemma 12 with the condition for state transition (2) gives the proof of Theorem 6.

6 Numerical Experiment

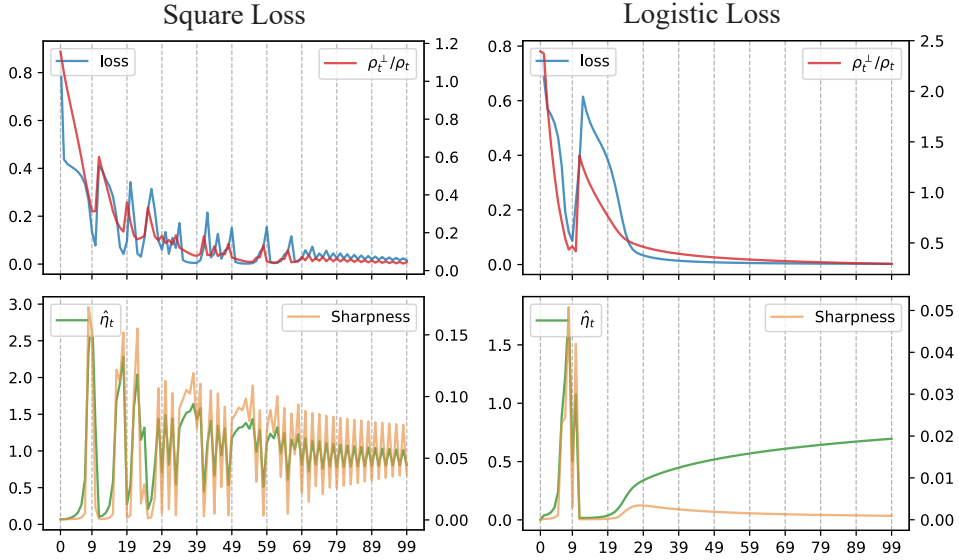


Fig. 3 Simulation experiments with square loss (left) and logistic loss (right). The x-axis represents the number of gradient descent iterations.

To verify our theoretical results, we conduct two simulation experiments using human-generated data. In general, reproducing loss spikes in the overparameterization case is challenging due to the existence of infinite solutions. To enable the occurrence

of loss spikes, we slice a Hilbert matrix, perform random rotations, and add Gaussian noise to construct our dataset. This ensures that the dataset has a very small margin and condition number. Our dataset contains 10 samples, and we set the feature dimension to 20 to align with the overparameterization setting. We train a batch-normalized linear model using both square loss and logistic loss.

As shown in Figure 3, the experimental results align with our theoretical understanding. Specifically, the loss can spike only when it is relatively low, and the spike always coincides with the divergence of ρ_t^\perp/ρ_t . This validates Lemmas 1 and 4, indicating that loss spikes are indeed caused by directional divergence.

Additionally, we record the value of the effective learning rate $\hat{\eta}_t$ at each iteration, calculated using the equation $\hat{\eta}_t = \alpha_t \cdot \rho_t / \|\mathbf{w}_t\|_\Sigma^2$. Almost every time, the sharp increase in $\hat{\eta}_t$ closely follows the loss spike, while its sharp decrease follows the redescend of the loss. This implies that if $\hat{\eta}_t$ becomes too large, divergence occurs; otherwise, the training loss continues to decline. This observation is consistent with our discussions in Section 5.

We also record the sharpness, defined as the maximal eigenvalue of the Hessian matrix. By observing the subfigures of the second row in Figure 3, we find that the changing trend of sharpness is very close to that of $\hat{\eta}_t$. This observation may inspire further investigation into the relationship between the effective learning rate and the optimization landscape.

7 Conclusion

This study elucidates the underlying mechanisms of loss spike. We demonstrate that batch normalization can induce loss spikes by causing directional divergence and loss spikes occur only when the loss has been reduced to a relatively low level. Future work can build on these findings to explore practical strategies for improving training stability in large-scale models, developing indicators for the occurrence of loss spikes during pretraining, and investigating the broader implications of effective learning rates on complex model architectures.

Appendix A Proof of Lemma 7

This is a property that any normalized models hold. We first prove this lemma.

Lemma 13 *Given a function $\mathcal{R}(\mathbf{w}) : \mathbb{R}^d \rightarrow \mathbb{R}$, if $\forall k \in \mathbb{R} \setminus \{0\}, \mathcal{R}(\mathbf{w}) = \mathcal{R}(k \cdot \mathbf{w})$, we have $\langle \mathbf{w}, \nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}) \rangle = 0$.*

Proof We have $\forall k \in \mathbb{R} \setminus \{0\}, \mathcal{R}(\mathbf{w}) = \mathcal{R}(k\mathbf{w})$. Then take derivatives with respect to k on both sides to obtain $\langle \nabla_{\mathbf{w}} \mathcal{R}(k \cdot \mathbf{w}), \mathbf{w} \rangle = 0$. By setting $k = 1$, we prove that $\langle \nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}), \mathbf{w} \rangle = 0$. \square

Lemma 7 (Directional Convergence and Divergence) Suppose there exists a reference direction $\hat{\mathbf{w}}$. Consider gradient descent (1) on an objective function $\mathcal{R}(\mathbf{w}, \alpha)$, where \mathbf{w} is parameterized by normalization and α represents the scaling factor of the normalization. We have the following direction convergence condition: if there exists a $t \geq 0$ such that

$$\rho_t > 0 \text{ and } \frac{\eta \rho_t}{\|\mathbf{w}_t\|} \|\nabla_{\mathbf{w}} \mathcal{R}_t\|^2 \leq -2 \langle \hat{\mathbf{w}}, \nabla_{\mathbf{w}} \mathcal{R}_t \rangle,$$

it holds that $(\rho_{t+1}^\perp)^2 \leq (\rho_t^\perp)^2$, and the following direction divergence condition: if there exists a $t \geq 0$ such that

$$0 < \rho_t^\perp / \rho_t \leq 1, \quad \alpha_t > 0 \text{ and } \frac{\eta}{\|\mathbf{w}_t\|} \|\nabla_{\mathbf{w}} \mathcal{R}_t\| \geq \frac{2\rho_t \rho_t^\perp}{\rho_t^2 - (\rho_t^\perp)^2},$$

it holds that $(\rho_{t+1}^\perp)^2 \geq (\rho_t^\perp)^2$.

Proof We prove the first result.

$$\begin{aligned} (\rho_{t+1}^\perp)^2 &= \left\| \hat{\mathbf{w}} - \frac{\rho_{t+1}}{\|\mathbf{w}_{t+1}\|} \mathbf{w}_{t+1} \right\|^2 \\ &\leq \left\| \hat{\mathbf{w}} - \frac{\rho_t}{\|\mathbf{w}_t\|} \mathbf{w}_{t+1} \right\|^2 \\ &= \left\| \hat{\mathbf{w}} - \frac{\rho_t}{\|\mathbf{w}_t\|} \mathbf{w}_t + \frac{\eta \rho_t}{\|\mathbf{w}_t\|} \nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}_t, \alpha_t) \right\|^2 \\ &= (\rho_t^\perp)^2 + \frac{\eta^2 \rho_t^2}{\|\mathbf{w}_t\|^2} \|\nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}_t, \alpha_t)\|^2 + \frac{2\eta \rho_t}{\|\mathbf{w}_t\|} \left(\hat{\mathbf{w}} - \frac{\rho_t}{\|\mathbf{w}_t\|} \mathbf{w}_t \right)^T \nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}_t, \alpha_t) \end{aligned}$$

where the inequality is since $\mathbf{w} \cdot \langle \mathbf{w}, \hat{\mathbf{w}} \rangle / \|\mathbf{w}\|^2$ is the projection of $\hat{\mathbf{w}}$ onto $\text{span}\{\mathbf{w}\}$ under the Euclidean inner product $\langle \cdot, \cdot \rangle$ and the last equation is because $\langle \mathbf{w}_t, \nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}_t, \alpha_t) \rangle = 0$ by Lemma 13. And then, we have

$$\begin{aligned} (\rho_{t+1}^\perp)^2 &\leq (\rho_t^\perp)^2 + \frac{\eta^2 \rho_t^2}{\|\mathbf{w}_t\|^2} \|\nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}_t, \alpha_t)\|^2 + \frac{2\eta \rho_t}{\|\mathbf{w}_t\|} \langle \hat{\mathbf{w}}, \nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}_t, \alpha_t) \rangle \\ &= (\rho_t^\perp)^2 + \frac{\eta \rho_t}{\|\mathbf{w}_t\|} \left(\frac{\eta \rho_t}{\|\mathbf{w}_t\|} \|\nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}_t, \alpha_t)\|^2 + 2 \langle \hat{\mathbf{w}}, \nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}_t, \alpha_t) \rangle \right) \quad \text{by } \rho_t > 0 \\ &\leq (\rho_t^\perp)^2, \end{aligned}$$

Next, we prove the second result. Since $(\rho_t^\perp)^2 + \rho_t^2 = \|\hat{\mathbf{w}}\|^2$, we check if $(\rho_{t+1}^\perp)^2 \geq (\rho_t^\perp)^2$ by verifying if $(\rho_{t+1}^\perp)^2 - (\rho_t^\perp)^2 \geq 0$. To simplify the calculation, we decompose $\hat{\mathbf{w}}$ orthogonally into two components: one in the $\text{span}\{\mathbf{w}_t\}$ and the other in a direction orthogonal to it:

$$\mathbf{e}_1 := \mathbf{w}_t / \|\mathbf{w}_t\|; \quad \mathbf{e}_2 := (\hat{\mathbf{w}} - \rho_t \mathbf{e}_1) / \rho_t^\perp.$$

It is easy to see that

$$\langle \mathbf{e}_1, \mathbf{e}_2 \rangle = 0; \quad \mathbf{w}_t = \|\mathbf{w}_t\| \mathbf{e}_1; \quad \hat{\mathbf{w}} = \rho_t \mathbf{e}_1 + \rho_t^\perp \mathbf{e}_2.$$

Recall that

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}} \mathcal{R}_t = \|\mathbf{w}_t\| \mathbf{e}_1 - \eta \nabla_{\mathbf{w}} \mathcal{R}_t$$

By Lemma 13, we know $\langle \mathbf{w}_t, \nabla_{\mathbf{w}} \mathcal{R}_t \rangle = 0$. We then calculate $\langle \mathbf{w}_{t+1}, \hat{\mathbf{w}} \rangle^2$ and $\|\mathbf{w}_{t+1}\|^2$.

$$\begin{aligned} \langle \mathbf{w}_{t+1}, \hat{\mathbf{w}} \rangle^2 &= \langle \|\mathbf{w}_t\| \mathbf{e}_1 - \eta \nabla_{\mathbf{w}} \mathcal{R}_t, \rho_t \mathbf{e}_1 + \rho_t^\perp \mathbf{e}_2 \rangle^2 \\ &= \left(\rho_t \|\mathbf{w}_t\| - \eta \rho_t^\perp \langle \nabla_{\mathbf{w}} \mathcal{R}_t, \mathbf{e}_2 \rangle \right)^2 \\ &= \rho_t^2 \|\mathbf{w}_t\|^2 + \eta^2 \left(\rho_t^\perp \right)^2 \langle \nabla_{\mathbf{w}} \mathcal{R}_t, \mathbf{e}_2 \rangle^2 - 2\eta \rho_t^\perp \rho_t \|\mathbf{w}_t\| \langle \nabla_{\mathbf{w}} \mathcal{R}_t, \mathbf{e}_2 \rangle; \\ \|\mathbf{w}_{t+1}\|^2 &= \|\mathbf{w}_t\|^2 + \eta^2 \|\nabla_{\mathbf{w}} \mathcal{R}_t\|^2 - 2\eta \langle \mathbf{w}_t, \nabla_{\mathbf{w}} \mathcal{R}_t \rangle \\ &= \|\mathbf{w}_t\|^2 + \eta^2 \|\nabla_{\mathbf{w}} \mathcal{R}_t\|^2, \end{aligned}$$

Then we check the sign of $\rho_{t+1}^2 - \rho_t^2$:

$$\begin{aligned} &\|\mathbf{w}_t\|^2 \|\mathbf{w}_{t+1}\|^2 \left(\rho_{t+1}^2 - \rho_t^2 \right) \\ &= \|\mathbf{w}_t\|^2 \langle \mathbf{w}_{t+1}, \hat{\mathbf{w}} \rangle^2 - \langle \mathbf{w}_t, \hat{\mathbf{w}} \rangle^2 \|\mathbf{w}_{t+1}\|^2 \\ &= \|\mathbf{w}_t\|^2 \left(\rho_t^2 \|\mathbf{w}_t\|^2 + \eta^2 \left(\rho_t^\perp \right)^2 \langle \nabla_{\mathbf{w}} \mathcal{R}_t, \mathbf{e}_2 \rangle^2 - 2\eta \rho_t^\perp \rho_t \|\mathbf{w}_t\| \langle \nabla_{\mathbf{w}} \mathcal{R}_t, \mathbf{e}_2 \rangle \right) \\ &\quad - \rho_t^2 \|\mathbf{w}_t\|^2 \left(\|\mathbf{w}_t\|^2 + \eta^2 \|\nabla_{\mathbf{w}} \mathcal{R}_t\|^2 \right) \\ &= \eta \|\mathbf{w}_t\|^2 \left(\eta \left(\rho_t^\perp \right)^2 \langle \nabla_{\mathbf{w}} \mathcal{R}_t, \mathbf{e}_2 \rangle^2 - 2\rho_t^\perp \rho_t \|\mathbf{w}_t\| \langle \nabla_{\mathbf{w}} \mathcal{R}_t, \mathbf{e}_2 \rangle - \rho_t^2 \eta \|\nabla_{\mathbf{w}} \mathcal{R}_t\|^2 \right) \\ &\leq \eta \|\mathbf{w}_t\|^2 \left(-\eta \left(\rho_t^2 - \left(\rho_t^\perp \right)^2 \right) \|\nabla_{\mathbf{w}} \mathcal{R}_t\|^2 + 2\rho_t^\perp \rho_t \|\mathbf{w}_t\| \|\nabla_{\mathbf{w}} \mathcal{R}_t\| \right) \\ &= \eta \|\mathbf{w}_t\|^2 \|\nabla_{\mathbf{w}} \mathcal{R}_t\| \left(-\eta \|\nabla_{\mathbf{w}} \mathcal{R}_t\| \left(\rho_t^2 - \left(\rho_t^\perp \right)^2 \right) + 2\rho_t^\perp \rho_t \|\mathbf{w}_t\| \right) \end{aligned}$$

Now, let's examine the signs within the parentheses. Recall $0 < \rho_t^\perp / \rho_t \leq 1$, it holds that

$$-\eta \left(\rho_t^2 - \left(\rho_t^\perp \right)^2 \right) \|\nabla_{\mathbf{w}} \mathcal{R}_t\|^2 \leq 0$$

This is the case that $\rho_t^\perp / \rho_t < 0$. Since $1 \geq \rho_t^\perp / \rho_t > 0$, it holds that $\rho_t^2 - \left(\rho_t^\perp \right)^2 > 0$. we know $\rho_t > 0$. Therefore, we can verify

$$\frac{\eta \|\nabla_{\mathbf{w}} \mathcal{R}_t\|}{\|\mathbf{w}_t\|} \geq \frac{2\rho_t \rho_t^\perp}{\rho_t^2 - \left(\rho_t^\perp \right)^2}$$

ensures that

$$-\eta \|\nabla_{\mathbf{w}} \mathcal{R}_t\| \left(\rho_t^2 - \left(\rho_t^\perp \right)^2 \right) + 2\rho_t^\perp \rho_t \|\mathbf{w}_t\| \leq 0.$$

This means $\rho_{t+1}^2 - \rho_t^2 \leq 0$. □

Appendix B Detailed Proof for Linear Regression

Lemma 1 (Decomposition of Mean Square Loss) Let $\ell = \ell_{squ}$, $\hat{\mathbf{w}} = \Sigma^{-1}\boldsymbol{\mu}$ and $\Sigma = \mathbf{I}$. Then, the following holds:

$$\mathcal{R}_t = (\alpha_t - \rho_t)^2 + (\rho_t^\perp)^2 + 1 - \|\hat{\mathbf{w}}\|^2$$

Proof Since $\Sigma = \mathbf{I}$, we have

$$\rho_t = \frac{\langle \hat{\mathbf{w}}, \mathbf{w}_t \rangle_\Sigma}{\|\mathbf{w}_t\|_\Sigma}; \quad \rho_t^\perp = \left\| \hat{\mathbf{w}} - \rho_t \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|_\Sigma} \right\|_\Sigma$$

Therefore, we have

$$\begin{aligned} \mathcal{R}_t &= \frac{1}{n} \left\| \alpha_t \frac{\tilde{\mathbf{X}}^T \mathbf{w}_t}{\|\mathbf{w}_t\|_\Sigma} - \mathbf{1}_n \right\|^2 \\ &= \alpha_t^2 + 1 - \frac{2\alpha_t}{n} \frac{\mathbf{w}_t^T \tilde{\mathbf{X}} \mathbf{1}}{\|\mathbf{w}_t\|_\Sigma} \\ &= \alpha_t^2 + 1 - 2\alpha_t \frac{\mathbf{w}_t^T \mathbf{u}}{\|\mathbf{w}_t\|_\Sigma} \\ &= \alpha_t^2 + 1 - 2\alpha_t \frac{\langle \mathbf{w}_t, \hat{\mathbf{w}}_t \rangle_\Sigma}{\|\mathbf{w}_t\|_\Sigma} \\ &= \alpha_t^2 + 1 - 2\alpha_t \rho_t \\ &= (\alpha_t - \rho_t)^2 + 1 - \rho_t^2 \\ &= (\alpha_t - \rho_t)^2 + \|\hat{\mathbf{w}}\|_\Sigma^2 - \rho_t^2 + 1 - \|\hat{\mathbf{w}}\|_\Sigma^2 \\ &= (\alpha_t - \rho_t)^2 + (\rho_t^\perp)^2 + 1 - \|\hat{\mathbf{w}}\|_\Sigma^2 \end{aligned}$$

Note that in Σ -inner product space, Pythagorean theorem also holds, which means $\rho_t^2 + (\rho_t^\perp)^2 = \|\hat{\mathbf{w}}\|_\Sigma^2$. It is easy to verify that a solution is reached when \mathbf{w} is colinear with $\hat{\mathbf{w}}$, which indicates that $\rho_t = \alpha_t = \|\hat{\mathbf{w}}\|$ and $\rho_t^\perp = 0$. Therefore, we have $\inf_{\mathbf{w}, \alpha} \mathcal{R}(\mathbf{w}, \alpha) = 1 - \|\hat{\mathbf{w}}\|_\Sigma^2$. This completes the proof. \square

Lemma 8 (The Dynamics of BN Linear Regression) Let $\ell = \ell_{squ}$, $\hat{\mathbf{w}} = \Sigma^{-1}\boldsymbol{\mu}$ and $\Sigma = \mathbf{I}$. Consider the gradient descent (1), it holds that

$$\begin{aligned} (1). \frac{\rho_{t+1}^\perp}{\rho_{t+1}} &= \frac{|\hat{\eta}_t - 1|}{1 + \hat{\eta}_t \left(\frac{\rho_t^\perp}{\rho_t} \right)^2} \frac{\rho_t^\perp}{\rho_t}; \\ (2). \alpha_{t+1} &= \alpha_t + \eta_\alpha (\rho_t - \alpha_t); \\ (3). \|\mathbf{w}_{t+1}\|^2 &= \|\mathbf{w}_t\|^2 + \frac{\eta^2 \alpha_t^2}{\|\mathbf{w}_t\|^2} \left(\rho_t^\perp \right)^2, \end{aligned}$$

where $\hat{\eta}_t$ is effective learning rate, defined as $\hat{\eta}_t := \eta \alpha_t \rho_t / \|\mathbf{w}_t\|^2$.

Proof (1). We first reformulate the gradient $\nabla_{\mathbf{w}} \mathcal{R}_t$:

$$\begin{aligned}
-\nabla_{\mathbf{w}} \mathcal{R}_t &= \frac{\alpha}{n \|\mathbf{w}\|_{\Sigma}} \left(\mathbf{I} - \frac{\Sigma \mathbf{w} \mathbf{w}^T}{\|\mathbf{w}\|_{\Sigma}^2} \right) \tilde{\mathbf{X}} \ell' \left(\frac{\alpha \tilde{\mathbf{X}}^T \mathbf{w}}{\|\mathbf{w}\|_{\Sigma}} \right) \\
&= \frac{\alpha}{n \|\mathbf{w}\|_{\Sigma}} \left(\mathbf{I} - \frac{\Sigma \mathbf{w} \mathbf{w}^T}{\|\mathbf{w}\|_{\Sigma}^2} \right) \tilde{\mathbf{X}} \left(\mathbf{1}_n - \alpha \cdot \frac{\tilde{\mathbf{X}}^T \mathbf{w}}{\|\mathbf{w}\|_{\Sigma}} \right) \\
&= \frac{\alpha_t}{\|\mathbf{w}_t\|_{\Sigma}} \left(\mathbf{I} - \frac{\Sigma \mathbf{w}_t \mathbf{w}_t^T}{\|\mathbf{w}_t\|_{\Sigma}^2} \right) \mathbf{u} \quad \text{recall } \mathbf{u} := \frac{1}{n} \tilde{\mathbf{X}} \mathbf{1}_n \\
&= \frac{\alpha_t}{\|\mathbf{w}_t\|_{\Sigma}} \Sigma \left(\Sigma^{-1} \mathbf{u} - \frac{\mathbf{w}_t^T \mathbf{u} \mathbf{w}_t}{\|\mathbf{w}_t\|_{\Sigma}^2} \right) \\
&= \frac{\alpha_t}{\|\mathbf{w}_t\|_{\Sigma}} \Sigma \left(\Sigma^{-1} \mathbf{u} - \frac{\mathbf{w}_t^T \Sigma \Sigma^{-1} \mathbf{u} \mathbf{w}_t}{\|\mathbf{w}_t\|_{\Sigma}^2} \right) \\
&= \frac{\alpha_t}{\|\mathbf{w}_t\|_{\Sigma}} \Sigma \left(\hat{\mathbf{w}} - \frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle_{\Sigma}}{\|\mathbf{w}_t\|_{\Sigma}^2} \mathbf{w}_t \right) \quad \text{recall } \hat{\mathbf{w}} := \Sigma^{-1} \mathbf{u} \\
&= \frac{\alpha_t}{\|\mathbf{w}_t\|} \left(\hat{\mathbf{w}} - \frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle}{\|\mathbf{w}_t\|^2} \mathbf{w}_t \right) \quad \text{by } \Sigma = \mathbf{I}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\|\nabla_{\mathbf{w}} \mathcal{R}_t\| &= \frac{|\alpha_t|}{\|\mathbf{w}_t\|} \rho_t^{\perp}; \\
-\langle \nabla_{\mathbf{w}} \mathcal{R}_t, \hat{\mathbf{w}} \rangle &= \frac{\alpha_t}{\|\mathbf{w}_t\|} \left\langle \hat{\mathbf{w}} - \frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle}{\|\mathbf{w}_t\|^2} \mathbf{w}_t, \hat{\mathbf{w}} \right\rangle = \frac{\alpha_t}{\|\mathbf{w}_t\|} \left(\|\hat{\mathbf{w}}\|^2 - \rho_t^2 \right) = \frac{\alpha_t}{\|\mathbf{w}_t\|} \left(\rho_t^{\perp} \right)^2.
\end{aligned}$$

Next, We decompose $\hat{\mathbf{w}}$ orthogonally into two components: one in the span $\{\mathbf{w}_t\}$ and the other in a direction orthogonal to it:

$$\mathbf{e}_1 = \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|}; \quad \mathbf{e}_2 = (\hat{\mathbf{w}} - \rho_t \mathbf{e}_1) / \rho_t^{\perp}.$$

Recall the gradient descent update:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}} \mathcal{R}_t = \|\mathbf{w}_t\| \mathbf{e}_1 - \eta \nabla_{\mathbf{w}} \mathcal{R}_t$$

We have

$$\begin{aligned}
\langle \mathbf{w}_{t+1}, \hat{\mathbf{w}} \rangle &= \langle \|\mathbf{w}_t\| \mathbf{e}_1 - \eta \nabla_{\mathbf{w}} \mathcal{R}_t, \rho_t \mathbf{e}_1 + \rho_t^{\perp} \mathbf{e}_2 \rangle \\
&= \rho_t \|\mathbf{w}_t\| - \eta \rho_t^{\perp} \langle \nabla_{\mathbf{w}} \mathcal{R}_t, \mathbf{e}_2 \rangle \\
&= \rho_t \|\mathbf{w}_t\| - \eta \langle \nabla_{\mathbf{w}} \mathcal{R}_t, \hat{\mathbf{w}} - \rho_t \mathbf{e}_1 \rangle \\
&= \rho_t \|\mathbf{w}_t\| - \eta \langle \nabla_{\mathbf{w}} \mathcal{R}_t, \hat{\mathbf{w}} \rangle \\
&= \rho_t \|\mathbf{w}_t\| + \frac{\eta \alpha_t}{\|\mathbf{w}_t\|} \left(\rho_t^{\perp} \right)^2
\end{aligned}$$

and

$$\begin{aligned}
\|\mathbf{w}_{t+1}\|^2 &= \|\mathbf{w}_t\|^2 + \eta^2 \|\nabla_{\mathbf{w}} \mathcal{R}_t\|^2 - 2\eta \langle \mathbf{w}_t, \nabla_{\mathbf{w}} \mathcal{R}_t \rangle \\
&= \|\mathbf{w}_t\|^2 + \eta^2 \|\nabla_{\mathbf{w}} \mathcal{R}_t\|^2 \\
&= \|\mathbf{w}_t\|^2 + \frac{\eta^2 \alpha_t^2}{\|\mathbf{w}_t\|^2} \left(\rho_t^{\perp} \right)^2,
\end{aligned}$$

where we apply $\langle \mathbf{w}_t, \nabla_{\mathbf{w}} \mathcal{R}_t \rangle = 0$ and $\langle \mathbf{e}_1, \nabla_{\mathbf{w}} \mathcal{R}_t \rangle = 0$. Then, by definition of ρ_t^\perp and ρ_t , we have

$$\begin{aligned}
\left(\frac{\rho_{t+1}^\perp}{\rho_{t+1}} \right)^2 &= \frac{\|\hat{\mathbf{w}}\|^2 - \rho_{t+1}^2}{\rho_{t+1}^2} \\
&= \frac{\|\mathbf{w}_{t+1}\|^2 \|\hat{\mathbf{w}}\|^2 - \langle \mathbf{w}_{t+1}, \hat{\mathbf{w}} \rangle^2}{\langle \mathbf{w}_{t+1}, \hat{\mathbf{w}} \rangle^2} \\
&= \frac{\left(\|\mathbf{w}_t\|^2 + \frac{\eta^2 \alpha_t^2}{\|\mathbf{w}_t\|^2} (\rho_t^\perp)^2 \right) \left(\rho_t^2 + (\rho_t^\perp)^2 \right) - \left(\rho_t \|\mathbf{w}_t\| + \frac{\eta \alpha_t}{\|\mathbf{w}_t\|} (\rho_t^\perp)^2 \right)^2}{\left(\rho_t \|\mathbf{w}_t\| + \frac{\eta \alpha_t}{\|\mathbf{w}_t\|} (\rho_t^\perp)^2 \right)^2} \\
&= (\rho_t^\perp)^2 \frac{\left(\frac{\eta^2 \alpha_t^2}{\|\mathbf{w}_t\|^2} \rho_t^2 + \|\mathbf{w}_t\|^2 - 2\eta \alpha_t \rho_t \right)}{\left(\rho_t \|\mathbf{w}_t\| + \frac{\eta \alpha_t}{\|\mathbf{w}_t\|} (\rho_t^\perp)^2 \right)^2} \\
&= \left(\frac{\rho_t^\perp}{\rho_t} \right)^2 \frac{(\hat{\eta}_t - 1)^2}{(1 + \hat{\eta}_t (\rho_t^\perp / \rho_t)^2)^2} \quad \text{remember } \hat{\eta}_t = \frac{\eta \alpha_t \rho_t}{\|\mathbf{w}_t\|^2}
\end{aligned}$$

(2). Then we consider the update of α_t , which is much simpler:

$$\begin{aligned}
\frac{\partial \mathcal{R}_t}{\partial \alpha} &= \frac{1}{n} \left(\frac{\tilde{\mathbf{X}}^T \mathbf{w}_t}{\|\mathbf{w}_t\|_\Sigma} \right)^T \ell' \left(\frac{\alpha_t \tilde{\mathbf{X}}^T \mathbf{w}_t}{\|\mathbf{w}_t\|_\Sigma} \right) \\
&= \frac{1}{n} \left(\frac{\tilde{\mathbf{X}}^T \mathbf{w}_t}{\|\mathbf{w}_t\|_\Sigma} \right)^T \left(\alpha_t \cdot \frac{\tilde{\mathbf{X}}^T \mathbf{w}_t}{\|\mathbf{w}_t\|_\Sigma} - \mathbf{1}_n \right) \\
&= \alpha_t - \frac{\mathbf{w}_t^T \mathbf{u}}{\|\mathbf{w}\|_\Sigma}
\end{aligned}$$

Therefore, we have

$$\alpha_{t+1} = \alpha_t - \eta_\alpha \frac{\partial \mathcal{R}_t}{\partial \alpha} = \alpha_t - \eta_\alpha \left(\alpha_t - \frac{\mathbf{w}_t^T \mathbf{u}}{\|\mathbf{w}_t\|_\Sigma} \right) = \alpha_t + \eta_\alpha (\rho_t - \alpha_t)$$

□

Lemma 14 Let $\ell = \ell_{sq}$ be the square loss and $\hat{\mathbf{w}} = \Sigma^{-1} \boldsymbol{\mu}$. Suppose $\Sigma = \mathbf{I}$. Consider the gradient descent (1) for $t > t_0$ with $\eta_\alpha \in (0, 1)$, where t_0 is such that $\rho_{t_0}^\perp / \rho_{t_0} \leq 1/\sqrt{3}$ and $0 < \alpha_{t_0} < \rho_{t_0}$. The following results hold:

1. **Condition of No Spike.** If $\frac{\eta}{\|\mathbf{w}_{t_0}\|^2} < \frac{2}{\|\hat{\mathbf{w}}\|^2}$, there shall be no spikes for any $t \geq t_0$.
2. **Condition of Spike.** There exists a constant C such that if η satisfies $\frac{8}{\|\hat{\mathbf{w}}\|^2} < \frac{\eta}{\|\mathbf{w}_{t_0}\|^2} \leq C$, then a spike will occur within at most $\Delta T_0 = \Theta \left(\ln \left(\frac{\eta_\alpha}{\eta \|\hat{\mathbf{w}}\|^2 (\rho_{t_0}^\perp / \rho_{t_0})^2} \right) / \eta_\alpha \right)$ iterations. Formally speaking, there exists $t_1 \in (t_0, t_0 + \Delta T_0]$ such that

$$\frac{\rho_{t+1}^\perp}{\rho_{t+1}} \leq \frac{\rho_t^\perp}{\rho_t} \quad \forall t \in [t_0, t_1) \quad \text{and} \quad \frac{\rho_{t_1+1}^\perp}{\rho_{t_1+1}} \geq \frac{\rho_{t_1}^\perp}{\rho_{t_1}}.$$

Theorem 2 (Condition of Spike) Let $\ell = \ell_{squ}$, $\hat{\mathbf{w}} = \Sigma^{-1}\mu$. Suppose $\Sigma = \mathbf{I}$. Consider the gradient descent (1) for $t > t_0$ with $\eta_\alpha \in (0, 1)$, where t_0 is such that $\rho_{t_0}^\perp/\rho_{t_0} \leq 1/\sqrt{3}$ and $0 < \alpha_{t_0} < \rho_{t_0}$. The following results hold:

1. **Condition of No Spike.** If $\frac{\eta}{\|\mathbf{w}_{t_0}\|^2} < \frac{2}{\|\hat{\mathbf{w}}\|^2}$, there shall be no spikes for any $t \geq t_0$.
2. **Condition of Spike.** There exists a constant C such that if η satisfies $\frac{8}{\|\hat{\mathbf{w}}\|^2} < \frac{\eta}{\|\mathbf{w}_{t_0}\|^2} \leq C$, then a spike will occur within at most $\Delta T_0 = \Theta\left(\ln\left(\frac{\eta_\alpha}{\eta\|\hat{\mathbf{w}}\|^2(\rho_{t_0}^\perp/\rho_{t_0})^2}\right)/\eta_\alpha\right)$ iterations. Formally speaking, there exists $t_1 \in (t_0, t_0 + \Delta T_0]$ such that

$$\frac{\rho_{t+1}^\perp}{\rho_{t+1}} \leq \frac{\rho_t^\perp}{\rho_t} \quad \forall t \in [t_0, t_1) \quad \text{and} \quad \frac{\rho_{t_1+1}^\perp}{\rho_{t_1+1}} \geq \frac{\rho_{t_1}^\perp}{\rho_{t_1}}.$$

where

$$C = \min\left(\frac{1}{\alpha_{t_0}\rho_{t_0}}, \frac{3}{16\|\hat{\mathbf{w}}\|^2} \frac{\eta_\alpha}{e^2(1 - \alpha_{t_0}/\rho_{t_0})}\right);$$

$$\Delta T_0 \leq \left\lceil \ln\left(\frac{\eta_\alpha\eta(1-k)\|\mathbf{w}_{t_0}\|^2}{4\eta^2\|\hat{\mathbf{w}}\|^2(\rho_{t_0}^\perp/\rho_{t_0})^2}\right)/\eta_\alpha + 1 \right\rceil.$$

Proof We first prove the condition of no spike, then the condition of spike.

Condition of No Spike We first prove $\alpha_t \leq \|\hat{\mathbf{w}}\| \quad \forall t \geq t_0$ by induction. When $t = t_0$, it holds that $\alpha_t \leq \rho_t \leq \|\hat{\mathbf{w}}\|$. Then we assume $\alpha_t \leq \|\hat{\mathbf{w}}\|$ for $t \geq t_0$ and consider α_{t+1} :

$$\alpha_{t+1} = \alpha_t + \eta_\alpha(\rho_t - \alpha_t) = (1 - \eta_\alpha)\alpha_t + \eta_\alpha\rho_t$$

Since $\eta_\alpha \in (0, 1)$, we observe that α_{t+1} is, in fact, the convex combination of ρ_t and α_t , both of which are smaller than $\|\hat{\mathbf{w}}\|$. Therefore, $\alpha_{t+1} \leq \|\hat{\mathbf{w}}\|$. By Lemma 8, if $\hat{\eta}_t$ is always smaller than $\frac{2}{1 - (\rho_t^\perp/\rho_t)^2}$, a spike shall never happen. We verify the value of $\hat{\eta}_t$. For any $t \geq 0$, we have

$$\hat{\eta}_t = \frac{\eta\alpha_t\rho_t}{\|\mathbf{w}_t\|^2} \leq \frac{\eta\|\hat{\mathbf{w}}\|^2}{\|\mathbf{w}_t\|^2} \leq \frac{\eta\|\hat{\mathbf{w}}\|^2}{\|\mathbf{w}_{t_0}\|^2} \leq 2 \leq \frac{2}{1 - (\rho_t^\perp/\rho_t)^2},$$

where the second inequality is by $\|\mathbf{w}_t\| \geq \|\mathbf{w}_{t_0}\| \quad \forall t_0 \leq t$ and the third inequality is by the upper bound of η .

Condition of Spike. By the upper bound of η , we can verify that following inequalities hold:

$$\hat{\eta}_{t_0} = \frac{\eta\alpha_{t_0}\rho_{t_0}}{\|\mathbf{w}_{t_0}\|^2} \leq \frac{2}{1 - (\rho_{t_0}^\perp/\rho_{t_0})^2},$$

which ensures that ρ_t^\perp/ρ_t is decreasing when $t = t_0$.

Without loss of generality, we treat the time interval $[t_0, t_0 + \Delta T_0]$ as $[0, \Delta T_0]$ by defining $t \leftarrow t - t_0$. This index shift does not affect the correctness of the proof. By the upper bound of η in condition (1), we have $\hat{\eta}_0 \leq \frac{2}{1 - (\rho_0^\perp/\rho_0)^2}$, which guarantees that dynamics is still in

Falling Edge state when $t = 0$. Next, we assume that ρ_t^\perp will be decreasing for all $t \geq 0$, then derive a contradiction to identify the spike. By this assumption, we have $\forall t \geq 0, \rho_{t+1}^\perp \leq \rho_t^\perp$ and $\rho_{t+1} \geq \rho_t$. By the update of α_t and $\eta_\alpha \in (0, 1)$, we have

$$\begin{aligned}\alpha_{t+1} &= \alpha_t + \eta_\alpha (\rho_t - \alpha_t) \\ &\geq \alpha_t + \eta_\alpha (\rho_0 - \alpha_t)\end{aligned}$$

Take the negative of both sides and add ρ_0 to obtain

$$\begin{aligned}\rho_0 - \alpha_{t+1} &\leq \rho_0 - \alpha_t - \eta_\alpha (\rho_0 - \alpha_t) \\ &\leq (1 - \eta_\alpha) (\rho_0 - \alpha_t) \\ &\leq \exp(-\eta_\alpha) (\rho_0 - \alpha_t) \quad \text{since } \eta_\alpha \in (0, 1) \\ &\leq \exp(-\eta_\alpha (t + 1)) (\rho_0 - \alpha_0)\end{aligned}$$

Then we obtain the lower bound of α_t :

$$\forall t \geq 0, \quad \alpha_t \geq \rho_0 - e^{-\eta_\alpha t} (\rho_0 - \alpha_0)$$

Next, we calculate the upper bound of $\|\mathbf{w}_t\|$:

$$\begin{aligned}\forall t \geq 0, \quad \|\mathbf{w}_t\|^2 &= \|\mathbf{w}_0\|^2 + \eta^2 \sum_{\tau=0}^{t-1} \frac{\alpha_\tau^2 (\rho_\tau^\perp)^2}{\|\mathbf{w}_\tau\|^2} \\ &\leq \|\mathbf{w}_0\|^2 + \frac{\eta^2 \|\hat{\mathbf{w}}\|^2}{\|\mathbf{w}_0\|^2} \sum_{\tau=0}^{t-1} (\rho_\tau^\perp)^2 \quad \text{since } \|\mathbf{w}_t\| \geq \|\mathbf{w}_0\| \\ &\leq \|\mathbf{w}_0\|^2 + \frac{\eta^2 \|\hat{\mathbf{w}}\|^2}{\|\mathbf{w}_0\|^2} (\rho_0^\perp)^2 t \quad \text{since } \rho_t^\perp \text{ is decreasing}\end{aligned}$$

Now, we can lower bound the effective learning rate $\hat{\eta}_t$:

$$\forall t \geq 0, \quad \hat{\eta}_t = \frac{\eta \alpha_t \rho_t}{\|\mathbf{w}_t\|^2} \geq \frac{\eta \alpha_t \rho_0}{\|\mathbf{w}_0\|^2 + \frac{\eta^2 \|\hat{\mathbf{w}}\|^2}{\|\mathbf{w}_0\|^2} (\rho_0^\perp)^2 t} \geq \frac{\eta \rho_0^2 - \eta \rho_0 (\rho_0 - \alpha_0) e^{-\eta_\alpha t}}{\|\mathbf{w}_0\|^2 + \frac{\eta^2 \|\hat{\mathbf{w}}\|^2}{\|\mathbf{w}_0\|^2} (\rho_0^\perp)^2 t}$$

By Lemma.8, if $\hat{\eta}_t$ is larger than $\frac{2}{1 - (\rho_t^\perp / \rho_t)^2}$, $\rho_{t+1}^\perp / \rho_{t+1}$ would be larger than ρ_t^\perp / ρ_t , which is, in fact, that loss spike occurs. We now investigate whether the lower bound of $\hat{\eta}_t$ can exceed the threshold at some t . Therefore, according to the Lemma.8, assuming ρ_t^\perp is always decreasing leads to a contradiction, indicating the presence of a spike. We investigate if there exists a $t > 0$ such that

$$\frac{\eta \rho_0^2 - \eta \rho_0 (\rho_0 - \alpha_0) e^{-\eta_\alpha t}}{\|\mathbf{w}_0\|^2 + \frac{\eta^2 \|\hat{\mathbf{w}}\|^2}{\|\mathbf{w}_0\|^2} (\rho_0^\perp)^2 t} \geq \frac{2}{1 - (\rho_t^\perp)^2 / \rho_t^2} \quad (\text{B1})$$

Since $\rho_0^\perp \geq \rho_t^\perp$, we have

$$(\text{B1}) \Leftarrow \frac{\eta \rho_0^2 - \eta \rho_0 (\rho_0 - \alpha_0) e^{-\eta_\alpha t}}{\|\mathbf{w}_0\|^2 + \frac{\eta^2 \|\hat{\mathbf{w}}\|^2}{\|\mathbf{w}_0\|^2} (\rho_0^\perp)^2 t} \geq \frac{2}{1 - (\rho_0^\perp)^2 / \rho_0^2} \quad (\text{B2})$$

Rearrange the inequility to obtain

$$\eta \rho_0^2 - \frac{2 \|\mathbf{w}_0\|^2}{1 - (\rho_0^\perp)^2 / \rho_0^2} > \frac{2}{1 - (\rho_0^\perp)^2 / \rho_0^2} \frac{\eta^2 \|\hat{\mathbf{w}}\|^2}{\|\mathbf{w}_0\|^2} (\rho_0^\perp)^2 t + \eta \rho_0 (\rho_0 - \alpha_0) e^{-\eta_\alpha t}$$

Devide ρ_0^2 on both side to obtain

$$\begin{aligned} \eta - \frac{2\|\mathbf{w}_0\|^2}{\rho_0^2 - (\rho_0^\perp)^2} &> \frac{2}{1 - (\rho_0^\perp)^2/\rho_0^2} \frac{\eta^2 \|\hat{\mathbf{w}}\|^2}{\|\mathbf{w}_0\|^2} \left(\rho_0^\perp/\rho_0\right)^2 t + \eta(1 - \alpha_0/\rho_0) e^{-\eta_\alpha t} \\ &= \frac{2}{1 - (\rho_0^\perp)^2/\rho_0^2} \frac{\eta^2 \|\hat{\mathbf{w}}\|^2}{\|\mathbf{w}_0\|^2} \left(\rho_0^\perp/\rho_0\right)^2 t + \eta(1 - k) e^{-\eta_\alpha t} \quad \text{we denote } k = \alpha_0/\rho_0 \in (0, 1) \end{aligned} \quad (\text{B3})$$

We take a look at the left hand side of the (B3):

$$\begin{aligned} \text{LHS of (B3)} &= \eta - \frac{2\|\mathbf{w}_0\|^2}{\rho_0^2 - (\rho_0^\perp)^2} \\ &= \eta - 2 \frac{\|\mathbf{w}_0\|^2}{\|\hat{\mathbf{w}}\|^2} \frac{1 + (\rho_0^\perp/\rho_0)^2}{1 - (\rho_0^\perp/\rho_0)^2} \\ &\geq \eta - \frac{4\|\mathbf{w}_0\|^2}{\|\hat{\mathbf{w}}\|^2} \quad \text{since } \rho_0^\perp/\rho_0 \leq 1/\sqrt{3} \\ &\geq \frac{1}{2}\eta \quad \text{by } \eta > 8\|\mathbf{w}_0\|^2/\|\hat{\mathbf{w}}\|^2 \end{aligned} \quad (\text{B4})$$

Then we look at the right hand side of the (B3):

$$\begin{aligned} \text{RHS of (B3)} &= \frac{2\left(\rho_0^\perp/\rho_0\right)^2}{1 - (\rho_0^\perp)^2/\rho_0^2} \frac{\eta^2 \|\hat{\mathbf{w}}\|^2}{\|\mathbf{w}_0\|^2} t + \eta(1 - k) e^{-\eta_\alpha t} \\ &\leq \underbrace{4 \frac{\eta^2 \|\hat{\mathbf{w}}\|^2}{\|\mathbf{w}_0\|^2} \left(\rho_0^\perp/\rho_0\right)^2 t}_B + \underbrace{\eta(1 - k) e^{-\eta_\alpha t}}_A \quad \text{since } \rho_0^\perp/\rho_0 \leq 1/\sqrt{2} \end{aligned}$$

We discuss the minimum of the function $f(t; a, b) = Ae^{-\eta_\alpha t} + Bt$. Its minimum is achieved by $t_{\min} = \ln(A\eta_\alpha/B)/\eta_\alpha$. As t_{\min} is unlikely to be an integer, we calculate $f(t)$ at the integers closest to t_{\min} . Since $f(t)$ is monotonically decreasing in $t < t_{\min}$. We have

$$\begin{aligned} f(\lfloor t_{\min} + 1 \rfloor) &\leq f(t_{\min} + 1) = f(t_{\min}) + f(t_{\min} + 1) - f(t_{\min}) \\ &= f(t_{\min}) + Ae^{-\eta_\alpha(t_{\min}+1)} + B(t_{\min} + 1) - Ae^{-\eta_\alpha t_{\min}} - Bt_{\min} \\ &= f(t_{\min}) + B \left(\frac{e^{-\eta_\alpha} - 1}{\eta_\alpha} + 1 \right) \leq f(t_{\min}) + B \quad \text{since } \eta_\alpha \in (0, 1) \end{aligned}$$

And $f_{\min} = f(t_{\min}) = B \ln(A\eta_\alpha e/B)/\eta_\alpha$.

$$\begin{aligned} \inf_{t>0} \{\text{RHS of (B3)}\} &\leq f(\lfloor t_{\min} + 1 \rfloor) \\ &\leq f(t_{\min}) + B \\ &= B \ln(A\eta_\alpha e/B)/\eta_\alpha + B \\ &= B \ln(A\eta_\alpha e^{1+\eta_\alpha}/B)/\eta_\alpha \\ &\leq \sqrt{e^{1+\eta_\alpha} AB/\eta_\alpha} \quad \text{since } \ln(x) \leq \sqrt{x} \quad \forall x > 0 \\ &= \sqrt{4 \frac{e\eta^2 \|\hat{\mathbf{w}}\|^2}{\eta_\alpha \|\mathbf{w}_0\|^2} \eta(1 - k) \left(\rho_0^\perp/\rho_0\right)} \\ &= \frac{2\eta \|\hat{\mathbf{w}}\|}{\|\mathbf{w}_0\|} \sqrt{\frac{\eta e^{1+\eta_\alpha} (1 - k)}{\eta_\alpha}} \left(\rho_0^\perp/\rho_0\right) \end{aligned} \quad (\text{B5})$$

To verify if there exists a t make the inequality (B1) establish, we have

$$\begin{aligned}
(B1) &\Leftarrow (B2) \Leftrightarrow (B3) \Leftrightarrow \text{LHS of (B3)} \geq \text{RHS of (B3)} \quad \text{for some } t \geq 1 \\
&\Leftarrow \text{LHS of (B3)} \geq \inf_{t>0} \{\text{RHS of (B3)}\} \\
&\Leftarrow \frac{1}{2}\eta \geq \frac{2\eta\|\hat{\mathbf{w}}\|}{\|\mathbf{w}_0\|} \sqrt{\frac{\eta e^{1+\eta_\alpha}(1-k)}{\eta_\alpha}} \left(\rho_0^\perp/\rho_0\right) \quad \text{by (B3) and (B5),} \\
&\Leftrightarrow \frac{\eta}{\|\mathbf{w}_0\|^2} \leq \frac{1}{16\|\hat{\mathbf{w}}\|^2} \frac{\eta_\alpha}{e^{1+\eta_\alpha}(1-k)} \left(\frac{\rho_0}{\rho_0^\perp}\right)^2 \\
&\Leftarrow \frac{\eta}{\|\mathbf{w}_0\|^2} \leq \frac{3}{16\|\hat{\mathbf{w}}\|^2} \frac{\eta_\alpha}{e^{1+\eta_\alpha}(1-k)} \quad \text{by } \rho_{t_0}^\perp/\rho_{t_0} \leq 1/\sqrt{3} \\
&\Leftarrow \frac{\eta}{\|\mathbf{w}_0\|^2} \leq \frac{3}{16\|\hat{\mathbf{w}}\|^2} \frac{\eta_\alpha}{e^2(1-k)}, \quad \text{by } \eta_\alpha \in (0,1)
\end{aligned}$$

which means after at most t_{\min} iterations, ρ_t^\perp will increase. This completes the proof. \square

Theorem 3 (Shape of the Loss Spike) *Let the settings of Theorem 2.2 hold, which guarantee the existence of a spike. Then, based on the results of Theorem 2.2, the Rising Edge will last for at most ΔT_1 iterations, then it returns to Falling Edge. Specifically, there exists a $t_2 \in (t_1, t_1 + \Delta T_1]$ such that*

$$\frac{\rho_{t+1}^\perp}{\rho_{t+1}} \geq \frac{\rho_t^\perp}{\rho_t} \quad \forall t \in [t_1, t_2) \quad \text{and} \quad \frac{\rho_{t_2+1}^\perp}{\rho_{t_2+1}} \leq \frac{\rho_{t_2}^\perp}{\rho_{t_2}},$$

where

$$\Delta T_1 = \left\lceil \frac{1}{4} \frac{\|\hat{\mathbf{w}}\|^4}{\alpha_{t_1}^2} \left(1/(\rho_{t_1}^\perp)^2 - 1/\rho_{t_1}^2\right)^2 \right\rceil + \left\lceil \frac{1}{4} \frac{\|\hat{\mathbf{w}}\|^2}{\rho_{t_1}^2} \left(\rho_{t_1}/\rho_{t_1}^\perp - \rho_{t_1}^\perp/\rho_{t_1}\right)^2 \right\rceil$$

Moreover, we define a time $\phi \in [t_1, t_2]$ as the first moment when α_t catches up with ρ_t , i.e., the time such that $\alpha_t \leq \rho_t \quad \forall t \in [t_1, \phi]$ and $\alpha_t \geq \rho_t \quad \forall t \in (\phi, t_2)$. Then, the dynamics of ρ_t^\perp/ρ_t is given by:

$$\forall t \in [t_1, \phi], \quad (\rho_t^\perp/\rho_t)^2 \leq 1 - \frac{2\rho_{t_1}^\perp \alpha_{t_1}}{\|\hat{\mathbf{w}}\|^2} \sqrt{t - t_1}; \quad \forall t > \phi, \quad (\rho_t^\perp/\rho_t)^2 \leq 1 - \frac{2\rho_{t_1}^\perp}{\|\hat{\mathbf{w}}\|} \sqrt{t - \phi}.$$

Proof By Lemma 15, we know that the *Rising Edge* will eventually terminate. We assume that *Rising Edge* ends after ΔT_1 iterations, i.e.,

$$\frac{\rho_t^\perp}{\rho_t} \leq \frac{\rho_{t+1}^\perp}{\rho_{t+1}} \quad \forall t \in [t_1, t_1 + \Delta T_1) \quad \text{and} \quad \frac{\rho_{t_1+\Delta T_1}^\perp}{\rho_{t_1+\Delta T_1}} \geq \frac{\rho_{t_1+\Delta T_1+1}^\perp}{\rho_{t_1+\Delta T_1+1}}.$$

And since ρ_t keeps increasing during $t \in [t_0, t_1)$ and $0 < \alpha_{t_0} < \rho_{t_0}$, it hold that $0 < \alpha_{t_1} < \rho_{t_1}$ by the gradient descent update of α_t . For simplicity, we treat the time interval $[t_1, t_1 + \Delta T_1)$ as $[0, \Delta T_1)$ by defining $t \leftarrow t - t_1$. Therefore, we have

$$\frac{\rho_t^\perp}{\rho_t} \leq \frac{\rho_{t+1}^\perp}{\rho_{t+1}} \quad \forall t \in [0, \Delta T_1), \quad \frac{\rho_{t_1}^\perp}{\rho_{t_1}} \geq \frac{\rho_{t_1+1}^\perp}{\rho_{t_1+1}} \quad \text{and} \quad 0 < \alpha_0 < \rho_0. \quad \text{provided } t \leftarrow t - t_1$$

It must hold that

$$\forall t \in [0, \Delta T_1), \quad \frac{\hat{\eta}_t - 1}{1 + \hat{\eta}_t \left(\frac{\rho_t^\perp}{\rho_t} \right)^2} \geq 1$$

by Lemma 8, since ρ_t^\perp / ρ_t keeps increasing for $t \in [0, \Delta T_1)$, which can be rewritten as

$$\forall t < \Delta T_1, \quad \left(\frac{\rho_t^\perp}{\rho_t} \right)^2 \leq 1 - \frac{2}{\hat{\eta}_t} \leq 1. \quad (\text{B6})$$

The inequality above can be used to evaluate the upper bound of $(\rho_t^\perp / \rho_t)^2$. Since $\alpha_0 < \rho_0$ while ρ_t continues to decrease, we divide $[0, \Delta T_1)$ into two phases, $P_1 = [0, \phi]$ and $P_2 = (\phi, \Delta T_1)$, such that for all $t \in P_1$, $\alpha_t \leq \rho_t$ and for all $t \in P_2$, $\alpha_t \geq \rho_t$. Note that P_2 may be empty, which means α_t never exceeds ρ_t during the *Rising Edge*.

The bound of α_t During P_1 , since α_t tracks ρ_t and ρ_t is always greater than α_t , it follows that α_t is increasing. From the definition of P_1 , we have $\alpha_t \leq \rho_t \leq \|\hat{\mathbf{w}}\|$, for all $t \in P_1$. Therefore, we obtain

$$\alpha_0 \leq \alpha_t \leq \|\hat{\mathbf{w}}\|, \quad \forall t \in P_1.$$

Next, consider P_2 . For any $t \in P_2$, we have $\alpha_t \geq \rho_t$. Given the update rule for α_t , it follows that α_t must decrease for $t \in P_2$. Thus, we have $\alpha_t \leq \alpha_\phi \leq \rho_\phi \leq \rho_0 \leq \|\hat{\mathbf{w}}\|$. Therefore, it holds that

$$\rho_t \leq \alpha_t \leq \|\hat{\mathbf{w}}\|, \quad \forall t \in P_2.$$

The bound of $\|\mathbf{w}_t\|^2$. Given any $t \in P_1$, we have

$$\|\mathbf{w}_t\|^2 = \|\mathbf{w}_0\|^2 + \eta^2 \sum_{\tau=0}^t \frac{\alpha_\tau^2 (\rho_\tau^\perp)^2}{\|\mathbf{w}_\tau\|^2} \geq \|\mathbf{w}_0\|^2 + \frac{\eta^2}{\|\mathbf{w}_t\|^2} \sum_{\tau=0}^t \alpha_\tau^2 (\rho_\tau^\perp)^2 \geq \|\mathbf{w}_0\|^2 + \frac{\eta^2 (\rho_0^\perp)^2}{\|\mathbf{w}_t\|^2} \alpha_0^2 t$$

We rearrange the inequality to obtain

$$\|\mathbf{w}_t\|^4 - \|\mathbf{w}_0\|^2 \|\mathbf{w}_t\|^2 - \eta^2 (\rho_0^\perp)^2 \alpha_0^2 t \geq 0$$

Solving the range of $\|\mathbf{w}_t\|^2$, we have

$$\|\mathbf{w}_t\|^2 \geq \left(\|\mathbf{w}_0\|^2 + \sqrt{\|\mathbf{w}_0\|^4 + 4\eta^2 (\rho_0^\perp)^2 \alpha_0^2 t} \right) / 2, \quad \forall t \in P_1$$

Then consider $t \in P_2$, we have

$$\|\mathbf{w}_t\|^2 \geq \|\mathbf{w}_0\|^2 + \frac{\eta^2}{\|\mathbf{w}_t\|^2} \sum_{\tau=0}^t \alpha_\tau^2 (\rho_\tau^\perp)^2 \geq \|\mathbf{w}_0\|^2 + \frac{\eta^2}{\|\mathbf{w}_t\|^2} \sum_{\tau=\phi}^t \alpha_\tau^2 (\rho_\tau^\perp)^2 \geq \|\mathbf{w}_0\|^2 + \frac{\eta^2 (\rho_0^\perp)^2}{\|\mathbf{w}_t\|^2} \rho_t^2 (t - \phi)$$

and

$$\|\mathbf{w}_t\|^2 \geq \left(\|\mathbf{w}_0\|^2 + \sqrt{\|\mathbf{w}_0\|^4 + 4\eta^2 (\rho_0^\perp)^2 \rho_t^2 (t - \phi)} \right) / 2, \quad \forall t \in P_2$$

The upper bound of $\hat{\eta}_t$. Given any $t \in P_1$, we have

$$\hat{\eta}_t = \eta \frac{\alpha_t \rho_t}{\|\mathbf{w}_t\|^2} \leq \frac{\eta \|\hat{\mathbf{w}}\|^2}{\left(\|\mathbf{w}_0\|^2 + \sqrt{\|\mathbf{w}_0\|^4 + 4\eta^2 (\rho_0^\perp)^2 \alpha_0^2 t} \right) / 2} \leq \frac{\eta \|\hat{\mathbf{w}}\|^2}{\sqrt{4\eta^2 (\rho_0^\perp)^2 \alpha_0^2 t} / 2} = \frac{\|\hat{\mathbf{w}}\|^2}{\rho_0^\perp \alpha_0} \frac{1}{\sqrt{t}}$$

And given any $t \in P_2$, we have

$$\begin{aligned}\hat{\eta}_t &= \eta \frac{\alpha_t \rho_t}{\|\mathbf{w}_t\|^2} \leq \frac{\eta \|\hat{\mathbf{w}}\| \rho_t}{\left(\|\mathbf{w}_0\|^2 + \sqrt{\|\mathbf{w}_0\|^4 + 4\eta^2 (\rho_0^\perp)^2 \rho_t^2 (t - \phi)} \right) / 2} \\ &\leq \frac{\eta \|\hat{\mathbf{w}}\| \rho_t}{\sqrt{4\eta^2 (\rho_0^\perp)^2 \rho_t^2 (t - \phi) / 2}} \\ &= \frac{\|\hat{\mathbf{w}}\|}{\rho_0^\perp} \frac{1}{\sqrt{t - \phi}}\end{aligned}$$

The minimum feasible ΔT_1 . By Lemma 8, we aim to find a t such that

$$\hat{\eta}_t \leq \frac{2}{1 - (\rho_t^\perp)^2 / \rho_t^2}, \quad (\text{B7})$$

which is the feasible solution of ΔT_1 . If P_2 is empty, which means $[0, \Delta T_1) = P_1$, then we have

$$\begin{aligned}(\text{B7}) &\Leftrightarrow \hat{\eta}_t \leq \frac{2}{1 - (\rho_0^\perp)^2 / \rho_0^2} \quad \text{since } \rho_t^\perp / \rho_t \text{ is increasing during } t \in [0, \Delta T_1) \\ &\Leftrightarrow \frac{\|\hat{\mathbf{w}}\|^2}{\rho_0^\perp \alpha_0} \frac{1}{\sqrt{t}} \leq \frac{2}{1 - (\rho_0^\perp)^2 / \rho_0^2} \quad \text{by the upper bound of } \hat{\eta}_t \\ &\Leftrightarrow t \geq \left\lceil \frac{1}{4} \left(1/(\rho_0^\perp)^2 - 1/\rho_0^2 \right)^2 \frac{\|\hat{\mathbf{w}}\|^4}{\alpha_0^2} \right\rceil.\end{aligned}$$

Next, consider the case that P_2 is not empty. In this case, it must be that $\phi < \left\lceil \frac{1}{4} \left(1/(\rho_0^\perp)^2 - 1/\rho_0^2 \right)^2 \frac{\|\hat{\mathbf{w}}\|^4}{\alpha_0^2} \right\rceil$, because otherwise the *Rising Edge* would terminate within P_1 . For $t \in P_2$, we have

$$\begin{aligned}(\text{B7}) &\Leftrightarrow \hat{\eta}_t \leq \frac{2}{1 - (\rho_0^\perp)^2 / \rho_0^2} \\ &\Leftrightarrow \frac{\|\hat{\mathbf{w}}\|}{\rho_0^\perp} \frac{1}{\sqrt{t - \phi}} \leq \frac{2}{1 - (\rho_0^\perp)^2 / \rho_0^2} \\ &\Leftrightarrow t \geq \phi + \frac{1}{4} \frac{\|\hat{\mathbf{w}}\|^2}{\rho_0^2} \left(\rho_0 / \rho_0^\perp - \rho_0^\perp / \rho_0 \right)^2 \\ &\Leftrightarrow t \geq \left\lceil \frac{1}{4} \left(1/(\rho_0^\perp)^2 - 1/\rho_0^2 \right)^2 \frac{\|\hat{\mathbf{w}}\|^4}{\alpha_0^2} \right\rceil + \left\lceil \frac{1}{4} \frac{\|\hat{\mathbf{w}}\|^2}{\rho_0^2} \left(\rho_0 / \rho_0^\perp - \rho_0^\perp / \rho_0 \right)^2 \right\rceil\end{aligned}$$

The upper bound of ρ_t^\perp / ρ_t . We give the upper bound of ρ_t^\perp / ρ_t . Recall $(\rho_t^\perp / \rho_t)^2 \leq 1 - 2/\hat{\eta}_t$, we have

$$(\rho_t^\perp / \rho_t)^2 \leq 1 - 2/\hat{\eta}_t \leq 1 - \frac{2\rho_0^\perp \alpha_0}{\|\hat{\mathbf{w}}\|^2} \sqrt{t}, \quad \forall t \in P_1$$

and

$$(\rho_t^\perp / \rho_t)^2 \leq 1 - 2/\hat{\eta}_t \leq 1 - \frac{2\rho_0^\perp}{\|\hat{\mathbf{w}}\|} \sqrt{t - \phi} \quad \forall t \in P_2$$

□

Lemma 15 (Existence of the Falling Edge of Loss Spike) *Let $\ell = \ell_{squ}$, $\hat{\mathbf{w}} = \Sigma^{-1}\boldsymbol{\mu}$ and $\Sigma = \mathbf{I}$. Consider the gradient descent (1). If the initial state satisfies $\hat{\eta}_0 > \frac{2}{1-(\rho_0^\perp/\rho_0)^2}$ and $\eta_\alpha \in (0, 1)$, then there must exist a $t > 0$ such that $\hat{\eta}_t < \frac{2}{1-(\rho_t^\perp/\rho_t)^2}$.*

Proof According to Lemma 8, the condition $\hat{\eta}_0 \geq \frac{2}{1-(\rho_0^\perp/\rho_0)^2}$ implies that ρ_0 is decreasing.

We assume that ρ_t^\perp is monotonically increasing. Then we prove the result by leading to a contradiction based on this assumption. Since ρ_t^\perp is bounded above by $\|\hat{\mathbf{w}}\|$, it must converge as $t \rightarrow \infty$. There are two possible cases: either ρ_t^\perp converges to $\|\hat{\mathbf{w}}\|$, or it does not.

Case 1. If $\rho_t^\perp \rightarrow \rho_\infty^\perp$ ($\rho_\infty^\perp < \|\hat{\mathbf{w}}\|$), we prove $\|\mathbf{w}_t\|$ can be large arbitrarily.

$$\begin{aligned}\|\mathbf{w}_t\|^2 &= \|\mathbf{w}_{t-1}\|^2 + \eta^2 \frac{\alpha_{t-1}^2 (\rho_{t-1}^\perp)^2}{\|\mathbf{w}_{t-1}\|^2} \\ &= \|\mathbf{w}_0\|^2 + \eta^2 \sum_{\tau=0}^{t-1} \frac{\alpha_\tau^2 (\rho_\tau^\perp)^2}{\|\mathbf{w}_\tau\|^2}\end{aligned}$$

We can assume that $\inf_{t \geq 0} \{\alpha_t\} = \alpha_{\min}$ is strictly larger than 0. Because if not, it means that there exists a time $t > 0$ such that α_t can be arbitrarily small, and therefore $\hat{\eta}_t$ can be small than 2, which is contradiction. Therefore, we have

$$\begin{aligned}\|\mathbf{w}_t\|^2 &= \|\mathbf{w}_0\|^2 + \eta^2 \sum_{\tau=0}^{t-1} \frac{\alpha_\tau^2 (\rho_\tau^\perp)^2}{\|\mathbf{w}_\tau\|^2} \\ &\geq \|\mathbf{w}_0\|^2 + \eta^2 \alpha_{\min}^2 \sum_{\tau=0}^{t-1} \frac{(\rho_\tau^\perp)^2}{\|\mathbf{w}_\tau\|^2} \\ &\geq \|\mathbf{w}_0\|^2 + \frac{\eta^2 \alpha_{\min}^2}{\|\mathbf{w}_t\|^2} \sum_{\tau=0}^{t-1} (\rho_\tau^\perp)^2 \quad \text{since } \|\mathbf{w}_t\| \text{ is monotonically increasing} \\ &\geq \|\mathbf{w}_0\|^2 + \frac{\eta^2 \alpha_{\min}^2}{\|\mathbf{w}_t\|^2} (\rho_0^\perp)^2 t \quad \text{by our assumption of increasing } \rho_t^\perp\end{aligned}$$

Rearrange the above inequality to obtain

$$\|\mathbf{w}_t\|^4 - \|\mathbf{w}_t\|^2 \|\mathbf{w}_0\|^2 - \eta^2 \inf_{t \geq 0} \{\alpha_\tau^2\} (\rho_0^\perp)^2 t \geq 0$$

Then we solve the above inequality:

$$\|\mathbf{w}_t\|^2 \geq \frac{1}{2} \left(\|\mathbf{w}_0\|^2 + \sqrt{\|\mathbf{w}_0\|^4 + 4\eta^2 \inf_{t \geq 0} \{\alpha_\tau^2\} (\rho_0^\perp)^2 t} \right)$$

Therefore, there always exists a large enough t such that $\|\mathbf{w}_t\|$ is very large and $\hat{\eta}_t \leq 2$, which means after that, ρ_t^\perp is no longer increasing. This leads to contradiction.

Case 2. If $\rho_t^\perp \rightarrow \|\hat{\mathbf{w}}\|$, we have $\rho_t \rightarrow 0$. Therefore $\hat{\eta}_t = \eta \frac{\alpha_t \rho_t}{\|\mathbf{w}_t\|^2} \rightarrow 0$. And before that, $\hat{\eta}_t$ should have already been less than 2 and ρ_t^\perp is decreasing. \square

Appendix C Detailed Proof for Logistic Regression

Lemma 9 Let ℓ be ℓ_{\log} and $\hat{\mathbf{w}}$ be the SVM solution as presented in Definition 1. Suppose Assumption 1 holds. Consider the gradient descent (1), for any $t \geq 0$, if $\rho_t > 0$, it holds that

$$\begin{aligned} (1). \quad & \left| \|\mathbf{w}_t\|_{\Sigma} - \gamma^2 \|\mathbf{w}_t\| \cdot \rho_t \right| \leq 2\sqrt{2}\lambda_{\max} \cdot \gamma \frac{\|\mathbf{w}_t\|^2}{\|\mathbf{w}_t\|_{\Sigma}} \cdot \rho_t^{\perp}; \\ (2). \quad & \left| y_i \langle \mathbf{w}_t, \mathbf{x}_i \rangle - \gamma^2 \|\mathbf{w}_t\| \cdot \rho_t \right| \leq \sqrt{\lambda_{\max}} \gamma \|\mathbf{w}_t\| \cdot \rho_t^{\perp}, \quad \forall i \in [n], \end{aligned}$$

Proof In the proof, we ignore the subscript of \mathbf{w}_t . And for further analysis, we introduce $\tilde{\rho}(\mathbf{w})$ and $\tilde{\rho}^{\perp}(\mathbf{w})$.

$$\tilde{\rho}(\mathbf{w}) := \frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle}{\|\hat{\mathbf{w}}\|}; \quad \tilde{\rho}^{\perp}(\mathbf{w}) := \left\| \mathbf{w} - \rho(\mathbf{w}) \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|} \right\|,$$

It is easy to see that $\tilde{\rho}(\mathbf{w})$ represents the length of \mathbf{w} in $\text{span}\{\hat{\mathbf{w}}\}$, while $\tilde{\rho}^{\perp}(\mathbf{w})$ corresponds to the length of \mathbf{w} in $\text{span}^{\perp}\{\hat{\mathbf{w}}\}$. Moreover, we note that $\rho(\mathbf{w})$ shares the same sign as $\tilde{\rho}(\mathbf{w})$, and they correspond to the respective sides in a pair of similar triangles. Therefore, we have

$$\begin{aligned} \cos \angle(\mathbf{w}, \hat{\mathbf{w}}) &= \rho(\mathbf{w}) / \|\hat{\mathbf{w}}\| = \tilde{\rho}(\mathbf{w}) / \|\mathbf{w}\|; \\ \sin \angle(\mathbf{w}, \hat{\mathbf{w}}) &= \tilde{\rho}^{\perp}(\mathbf{w}) / \|\hat{\mathbf{w}}\| = \tilde{\rho}^{\perp}(\mathbf{w}) / \|\mathbf{w}\|. \end{aligned}$$

We decompose \mathbf{w} orthogonally into two components: one in the same direction as $\hat{\mathbf{w}}$, and the other in a direction orthogonal to it. Recall the definition of $\tilde{\rho}(\mathbf{w})$ and $\tilde{\rho}^{\perp}(\mathbf{w})$, we can define

$$\tilde{\mathbf{e}}_1 = \hat{\mathbf{w}} / \|\hat{\mathbf{w}}\|; \quad \tilde{\mathbf{e}}_2 = (\mathbf{w} - \tilde{\rho}(\mathbf{w})\tilde{\mathbf{e}}_1) / \tilde{\rho}^{\perp}(\mathbf{w}),$$

and by Cauchy Inequality we have the following bounds

$$\begin{aligned} \|\tilde{\mathbf{e}}_1\|_{\Sigma}^2 &= \left(\frac{\|\hat{\mathbf{w}}\|_{\Sigma}}{\|\hat{\mathbf{w}}\|} \right)^2 = \gamma^2; \\ |\langle \tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2 \rangle_{\Sigma}| &\leq \lambda_{\max}; \\ \lambda_{\min} &\leq \|\tilde{\mathbf{e}}_2\|_{\Sigma}^2 \leq \lambda_{\max}. \end{aligned} \tag{C8}$$

(1). Therefore

$$\begin{aligned} \|\mathbf{w}\|_{\Sigma}^2 &= \left\| \tilde{\rho}(\mathbf{w})\tilde{\mathbf{e}}_1 + \tilde{\rho}^{\perp}(\mathbf{w})\tilde{\mathbf{e}}_2 \right\|_{\Sigma}^2 \\ &= (\tilde{\rho}(\mathbf{w}))^2 \|\tilde{\mathbf{e}}_1\|_{\Sigma}^2 + \left(\tilde{\rho}^{\perp}(\mathbf{w}) \right)^2 \|\tilde{\mathbf{e}}_2\|_{\Sigma}^2 + 2\tilde{\rho}(\mathbf{w})\tilde{\rho}^{\perp}(\mathbf{w})\langle \tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2 \rangle_{\Sigma} \\ &= \gamma^2 (\tilde{\rho}(\mathbf{w}))^2 + \left(\tilde{\rho}^{\perp}(\mathbf{w}) \right)^2 \|\tilde{\mathbf{e}}_2\|_{\Sigma}^2 + 2\tilde{\rho}(\mathbf{w})\tilde{\rho}^{\perp}(\mathbf{w})\langle \tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2 \rangle_{\Sigma} \end{aligned}$$

Plug the bounds in (C8) into above equation, we have

$$\begin{aligned} \|\mathbf{w}\|_{\Sigma}^2 &\leq (\tilde{\rho}(\mathbf{w}))^2 \gamma^2 + \lambda_{\max} \left(\tilde{\rho}^{\perp}(\mathbf{w}) \right)^2 + 2\lambda_{\max} \tilde{\rho}(\mathbf{w}) \tilde{\rho}^{\perp}(\mathbf{w}) \\ \|\mathbf{w}\|_{\Sigma}^2 &\geq (\tilde{\rho}(\mathbf{w}))^2 \gamma^2 + \lambda_{\min} \left(\tilde{\rho}^{\perp}(\mathbf{w}) \right)^2 - 2\lambda_{\max} \tilde{\rho}(\mathbf{w}) \tilde{\rho}^{\perp}(\mathbf{w}) \\ &\geq (\tilde{\rho}(\mathbf{w}))^2 \gamma^2 - \lambda_{\max} \left(\tilde{\rho}^{\perp}(\mathbf{w}) \right)^2 - 2\lambda_{\max} \tilde{\rho}(\mathbf{w}) \tilde{\rho}^{\perp}(\mathbf{w}) \end{aligned}$$

Combine them to give

$$\begin{aligned} \left| \|\mathbf{w}\|_{\Sigma}^2 - (\tilde{\rho}(\mathbf{w}))^2 \gamma^2 \right| &\leq \lambda_{\max} \left(\tilde{\rho}^{\perp}(\mathbf{w}) \right)^2 + 2\lambda_{\max} \tilde{\rho}(\mathbf{w}) \tilde{\rho}^{\perp}(\mathbf{w}) \\ &\leq 2\lambda_{\max} \tilde{\rho}^{\perp}(\mathbf{w}) \left(\tilde{\rho}^{\perp}(\mathbf{w}) + \tilde{\rho}(\mathbf{w}) \right) \end{aligned}$$

Recall the definitions of $\sin \angle(\mathbf{w}, \hat{\mathbf{w}})$ and $\cos \angle(\mathbf{w}, \hat{\mathbf{w}})$, we have

$$\begin{aligned} \left| \|\mathbf{w}\|_{\Sigma}^2 - \cos^2 \angle(\mathbf{w}, \hat{\mathbf{w}}) \|\mathbf{w}\|^2 \gamma^2 \right| &\leq 2\lambda_{\max} \|\mathbf{w}\|^2 \sin \angle(\mathbf{w}, \hat{\mathbf{w}}) (\sin \angle(\mathbf{w}, \hat{\mathbf{w}}) + \cos \angle(\mathbf{w}, \hat{\mathbf{w}})) \\ &\leq 2\sqrt{2}\lambda_{\max} \|\mathbf{w}\|^2 \sin \angle(\mathbf{w}, \hat{\mathbf{w}}), \end{aligned}$$

where the second inequality is by the range of $\angle(\mathbf{w}, \hat{\mathbf{w}})$. By our definition, $\angle(\mathbf{w}, \hat{\mathbf{w}}) \in [0, \pi]$. Next, we apply $\sin \angle(\mathbf{w}, \hat{\mathbf{w}}) = \rho^{\perp}(\mathbf{w}) / \hat{\mathbf{w}}$ and $\cos \angle(\mathbf{w}, \hat{\mathbf{w}}) = \rho(\mathbf{w}) / \hat{\mathbf{w}}$ to give

$$\left| \|\mathbf{w}\|_{\Sigma}^2 - \left(\gamma^2 \|\mathbf{w}\| \cdot \rho(\mathbf{w}) \right)^2 \right| \leq 2\sqrt{2}\lambda_{\max} \gamma \|\mathbf{w}\|^2 \cdot \rho^{\perp}(\mathbf{w})$$

Then we bound $\left| \|\mathbf{w}\|_{\Sigma} - \gamma^2 \|\mathbf{w}\| \cdot \rho(\mathbf{w}) \right|$,

$$\begin{aligned} \left| \|\mathbf{w}\|_{\Sigma} - \gamma^2 \|\mathbf{w}\| \cdot \rho(\mathbf{w}) \right| &= \frac{\left| \|\mathbf{w}\|_{\Sigma}^2 - \left(\gamma^2 \|\mathbf{w}\| \cdot \rho(\mathbf{w}) \right)^2 \right|}{\left| \|\mathbf{w}\|_{\Sigma} + \gamma^2 \|\mathbf{w}\| \cdot \rho(\mathbf{w}) \right|} \\ &\leq \left| \|\mathbf{w}\|_{\Sigma}^2 - \left(\gamma^2 \|\mathbf{w}\| \cdot \rho(\mathbf{w}) \right)^2 \right| / \|\mathbf{w}\|_{\Sigma} \\ &\leq 2\sqrt{2}\lambda_{\max} \gamma \|\mathbf{w}\|^2 \cdot \rho^{\perp}(\mathbf{w}) / \|\mathbf{w}\|_{\Sigma} \end{aligned}$$

(2). For the second result, by Cauchy inequality, for any $i \in [n]$, it holds that

$$y_i \langle \mathbf{w}, \mathbf{x}_i \rangle = y_i \tilde{\rho}(\mathbf{w}) \tilde{\rho}(\mathbf{x}_i) + y_i \left\langle \mathcal{P}^{\perp}(\mathbf{w}), \mathcal{P}^{\perp}(\mathbf{x}_i) \right\rangle \begin{cases} \leq \tilde{\rho}(\mathbf{w}) \gamma + \|\mathbf{x}_i\| \cdot \tilde{\rho}^{\perp}(\mathbf{w}) \\ \geq \tilde{\rho}(\mathbf{w}) \gamma - \|\mathbf{x}_i\| \cdot \tilde{\rho}^{\perp}(\mathbf{w}) \end{cases},$$

we have

$$|y_i \langle \mathbf{w}, \mathbf{x}_i \rangle - \tilde{\rho}(\mathbf{w}) \gamma| \leq \|\mathbf{x}_i\| \cdot \tilde{\rho}^{\perp}(\mathbf{w}) \leq \sqrt{\lambda_{\max}} \cdot \tilde{\rho}^{\perp}(\mathbf{w})$$

Recall that $\tilde{\rho}^{\perp}(\mathbf{w}) \cdot \|\hat{\mathbf{w}}\| = \rho^{\perp}(\mathbf{w}) \cdot \|\mathbf{w}\|$ and $\tilde{\rho}(\mathbf{w}) \cdot \|\hat{\mathbf{w}}\| = \rho(\mathbf{w}) \cdot \|\mathbf{w}\|$, we have the second result. \square

Lemma 4 (Upper Bound of Logistic Loss) *Let ℓ be ℓ_{\log} and $\hat{\mathbf{w}}$ be the SVM solution as presented in Definition 1. Suppose Assumption 1 holds. Consider the gradient descent (1), for any $t \geq 0$, if $\rho_t > 0$ and $\alpha_t > 0$, it holds that*

$$\mathcal{R}_t \leq \ell(\alpha_t) + \alpha_t \left| \ell' \left(\left[1 - C_0 \gamma \cdot \rho_t^{\perp} \right] \alpha_t \right) \right| \cdot C_0 \gamma \cdot \rho_t^{\perp},$$

where C_0 (defined in 9) is

$$C_0 := \frac{\sqrt{\lambda_{\max}}}{\sqrt{\lambda_{\min}}} + 2\sqrt{2} \frac{\lambda_{\max}}{\lambda_{\min}}.$$

Proof In the proof, we ignore the subscript of \mathbf{w}_t and α_t . We first bound the difference between $|\ell(\alpha \cdot \langle \mathbf{w}, \mathbf{x}_i y_i \rangle / \|\mathbf{w}\|_{\Sigma})|$ and $|\ell(\alpha)|$. For any $i \in [n]$, we have

$$\left| |\ell(\alpha \cdot \langle \mathbf{w}, \mathbf{x}_i y_i \rangle / \|\mathbf{w}\|_{\Sigma})| - |\ell(\alpha)| \right| \leq \max \left(|\ell'(\alpha)|, \left| \ell' \left(\alpha \frac{\langle \mathbf{w}, \mathbf{x}_i y_i \rangle}{\|\mathbf{w}\|_{\Sigma}} \right) \right| \right) |\alpha \cdot \langle \mathbf{w}, \mathbf{x}_i y_i \rangle / \|\mathbf{w}\|_{\Sigma} - \alpha|$$

Since $\alpha > 0$, we have

$$\begin{aligned} |\alpha \langle \mathbf{w}, \mathbf{x}_i y_i \rangle / \|\mathbf{w}\|_{\Sigma} - \alpha| &= \alpha |\langle \mathbf{w}, \mathbf{x}_i y_i \rangle - \|\mathbf{w}\|_{\Sigma}| / \|\mathbf{w}\|_{\Sigma} \\ &\leq \alpha \frac{|\langle \mathbf{w}, \mathbf{x}_i y_i \rangle - \gamma^2 \|\mathbf{w}\| \cdot \rho(\mathbf{w})|}{\|\mathbf{w}\|_{\Sigma}} + \alpha \frac{|\|\mathbf{w}\|_{\Sigma} - \gamma^2 \|\mathbf{w}\| \cdot \rho(\mathbf{w})|}{\|\mathbf{w}\|_{\Sigma}} \\ &\leq \alpha \sqrt{\lambda_{\max}} \frac{\|\mathbf{w}\|}{\|\mathbf{w}\|_{\Sigma}} \gamma \cdot \rho^{\perp}(\mathbf{w}) + \alpha \cdot 2\sqrt{2} \lambda_{\max} \frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|_{\Sigma}^2} \gamma \cdot \rho^{\perp}(\mathbf{w}) \quad \text{by Lemma 9} \\ &\leq \alpha \frac{\sqrt{\lambda_{\max}}}{\sqrt{\lambda_{\min}}} \frac{\|\mathbf{w}\|}{\|\mathbf{w}\|_{\Sigma}} \gamma \cdot \rho^{\perp}(\mathbf{w}) + \alpha \cdot 2\sqrt{2} \frac{\lambda_{\max}}{\lambda_{\min}} \gamma \cdot \rho^{\perp}(\mathbf{w}) \\ &= \alpha C_0 \gamma \cdot \rho^{\perp}(\mathbf{w}), \end{aligned} \tag{C9}$$

where we denote $\sqrt{\lambda_{\max}}/\sqrt{\lambda_{\min}} + 2\sqrt{2}\lambda_{\max}/\lambda_{\min}$ as C_0 . We have

$$\begin{aligned} \left| \ell' \left(\alpha \frac{\langle \mathbf{w}, \mathbf{x}_i y_i \rangle}{\|\mathbf{w}\|_{\Sigma}} \right) \right| &= \left| \ell' \left(\alpha \frac{\langle \mathbf{w}, \mathbf{x}_i y_i \rangle}{\|\mathbf{w}\|_{\Sigma}} - \alpha + \alpha \right) \right| \\ &\leq \left| \ell' \left(\alpha - \alpha C_0 \gamma \cdot \rho^{\perp}(\mathbf{w}) \right) \right| \quad \text{since } |\ell'(\cdot)| \text{ is a decreasing function} \\ &= \left| \ell' \left([1 - C_0 \gamma \cdot \rho^{\perp}(\mathbf{w})] \alpha \right) \right| \end{aligned}$$

Therefore, we have

$$\begin{aligned} \max \left(|\ell'(\alpha)|, \left| \ell' \left(\alpha \frac{\langle \mathbf{w}, \mathbf{x}_i y_i \rangle}{\|\mathbf{w}\|_{\Sigma}} \right) \right| \right) &\leq \max \left(|\ell'(\alpha)|, \left| \ell' \left([1 - C_0 \gamma \cdot \rho^{\perp}(\mathbf{w})] \alpha \right) \right| \right) \\ &\leq \left| \ell' \left([1 - C_0 \gamma \cdot \rho^{\perp}(\mathbf{w})] \alpha \right) \right| \end{aligned} \tag{C10}$$

Now we are ready to bound $\mathcal{R}(\mathbf{w}, \alpha)$:

$$\begin{aligned} &|\mathcal{R}(\mathbf{w}, \alpha) - |\ell'(\alpha)|| \\ &= \left| \frac{1}{n} \sum_{i=1}^n \ell \left(\alpha \cdot \frac{\langle \mathbf{w}, \mathbf{x}_i y_i \rangle}{\|\mathbf{w}\|_{\Sigma}} \right) - \ell(\alpha) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^n \left| \ell \left(\alpha \cdot \frac{\langle \mathbf{w}, \mathbf{x}_i y_i \rangle}{\|\mathbf{w}\|_{\Sigma}} \right) - \ell(\alpha) \right| \\ &\leq \max \left(\left| \ell' \left(\alpha \cdot \frac{\langle \mathbf{w}, \mathbf{x}_1 y_1 \rangle}{\|\mathbf{w}\|_{\Sigma}} \right) \right|, \dots, \left| \ell' \left(\alpha \cdot \frac{\langle \mathbf{w}, \mathbf{x}_n y_n \rangle}{\|\mathbf{w}\|_{\Sigma}} \right) \right|, |\ell'(\alpha)| \right) \frac{1}{n} \sum_{i=1}^n \left| \alpha \cdot \frac{\langle \mathbf{w}, \mathbf{x}_i y_i \rangle}{\|\mathbf{w}\|_{\Sigma}} - \alpha \right| \\ &\leq \frac{1}{n} \left| \ell' \left([1 - C_0 \gamma \cdot \rho^{\perp}(\mathbf{w})] \alpha \right) \right| \sum_{i=1}^n \left| \alpha \cdot \frac{\langle \mathbf{w}, \mathbf{x}_i y_i \rangle}{\|\mathbf{w}\|_{\Sigma}} - \alpha \right| \quad \text{by (C10)} \\ &\leq \left| \ell' \left([1 - C_0 \gamma \cdot \rho^{\perp}(\mathbf{w})] \alpha \right) \right| \alpha C_0 \gamma \cdot \rho^{\perp}(\mathbf{w}) \quad \text{by (C9)} \end{aligned}$$

□

Before proving Lemma 5, we need to introduce Assumption 2, Definition 2 Lemma 16. One can refer to Section 3.1 of Wu et al. (2024b) for details about these techniques.

Assumption 2 (Non-degenerate data) *Let $\hat{\mathbf{w}}$ be the SVM solution as presented in Definition 1. And let \mathcal{S} is the support vector set of the dataset. Assume that there exists $\alpha_i > 0, \forall i \in \mathcal{S}$ such that $\hat{\mathbf{w}} = \sum_{i \in \mathcal{S}} \alpha_i y_i \mathbf{x}_i$.*

Definition 2 (Margin Offset) *Suppose Assumption 2 and 1 hold. There exists the margin offset $b > 0$ such that*

$$-b := \max_{\mathbf{w} \in \text{span}^\perp\{\hat{\mathbf{w}}\} \cap \text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_n\}} \min_{i \in [n]} y_i \cdot \langle \mathbf{x}_i, \mathbf{w} \rangle / \|\mathbf{w}\|.$$

Lemma 16 *Suppose Assumption 2 and 1 hold. Definition 2 immediately implies that: for any $\mathbf{w} \in \text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ such that $\mathbf{w}^T \hat{\mathbf{w}} = 0$, there exist $i \in [n]$ such that $y_i \cdot \langle \mathbf{x}_i, \mathbf{w} \rangle \leq -b \cdot \|\mathbf{w}\|$.*

Proof Assumption 1 suggests that all sample \mathbf{x}_i in the dataset is support vector. Therefore we have $\text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_n\} = \text{span}\{\mathbf{x}_i | \mathbf{x}_i \in \mathcal{S}\}$.

If Assumption 2 holds, we can prove for any $\mathbf{v} \in \text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ such that $\langle \mathbf{v}, \hat{\mathbf{w}} \rangle = 0$, there exist $i, j \in \mathcal{S}$, such that $y_i \cdot \langle \mathbf{x}_i, \mathbf{v} \rangle < 0, y_j \cdot \langle \mathbf{x}_j, \mathbf{v} \rangle > 0$. To see this, we have

$$0 = \langle \mathbf{v}, \hat{\mathbf{w}} \rangle = \sum_{i \in \mathcal{S}} \alpha_i y_i \langle \mathbf{v}, \mathbf{x}_i \rangle$$

Since $\mathbf{v} \in \text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_n\} = \text{span}\{\mathbf{x}_i | \mathbf{x}_i \in \mathcal{S}\}$, there must exist $i, j \in [n]$, such that $y_i \cdot \langle \mathbf{x}_i, \mathbf{v} \rangle < 0, y_j \cdot \langle \mathbf{x}_j, \mathbf{v} \rangle > 0$. Therefore, the constant b as described in this lemma always exists. \square

Lemma 5 (Lower Bound of Logistic Loss) *Let ℓ be ℓ_{\log} and $\hat{\mathbf{w}}$ be the SVM solution as presented in Definition 1. Suppose Assumption 1 and 2 hold. Consider the gradient descent (1), for all $t \geq 0$, if $\alpha_t > 0$, it holds that*

$$\mathcal{R}_t \geq \frac{1}{n} \ell \left(\frac{\alpha_t}{\sqrt{\lambda_{\min}}} \left(\rho_t \gamma^2 - \rho_t^\perp b \gamma \right) \right),$$

where $b > 0$ is the margin offset (see Definition 2 in Appendix).

Proof Consider \mathcal{R}_t , we have

$$\mathcal{R}_t = \frac{1}{n} \sum_{i=1}^n \ell \left(\alpha_t \frac{\langle \mathbf{w}_t, \mathbf{x}_i y_i \rangle}{\|\mathbf{w}_t\|_{\Sigma}} \right)$$

By Definition 2 and Lemma 16, there exists a $j \in [n]$ such that

$$\begin{aligned} \left\langle y_j \mathbf{x}_j, \left(\mathbf{I} - \frac{\hat{\mathbf{w}} \hat{\mathbf{w}}^T}{\|\hat{\mathbf{w}}\|^2} \right) \mathbf{w}_t \right\rangle &\leq -b \left\| \left(\mathbf{I} - \frac{\hat{\mathbf{w}} \hat{\mathbf{w}}^T}{\|\hat{\mathbf{w}}\|^2} \right) \mathbf{w}_t \right\| \\ &= -b \frac{\left\| \left(\mathbf{I} - \frac{\hat{\mathbf{w}} \hat{\mathbf{w}}^T}{\|\hat{\mathbf{w}}\|^2} \right) \mathbf{w}_t \right\|}{\left\| \left(\mathbf{I} - \frac{\mathbf{w}_t \mathbf{w}_t^T}{\|\mathbf{w}_t\|^2} \right) \hat{\mathbf{w}} \right\|} \left\| \left(\mathbf{I} - \frac{\mathbf{w}_t \mathbf{w}_t^T}{\|\mathbf{w}_t\|^2} \right) \hat{\mathbf{w}} \right\| \\ &= -b \frac{\|\mathbf{w}_t\|}{\|\hat{\mathbf{w}}\|} \rho_t^{\perp} = -\rho_t^{\perp} b \gamma \|\mathbf{w}_t\| \end{aligned}$$

Then, we have

$$\begin{aligned} \langle \mathbf{w}_t, \mathbf{x}_i y_i \rangle &= \left\langle \left(\frac{\hat{\mathbf{w}} \hat{\mathbf{w}}^T}{\|\hat{\mathbf{w}}\|^2} \right) \mathbf{w}_t, \mathbf{x}_i y_i \right\rangle + \left\langle \left(\mathbf{I} - \frac{\hat{\mathbf{w}} \hat{\mathbf{w}}^T}{\|\hat{\mathbf{w}}\|^2} \right) \mathbf{w}_t, \mathbf{x}_i y_i \right\rangle \\ &\leq \gamma^2 \mathbf{w}_t^T \hat{\mathbf{w}} - \rho_t^{\perp} b \gamma \|\mathbf{w}_t\| \\ &= \gamma^2 \|\mathbf{w}_t\| \rho_t - \rho_t^{\perp} b \gamma \|\mathbf{w}_t\| \end{aligned}$$

Then

$$\mathcal{R}_t \geq \frac{1}{n} \ell \left(\alpha_t \frac{\langle \mathbf{w}_t, \mathbf{x}_j y_j \rangle}{\|\mathbf{w}_t\|_{\Sigma}} \right) \geq \frac{1}{n} \ell \left(\alpha_t \frac{\|\mathbf{w}_t\|}{\|\mathbf{w}_t\|_{\Sigma}} \left(\rho_t \gamma^2 - \rho_t^{\perp} b \gamma \right) \right) \geq \frac{1}{n} \ell \left(\frac{\alpha_t}{\sqrt{\lambda_{\min}}} \left(\rho_t \gamma^2 - \rho_t^{\perp} b \gamma \right) \right)$$

□

Lemma 10 Let ℓ be ℓ_{\log} and $\hat{\mathbf{w}}$ be the SVM solution as defined in Definition 1. Suppose Assumption 1 holds. Consider the gradient descent (1), for any $t \geq 0$, it holds that $-\alpha_t \cdot \langle \hat{\mathbf{w}}, \nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}, \alpha_t) \rangle \geq 0$; and if $\alpha_t > 0$, we have

$$-\langle \hat{\mathbf{w}}, \nabla_{\mathbf{w}} \mathcal{R}_t \rangle \geq \frac{\lambda_{\min}}{8} \frac{\alpha_t e^{-\alpha_t}}{\|\mathbf{w}_t\|_{\Sigma}} \left(\rho_t^{\perp} \right)^2.$$

Proof In this proof, we ignore the subscript of \mathbf{w}_t and α_t . Recall that

$$\nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}, \alpha) = \frac{\alpha}{n \|\mathbf{w}\|_{\Sigma}} \left(\mathbf{I} - \frac{\Sigma \mathbf{w} \mathbf{w}^T}{\|\mathbf{w}\|_{\Sigma}^2} \right) \tilde{\mathbf{X}} \ell' \left(\frac{\alpha \tilde{\mathbf{X}}^T \mathbf{w}}{\|\mathbf{w}\|_{\Sigma}} \right)$$

We have

$$\begin{aligned}
-\langle \hat{\mathbf{w}}, \nabla_{\mathbf{w}} \mathcal{R} \rangle &= -\frac{\alpha}{n\|\mathbf{w}\|_{\Sigma}} \hat{\mathbf{w}}^T \left(\mathbf{I} - \frac{\Sigma \mathbf{w} \mathbf{w}^T}{\|\mathbf{w}\|_{\Sigma}^2} \right) \tilde{\mathbf{X}} \ell' \\
&= -\frac{\alpha}{n\|\mathbf{w}\|_{\Sigma}} \hat{\mathbf{w}}^T \tilde{\mathbf{X}} \left(\mathbf{I} - \frac{\tilde{\mathbf{X}}^T \mathbf{w} (\tilde{\mathbf{X}}^T \mathbf{w})^T}{\|\tilde{\mathbf{X}}^T \mathbf{w}\|^2} \right) \ell' \\
&= -\frac{\alpha}{n\|\mathbf{w}\|_{\Sigma}} \mathbf{1}^T \left(\mathbf{I} - \frac{\mathbf{m} \mathbf{m}^T}{\|\mathbf{m}\|^2} \right) \ell', \quad \text{since } \langle \hat{\mathbf{w}}, \mathbf{x}_i y_i \rangle = 1 \quad \forall i \in [n]
\end{aligned}$$

where we denote $\mathbf{m} := \tilde{\mathbf{X}}^T \mathbf{w}$. Note that $\|\mathbf{m}\|^2 = n\|\mathbf{w}\|_{\Sigma}^2$. We have

$$\begin{aligned}
-\langle \hat{\mathbf{w}}, \nabla_{\mathbf{w}} \mathcal{R} \rangle &= -\frac{\alpha}{n^2\|\mathbf{w}\|_{\Sigma}^3} \mathbf{1}^T (\mathbf{m}^T \mathbf{m} \mathbf{I} - \mathbf{m} \mathbf{m}^T) \ell' \\
&= -\frac{\alpha}{2n^2\|\mathbf{w}\|_{\Sigma}^3} \text{Tr} \left(\ell' \mathbf{m}^T - \mathbf{m} \ell'^T \right)^T (\mathbf{1} \mathbf{m}^T - \mathbf{m} \mathbf{1}^T) \\
&= \frac{\alpha}{2n^2\|\mathbf{w}\|_{\Sigma}^3} \sum_{i=1}^n \sum_{j=1}^n (|[\ell']_i| |\mathbf{m}_j - \mathbf{m}_i| |[\ell']_j|) (\mathbf{m}_j - \mathbf{m}_i) \\
&= \frac{\alpha}{2n^2\|\mathbf{w}\|_{\Sigma}^3} \sum_{i=1}^n \sum_{j=1}^n |[\ell']_i| |[\ell']_j| (\mathbf{m}_j - \mathbf{m}_i)^2 \frac{|[\ell']_j|^{-1} \mathbf{m}_j - \mathbf{m}_i |[\ell']_i|^{-1}}{\mathbf{m}_j - \mathbf{m}_i}
\end{aligned} \tag{C11}$$

For the last term of above equation, we have

$$\begin{aligned}
\frac{|[\ell']_j|^{-1} \mathbf{m}_j - \mathbf{m}_i |[\ell']_i|^{-1}}{\mathbf{m}_j - \mathbf{m}_i} &= \frac{(1 + \exp(\alpha \mathbf{m}_j / \|\mathbf{w}\|_{\Sigma})) \mathbf{m}_j - (1 + \exp(\alpha \mathbf{m}_i / \|\mathbf{w}\|_{\Sigma})) \mathbf{m}_i}{\mathbf{m}_j - \mathbf{m}_i} \\
&= 1 + \frac{\exp(\alpha \mathbf{m}_j / \|\mathbf{w}\|_{\Sigma}) \mathbf{m}_j - \exp(\alpha \mathbf{m}_i / \|\mathbf{w}\|_{\Sigma}) \mathbf{m}_i}{\mathbf{m}_j - \mathbf{m}_i}.
\end{aligned}$$

Multiply both the numerator and denominator by $\alpha / \|\mathbf{w}\|_{\Sigma}$, and express $\mathbf{m}_i \alpha / \|\mathbf{w}\|_{\Sigma}$ as a and $\mathbf{m}_j \alpha / \|\mathbf{w}\|_{\Sigma}$ as b , we can obtain:

$$\frac{|[\ell']_j|^{-1} \mathbf{m}_j - \mathbf{m}_i |[\ell']_i|^{-1}}{\mathbf{m}_j - \mathbf{m}_i} = 1 + \frac{be^b - ae^a}{b-a} \geq 1 + \exp(\max(a, b)) = \max(|[\ell']_i|^{-1}, |[\ell']_j|^{-1}). \tag{C12}$$

Plug (C12) into (C11) to give

$$-\langle \hat{\mathbf{w}}, \nabla_{\mathbf{w}} \mathcal{R} \rangle = \frac{\alpha}{2n^2\|\mathbf{w}\|_{\Sigma}^3} \sum_{i=1}^n \sum_{j=1}^n (\mathbf{m}_j - \mathbf{m}_i)^2 \max(|[\ell']_i|, |[\ell']_j|)$$

Now we proved that $-\langle \hat{\mathbf{w}}, \nabla_{\mathbf{w}} \mathcal{R} \rangle$ and α always has the same sign. In the left part of this proof, we only focus on the case that $\alpha > 0$. Further, by the Lemma C.1 of (Cao et al. 2023), we have

$$-\langle \hat{\mathbf{w}}, \nabla_{\mathbf{w}} \mathcal{R} \rangle \geq \frac{\alpha}{8n^2\|\mathbf{w}\|_{\Sigma}^3} \frac{\sum_{i=1}^n |[\ell']_i|}{n} \sum_{i=1}^n \sum_{j=1}^n (\mathbf{m}_j - \mathbf{m}_i)^2 \geq 0 \tag{C13}$$

Then consider $\sum_{i=1}^n |[\ell']_i|/n$.

$$\begin{aligned}
\frac{\sum_{i=1}^n |[\ell']_i|}{n} &= \frac{1}{n} \sum_{i=1}^n \left[1 + \exp \left(\frac{\alpha \cdot \langle \mathbf{w}, \mathbf{x}_i y_i \rangle}{\|\mathbf{w}\|_{\Sigma}} \right) \right]^{-1} \\
&\geq \frac{1}{n} \sum_{i=1}^n \left[1 + \exp \left(\frac{\alpha \cdot |\langle \mathbf{w}, \mathbf{x}_i y_i \rangle|}{\|\mathbf{w}\|_{\Sigma}} \right) \right]^{-1} \quad \text{since } |\ell'(\cdot)| \text{ is decreasing} \\
&\geq \left[1 + \exp \left(\alpha \cdot \frac{1}{n} \sum_{i=1}^n \frac{|\langle \mathbf{w}, \mathbf{x}_i y_i \rangle|}{\|\mathbf{w}\|_{\Sigma}} \right) \right]^{-1} \quad \text{since } |\ell'(\cdot)| \text{ is convex in } [0, \infty) \\
&\geq [1 + \exp(\alpha)]^{-1} \\
&\geq \exp(-\alpha)/2.
\end{aligned} \tag{C14}$$

Then we relate $\sum_{i=1}^n \sum_{j=1}^n (\mathbf{m}_j - \mathbf{m}_i)^2$ with ρ_t^\perp .

$$\begin{aligned}
\sum_{i=1}^n \sum_{j=1}^n (\mathbf{m}_j - \mathbf{m}_i)^2 &= \text{Tr} \left(\mathbf{m} \mathbf{1}^T - \mathbf{1} \mathbf{m}^T \right)^T \left(\mathbf{m} \mathbf{1}^T - \mathbf{1} \mathbf{m}^T \right) \\
&= 2 \left(\|\mathbf{1}\|^2 \|\mathbf{m}\|^2 - (\mathbf{1}^T \mathbf{m})^2 \right) \\
&= 2 \|\mathbf{m}\|^2 \mathbf{1}^T \left(\mathbf{I} - \mathbf{m} \mathbf{m}^T / \|\mathbf{m}\|^2 \right) \mathbf{1}
\end{aligned}$$

Note that $\mathbf{I} - \mathbf{m} \mathbf{m}^T / \|\mathbf{m}\|^2$ is a projection matrix, we have

$$\sum_{i=1}^n \sum_{j=1}^n (\mathbf{m}_j - \mathbf{m}_i)^2 = 2 \|\mathbf{m}\|^2 \left\| \left(\mathbf{I} - \frac{\mathbf{m} \mathbf{m}^T}{\|\mathbf{m}\|^2} \right) \mathbf{1} \right\|^2$$

Recall that $\tilde{\mathbf{X}}^T \hat{\mathbf{w}} = \mathbf{1}$ and $\mathbf{m} = \tilde{\mathbf{X}}^T \mathbf{w}$, it holds that

$$\begin{aligned}
\sum_{i=1}^n \sum_{j=1}^n (\mathbf{m}_j - \mathbf{m}_i)^2 &= 2n \|\mathbf{w}\|_{\Sigma}^2 \left\| \left(\mathbf{I} - \frac{\tilde{\mathbf{X}}^T \mathbf{w} (\tilde{\mathbf{X}}^T \mathbf{w})^T}{\|\tilde{\mathbf{X}}^T \mathbf{w}\|^2} \right) \tilde{\mathbf{X}}^T \hat{\mathbf{w}} \right\|^2 \\
&= 2n \|\mathbf{w}\|_{\Sigma}^2 \left\| \tilde{\mathbf{X}}^T \left(\hat{\mathbf{w}} - \frac{\mathbf{w}^T \tilde{\mathbf{X}} \tilde{\mathbf{X}}^T \hat{\mathbf{w}}}{\|\tilde{\mathbf{X}}^T \mathbf{w}\|^2} \mathbf{w} \right) \right\|^2 \\
&= 2n \|\mathbf{w}\|_{\Sigma}^2 \left\| \tilde{\mathbf{X}}^T \left(\hat{\mathbf{w}} - \frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle_{\Sigma}}{\|\mathbf{w}\|_{\Sigma}^2} \mathbf{w} \right) \right\|^2 \\
&= 2n^2 \|\mathbf{w}\|_{\Sigma}^2 \left\| \hat{\mathbf{w}} - \frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle_{\Sigma}}{\|\mathbf{w}\|_{\Sigma}^2} \mathbf{w} \right\|_{\Sigma}^2
\end{aligned} \tag{C15}$$

We note that $\mathbf{w} \cdot \langle \mathbf{w}, \hat{\mathbf{w}} \rangle / \|\mathbf{w}\|^2$ is the projection of $\hat{\mathbf{w}}$ onto $\text{span}\{\mathbf{w}\}$ under the Euclidean inner product $\langle \cdot, \cdot \rangle$, therefore we have

$$\left\| \hat{\mathbf{w}} - \frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle_{\Sigma}}{\|\mathbf{w}\|_{\Sigma}^2} \mathbf{w} \right\|_{\Sigma}^2 \geq \lambda_{\min} \left\| \hat{\mathbf{w}} - \frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle_{\Sigma}}{\|\mathbf{w}\|_{\Sigma}^2} \mathbf{w} \right\|^2 \geq \lambda_{\min} \left\| \hat{\mathbf{w}} - \frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle}{\|\mathbf{w}\|^2} \mathbf{w} \right\|^2 = \lambda_{\min} \left(\rho^\perp(\mathbf{w}) \right)^2 \tag{C16}$$

Plugging (C14), (C15), (C16) into (C13) gives us

$$-\langle \hat{\mathbf{w}}, \nabla_{\mathbf{w}} \mathcal{R} \rangle \geq \frac{\lambda_{\min}}{8} \frac{\alpha e^{-\alpha}}{\|\mathbf{w}\|_{\Sigma}} \left(\rho^\perp(\mathbf{w}) \right)^2$$

□

Lemma 11 Let ℓ be ℓ_{\log} and $\hat{\mathbf{w}}$ be the SVM solution as defined in Definition 1. Suppose Assumption 1 holds. Consider the gradient descent (1), for any $t \geq 0$, if $\alpha_t > 0$, it holds that

$$\frac{\lambda_{\min}}{4} \frac{\alpha_t e^{-\alpha_t}}{\|\mathbf{w}_t\|_{\Sigma}} \rho_t^{\perp} \leq \frac{\sqrt{\lambda_{\min}}}{4} \frac{\alpha_t e^{-\alpha_t}}{\|\mathbf{w}_t\|_{\Sigma}} \rho_t^{\perp, \Sigma} \leq \|\nabla_{\mathbf{w}} \mathcal{R}_t\| \leq \frac{\alpha_t}{\|\mathbf{w}_t\|_{\Sigma}} \max\left(\sqrt{\lambda_{\max}}, \lambda_{\max}(\alpha_t + 1) \rho_t^{\perp}\right),$$

where $\rho_t^{\perp, \Sigma}$ is defined as

$$\rho_t^{\perp, \Sigma} := \rho^{\perp, \Sigma}(\mathbf{w}_t) := \left\| \hat{\mathbf{w}} - \rho_t \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|_{\Sigma}} \right\|_{\Sigma}.$$

Proof First Upper Bound We have

$$\begin{aligned} \|\nabla_{\mathbf{w}} \mathcal{R}\| &= \frac{\alpha}{n\|\mathbf{w}\|_{\Sigma}} \left\| \left(\mathbf{I} - \frac{\Sigma \mathbf{w} \mathbf{w}^T}{\|\mathbf{w}\|_{\Sigma}^2} \right) \tilde{\mathbf{X}} \ell' \right\| = \frac{\alpha}{n\|\mathbf{w}\|_{\Sigma}} \left\| \tilde{\mathbf{X}} \left(\mathbf{I} - \frac{\mathbf{m} \mathbf{m}^T}{\|\mathbf{m}\|^2} \right) \ell' \right\| \\ &\leq \frac{\alpha}{\sqrt{n}\|\mathbf{w}\|_{\Sigma}} \sqrt{\lambda_{\max}} \|\ell'\| \\ &\leq \frac{\alpha}{\|\mathbf{w}\|_{\Sigma}} \sqrt{\lambda_{\max}}, \end{aligned}$$

where the second inequality is since $|\ell'(\cdot)| \leq 1$ and $\|\ell'\| \leq \sqrt{n}$.

Second Upper Bound We introduce notation $|\ell'|$:

$$|\ell'| = \begin{bmatrix} \ell'(\alpha \cdot \langle \mathbf{w}, y_i \mathbf{x}_i \rangle / \|\mathbf{w}\|_{\Sigma}) \\ \vdots \\ \ell'(\alpha \cdot \langle \mathbf{w}, y_i \mathbf{x}_i \rangle / \|\mathbf{w}\|_{\Sigma}) \end{bmatrix}$$

Next, we bound the difference between $|\ell'|$ and $|\hat{\ell}'|$, where $|\hat{\ell}'|$ is

$$\begin{aligned} |\hat{\ell}'| &:= |\ell'| \left(\cos_{\Sigma} \angle(\mathbf{w}, \hat{\mathbf{w}}) \cdot \frac{\alpha \tilde{\mathbf{X}}^T \hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|_{\Sigma}} \right) = |\ell'| (\cos_{\Sigma} \angle(\mathbf{w}, \hat{\mathbf{w}}) \cdot \alpha) \mathbf{1}, \\ \text{and } \cos_{\Sigma} \angle(\mathbf{w}, \hat{\mathbf{w}}) &:= \frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle_{\Sigma}}{\|\mathbf{w}\|_{\Sigma} \|\hat{\mathbf{w}}\|_{\Sigma}}. \end{aligned}$$

Then, we have

$$\begin{aligned} \|\nabla_{\mathbf{w}} \mathcal{R}\| &= \frac{\alpha}{n\|\mathbf{w}\|_{\Sigma}} \left\| \left(\mathbf{I} - \frac{\Sigma \mathbf{w} \mathbf{w}^T}{\|\mathbf{w}\|_{\Sigma}^2} \right) \tilde{\mathbf{X}} \ell' \right\| \\ &= \frac{\alpha}{n\|\mathbf{w}\|_{\Sigma}} \left\| \tilde{\mathbf{X}} \left(\mathbf{I} - \frac{\mathbf{m} \mathbf{m}^T}{\|\mathbf{m}\|^2} \right) \ell' \right\| \\ &\leq \sqrt{\lambda_{\max}} \frac{\alpha}{\sqrt{n}\|\mathbf{w}\|_{\Sigma}} \left(\left\| \left(\mathbf{I} - \frac{\mathbf{m} \mathbf{m}^T}{\|\mathbf{m}\|^2} \right) (|\ell'| - |\hat{\ell}'|) \right\| + \right. \\ &\quad \left. |\ell'| (\cos_{\Sigma} \angle(\mathbf{w}, \hat{\mathbf{w}}) \cdot \alpha) \left\| \left(\mathbf{I} - \frac{\mathbf{m} \mathbf{m}^T}{\|\mathbf{m}\|^2} \right) \mathbf{1} \right\| \right), \end{aligned} \tag{C17}$$

where we denote $\mathbf{m} = \tilde{\mathbf{X}}^T \mathbf{w}$. For the first term in above inequality, we have

$$\begin{aligned}
& \left\| \left(\mathbf{I} - \frac{\mathbf{m}\mathbf{m}^T}{\|\mathbf{m}\|^2} \right) (|\ell'| - |\hat{\ell}'|) \right\| \\
& \leq \left\| |\ell'| - |\hat{\ell}'| \right\| \\
& = |\ell''(z)| \left\| \frac{\alpha \tilde{\mathbf{X}}^T \mathbf{w}}{\|\mathbf{w}\|_\Sigma} - \cos_\Sigma \angle(\mathbf{w}, \hat{\mathbf{w}}) \cdot \frac{\alpha \tilde{\mathbf{X}}^T \hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|_\Sigma} \right\| \\
& \leq \max \{ |[\ell']_1|, \dots, |[\ell']_n|, |\ell'(\cos_\Sigma \angle(\mathbf{w}, \hat{\mathbf{w}}) \cdot \alpha)| \} \cdot \left\| \frac{\alpha \tilde{\mathbf{X}}^T \mathbf{w}}{\|\mathbf{w}\|_\Sigma} - \cos_\Sigma \angle(\mathbf{w}, \hat{\mathbf{w}}) \cdot \frac{\alpha \tilde{\mathbf{X}}^T \hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|_\Sigma} \right\|,
\end{aligned}$$

where the equation is by Mean Value Theorem and z is between $\alpha \cdot \langle \mathbf{w}, \mathbf{x}_i y_i \rangle$ and $\cos_\Sigma \angle(\mathbf{w}, \hat{\mathbf{w}})$.
 α . Next, we have

$$\begin{aligned}
\left\| \frac{\alpha \tilde{\mathbf{X}}^T \mathbf{w}}{\|\mathbf{w}\|_\Sigma} - \cos_\Sigma \angle(\mathbf{w}, \hat{\mathbf{w}}) \cdot \frac{\alpha \tilde{\mathbf{X}}^T \hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|_\Sigma} \right\| & \leq \frac{\alpha}{\|\mathbf{w}\|_\Sigma} \left\| \tilde{\mathbf{X}}^T \mathbf{w} - \frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle_\Sigma}{\|\hat{\mathbf{w}}\|_\Sigma^2} \tilde{\mathbf{X}}^T \hat{\mathbf{w}} \right\| \\
& = \frac{\alpha}{\|\mathbf{w}\|_\Sigma} \left\| \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{\|\mathbf{1}\|^2} \right) \mathbf{m} \right\| \\
& = \frac{\alpha}{\|\mathbf{w}\|_\Sigma} \frac{\|\mathbf{m}\|}{\|\mathbf{1}\|} \left\| \left(\mathbf{I} - \frac{\mathbf{m}\mathbf{m}^T}{\|\mathbf{m}\|^2} \right) \mathbf{1} \right\| \\
& = \alpha \left\| \left(\mathbf{I} - \frac{\mathbf{m}\mathbf{m}^T}{\|\mathbf{m}\|^2} \right) \mathbf{1} \right\|
\end{aligned}$$

Now take all of above bounds into (C17) together

$$\begin{aligned}
\|\nabla_{\mathbf{w}} \mathcal{R}\| & \leq \sqrt{\lambda_{\max}} \frac{\alpha}{\sqrt{n} \|\mathbf{w}\|_\Sigma} \left(\alpha \max \{ |[\ell']_1|, \dots, |[\ell']_n|, |\ell'(\cos_\Sigma \angle(\mathbf{w}, \hat{\mathbf{w}}) \cdot \alpha)| \} + \right. \\
& \quad \left. |\ell'(\cos_\Sigma \angle(\mathbf{w}, \hat{\mathbf{w}}) \cdot \alpha)| \right) \cdot \left\| \left(\mathbf{I} - \frac{\mathbf{m}\mathbf{m}^T}{\|\mathbf{m}\|^2} \right) \mathbf{1} \right\| \\
& \leq \sqrt{\lambda_{\max}} \frac{\alpha}{\sqrt{n} \|\mathbf{w}\|_\Sigma} (\alpha + 1) \left\| \left(\mathbf{I} - \frac{\mathbf{m}\mathbf{m}^T}{\|\mathbf{m}\|^2} \right) \mathbf{1} \right\|
\end{aligned}$$

By (C15), we have

$$\left\| \left(\mathbf{I} - \frac{\mathbf{m}\mathbf{m}^T}{\|\mathbf{m}\|^2} \right) \mathbf{1} \right\|^2 = n \left\| \hat{\mathbf{w}} - \frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle_\Sigma}{\|\mathbf{w}\|_\Sigma^2} \mathbf{w} \right\|_\Sigma^2$$

And note that $\mathbf{w} \cdot \langle \mathbf{w}, \hat{\mathbf{w}} \rangle / \|\mathbf{w}\|^2$ is the projection of $\hat{\mathbf{w}}$ onto $\text{span}\{\mathbf{w}\}$ under the Euclidean inner product $\langle \cdot, \cdot \rangle$, therefore we have

$$\left\| \hat{\mathbf{w}} - \frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle_\Sigma}{\|\mathbf{w}\|_\Sigma^2} \mathbf{w} \right\|_\Sigma^2 \leq \left\| \hat{\mathbf{w}} - \frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle}{\|\mathbf{w}\|^2} \mathbf{w} \right\|_\Sigma^2 \leq \lambda_{\max} \left\| \hat{\mathbf{w}} - \frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle}{\|\mathbf{w}\|^2} \mathbf{w} \right\|^2 = \lambda_{\max} (\rho_t^\perp)^2 \quad (\text{C18})$$

Lower Bound We have

$$\|\nabla_{\mathbf{w}} \mathcal{R}\| = \frac{\alpha}{n \|\mathbf{w}\|_\Sigma} \left\| \tilde{\mathbf{X}} \left(\mathbf{I} - \frac{\mathbf{m}\mathbf{m}^T}{\|\mathbf{m}\|^2} \right) \ell' \right\| \geq \sqrt{\lambda_{\min}} \frac{\alpha}{\sqrt{n} \|\mathbf{w}\|_\Sigma} \left\| \left(\mathbf{I} - \frac{\mathbf{m}\mathbf{m}^T}{\|\mathbf{m}\|^2} \right) \ell' \right\| \quad (\text{C19})$$

By lagrange's identity, we have

$$\begin{aligned}
\left\| \left(\mathbf{I} - \frac{\mathbf{m}\mathbf{m}^T}{\|\mathbf{m}\|^2} \right) \ell' \right\|^2 &= \frac{1}{2\|\mathbf{m}\|^2} \sum_{i=1}^n \sum_{j=1}^n ([\ell']_i \mathbf{m}_j - [\ell']_j \mathbf{m}_i)^2 \\
&= \frac{1}{2\|\mathbf{m}\|^2} \sum_{i=1}^n \sum_{j=1}^n ([\ell']_i [\ell']_j)^2 (\mathbf{m}_j - \mathbf{m}_i)^2 \left(\frac{[\ell']_j^{-1} \mathbf{m}_j - [\ell']_i^{-1} \mathbf{m}_i}{\mathbf{m}_j - \mathbf{m}_i} \right)^2, \\
&\geq \frac{1}{2\|\mathbf{m}\|^2} \sum_{i=1}^n \sum_{j=1}^n \max([\ell']_i^2, [\ell']_j^2) (\mathbf{m}_j - \mathbf{m}_i)^2 \quad \text{same as (C12)}
\end{aligned}$$

By Lemma C.1 of (Cao et al. 2023). We have

$$\begin{aligned}
\left\| \left(\mathbf{I} - \frac{\mathbf{m}\mathbf{m}^T}{\|\mathbf{m}\|^2} \right) \ell' \right\|^2 &\geq \frac{1}{2\|\mathbf{m}\|^2} \frac{1}{4n} \sum_{i=1}^n [\ell']_i^2 \sum_{j=1}^n (\mathbf{m}_j - \mathbf{m}_i)^2 \\
&\geq \frac{1}{4} \left(\sum_{i=1}^n [\ell']_i^2 \right) \left\| \hat{\mathbf{w}} - \frac{\langle \mathbf{w}, \hat{\mathbf{w}} \rangle_{\Sigma}}{\|\mathbf{w}\|_{\Sigma}^2} \mathbf{w} \right\|_{\Sigma}^2 \quad \text{by (C15)} \\
&\geq \frac{\lambda_{\min}}{4} \left(\sum_{i=1}^n [\ell']_i^2 \right) (\rho_t^{\perp})^2 \quad \text{by (C16)}
\end{aligned} \tag{C20}$$

Plug (C20) into (C19) to give

$$\|\nabla_{\mathbf{w}} \mathcal{R}\| \geq \frac{\lambda_{\min}}{2} \frac{\alpha}{\|\mathbf{w}\|_{\Sigma}} \cdot \frac{1}{\sqrt{n}} \|\ell'\| \cdot \rho_t^{\perp}$$

Then we look at $\frac{1}{\sqrt{n}} \|\ell'\|$

$$\begin{aligned}
\frac{1}{n} \|\ell'\|^2 &= \frac{1}{n} \sum_{i=1}^n \left[1 + \exp \left(\frac{\alpha \cdot \langle \mathbf{w}, \mathbf{x}_i y_i \rangle}{\|\mathbf{w}\|_{\Sigma}} \right) \right]^{-2} \\
&\geq \frac{1}{n} \sum_{i=1}^n \left[1 + \exp \left(\frac{\alpha \cdot |\langle \mathbf{w}, \mathbf{x}_i y_i \rangle|}{\|\mathbf{w}\|_{\Sigma}} \right) \right]^{-2} \quad \text{since } (\ell'(\cdot))^2 \text{ is decreasing} \\
&\geq \left[1 + \exp \left(\alpha \cdot \frac{1}{n} \sum_{i=1}^n \frac{|\langle \mathbf{w}, \mathbf{x}_i y_i \rangle|}{\|\mathbf{w}\|_{\Sigma}} \right) \right]^{-2} \quad \text{since } (\ell'(\cdot))^2 \text{ is convex in } [0, \infty) \\
&\geq [1 + \exp(\alpha)]^{-2} \\
&\geq \exp(-2\alpha)/4.
\end{aligned}$$

□

Lemma 12 Let ℓ be ℓ_{\log} and $\hat{\mathbf{w}}$ be the SVM solution as defined in Definition 1. Suppose Assumption 1 holds and $\lambda_{\max} \geq 1$. Consider the gradient descent (1), for any $\tan_{\min} > 0$, if and $\alpha_0 \leq \frac{1}{3} \log(\lambda_{\max})$,

$$\frac{\eta}{\|\mathbf{w}_0\|^2} \geq C_1 \cdot \left(1 + \frac{1}{\tan_{\min}^2} \right) \gamma \quad \text{and} \quad \eta_{\alpha} \leq C_3 \cdot \frac{\tan_{\min}^2}{1 + \tan_{\min}^2} \frac{\|\mathbf{w}_0\|^2}{\eta \gamma},$$

then there exists $t_0 < T_0$ such that $(\rho_{t_0}^\perp / \rho_{t_0})^2 \leq \tan_{\min}^2$, and for any $t < T_0$, we have $\frac{1}{2}\alpha_0 \leq \alpha_t \leq \frac{3}{2}\alpha_0$, where C_1, C_2, C_3 are some constants and

$$T_0 := C_2 \cdot \left(1 + \frac{1}{\tan_{\min}^2}\right) \frac{\eta\gamma}{\|\mathbf{w}_0\|^2}; \quad C := \frac{32\sqrt{\lambda_{\max}} \exp(3\alpha_0/2)}{\sqrt{\lambda_{\min}} \alpha_0};$$

$$C_1 := 2C \cdot \left(1 + 36 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} e^{2\alpha_0} (\rho_0^{\perp, \Sigma})^{-2}\right); \quad C_2 := 2C \cdot \alpha_0^2 \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{3/2}; \quad C_3 := \frac{\alpha_0}{2C_2}.$$

We mention that $\rho_0^{\perp, \Sigma}$ is quantity related to initial parameter, which is defined as

$$\rho_0^{\perp, \Sigma} = \left\| \hat{\mathbf{w}} - \frac{\langle \mathbf{w}_0, \hat{\mathbf{w}} \rangle_{\Sigma}}{\|\mathbf{w}_0\|_{\Sigma}^2} \mathbf{w}_0 \right\|_{\Sigma}.$$

Proof Given any $T_0 > 0$, we choose a small enough η_α to control the growth of α_t . We have

$$|\alpha_{t+1} - \alpha_t| = \left| -\eta_\alpha \frac{\partial \mathcal{R}}{\partial \alpha}(\mathbf{w}_t, \alpha_t) \right| = \frac{\eta_\alpha}{n \|\mathbf{w}_t\|_{\Sigma}} \left| \mathbf{w}_t^T \tilde{\mathbf{X}} \ell' \left(\frac{\alpha_t \tilde{\mathbf{X}}^T \mathbf{w}_t}{\|\mathbf{w}_t\|_{\Sigma}} \right) \right|$$

$$\leq \frac{\eta_\alpha}{n} \|\ell'_t\| \frac{\|\tilde{\mathbf{X}}^T \mathbf{w}_t\|}{\|\mathbf{w}_t\|_{\Sigma}} \leq \frac{\eta_\alpha}{n} \|\mathbf{1}\| \frac{\|\tilde{\mathbf{X}}^T \mathbf{w}_t\|}{\|\mathbf{w}_t\|_{\Sigma}} = \eta_\alpha. \quad \text{since } |\ell'(\cdot)| \leq 1$$

If $\eta_\alpha \leq \frac{\alpha_0}{2T_0}$, for any $t \in [0, T_0)$, we have

$$|\alpha_t - \alpha_0| \leq t\eta_\alpha \leq T_0\eta_\alpha \leq \frac{\alpha_0}{2}, \quad \frac{1}{2}\alpha_0 \leq \alpha_t \leq \frac{3}{2}\alpha_0$$

Then, we calculate the bounds of $\|\mathbf{w}_{T_0}\|$ and $\|\mathbf{w}_1\|$. By Lemma 11 (refer to the version in Appendix), we have the lower bound of $\|\mathbf{w}_1\|^2$

$$\begin{aligned} \|\mathbf{w}_1\|^2 &= \|\mathbf{w}_0\|^2 + \eta^2 \left\| \frac{\partial \mathcal{R}}{\partial \mathbf{w}}(\mathbf{w}_0, \alpha_0) \right\|^2 \\ &\geq \|\mathbf{w}_0\|^2 + \frac{\eta^2}{16} \lambda_{\min} \frac{1}{\|\mathbf{w}_0\|_{\Sigma}^2} \frac{\alpha_0^2}{e^{2\alpha_0}} (\rho_0^{\perp, \Sigma})^2 \\ &\geq \|\mathbf{w}_0\|^2 + \frac{\eta^2}{16} \frac{\lambda_{\min}}{\lambda_{\max}} \frac{1}{\|\mathbf{w}_0\|^2} \frac{\alpha_0^2}{e^{2\alpha_0}} (\rho_0^{\perp, \Sigma})^2 \\ &\geq \frac{\eta^2}{16} \frac{\lambda_{\min}}{\lambda_{\max}} \frac{1}{\|\mathbf{w}_0\|^2} \frac{\alpha_0^2}{e^{2\alpha_0}} (\rho_0^{\perp, \Sigma})^2 \end{aligned}$$

And we apply the first upper bound of gradient in Lemma 11 to obtain the upper bound of $\|\mathbf{w}_{T_0}\|$:

$$\|\mathbf{w}_1\|^2 = \|\mathbf{w}_0\|^2 + \eta^2 \left\| \frac{\partial \mathcal{R}}{\partial \mathbf{w}}(\mathbf{w}_0, \alpha_0) \right\|^2 \leq \|\mathbf{w}_0\|^2 + \frac{\eta^2 \lambda_{\max}}{\lambda_{\min}} \frac{\alpha_0^2}{\|\mathbf{w}_0\|^2},$$

and the upper bound of $\|\mathbf{w}_1\|^2$:

$$\begin{aligned}
\|\mathbf{w}_{T_0}\|^2 &= \|\mathbf{w}_1\|^2 + \eta^2 \sum_{\tau=1}^{T_0-1} \left\| \frac{\partial \mathcal{R}}{\partial \mathbf{w}}(\mathbf{w}_\tau, \alpha_\tau) \right\|^2 \\
&\leq \|\mathbf{w}_1\|^2 + \frac{\eta^2 \lambda_{\max}}{\lambda_{\min}} \sum_{\tau=1}^{T_0-1} \frac{\alpha_\tau^2}{\|\mathbf{w}_\tau\|^2} \\
&\leq \|\mathbf{w}_1\|^2 + (T_0 - 1) \eta^2 \frac{\lambda_{\max}}{\lambda_{\min}} \max_{\tau \in [1, T_0)} \{\alpha_\tau^2\} \frac{1}{\|\mathbf{w}_1\|^2} \\
&\leq \|\mathbf{w}_1\|^2 + \frac{9}{4} \frac{\alpha_0^2 \lambda_{\max}}{\lambda_{\min}} \frac{1}{\|\mathbf{w}_1\|^2} T_0 \eta^2.
\end{aligned}$$

Recall the bound in Lemma 10, we have

$$-\left\langle \hat{\mathbf{w}}, \frac{\partial \mathcal{R}}{\partial \mathbf{w}}(\mathbf{w}, \alpha) \right\rangle \geq \lambda_{\min} \frac{\alpha}{\|\mathbf{w}\|_{\Sigma}} \frac{e^{-\alpha}}{8} \left(\rho^\perp(\mathbf{w}) \right)^2$$

Therefore, by gradient descent update, we have for any $t \in [0, T_0]$

$$\begin{aligned}
\langle \mathbf{w}_{t+1}, \hat{\mathbf{w}} \rangle - \langle \mathbf{w}_t, \hat{\mathbf{w}} \rangle &= -\eta \left\langle \frac{\partial \mathcal{R}}{\partial \mathbf{w}}(\mathbf{w}_t, \alpha_t), \hat{\mathbf{w}} \right\rangle \\
&\geq \eta \frac{\lambda_{\min}}{8} \frac{1}{\|\mathbf{w}_t\|_{\Sigma}} \frac{\alpha_t}{e^{\alpha_t}} \left(\rho_t^\perp \right)^2 \\
&\geq \eta \frac{\lambda_{\min}}{8\sqrt{\lambda_{\max}}} \frac{1}{\|\mathbf{w}_t\|} \frac{\alpha_t}{e^{\alpha_t}} \left(\rho_t^\perp \right)^2 \\
&\geq \eta \frac{\alpha_0}{16 \exp(\frac{3}{2}\alpha_0)} \frac{\lambda_{\min}}{\sqrt{\lambda_{\max}}} \frac{1}{\|\mathbf{w}_t\|} \left(\rho_t^\perp \right)^2
\end{aligned}$$

The expression can be rearranged to obtain

$$\left(\rho_t^\perp \right)^2 \leq \frac{16\sqrt{\lambda_{\max}}}{\lambda_{\min}} \frac{\exp(3\alpha_0/2)}{\alpha_0} \frac{\|\mathbf{w}_t\|}{\eta} (\langle \mathbf{w}_{t+1}, \hat{\mathbf{w}} \rangle - \langle \mathbf{w}_t, \hat{\mathbf{w}} \rangle)$$

Further, we have

$$\begin{aligned}
\min_{\tau \in [0, T_0)} \left(\rho_\tau^\perp \right)^2 &\leq \frac{1}{T_0} \sum_{\tau=0}^{T_0-1} \left(\rho_\tau^\perp \right)^2 \\
&\leq \frac{16\sqrt{\lambda_{\max}}}{\lambda_{\min}} \frac{\exp(3\alpha_0/2)}{\alpha_0} \frac{1}{\eta} \sum_{\tau=0}^{T_0-1} \|\mathbf{w}_\tau\| (\langle \mathbf{w}_{\tau+1}, \hat{\mathbf{w}} \rangle - \langle \mathbf{w}_\tau, \hat{\mathbf{w}} \rangle) \\
&\leq \frac{16\sqrt{\lambda_{\max}}}{\lambda_{\min}} \frac{\exp(3\alpha_0/2)}{\alpha_0} \max_{\tau \in [0, T_0)} \{\|\mathbf{w}_\tau\|\} \frac{1}{\eta} \sum_{\tau=0}^{T_0-1} (\langle \mathbf{w}_{\tau+1}, \hat{\mathbf{w}} \rangle - \langle \mathbf{w}_\tau, \hat{\mathbf{w}} \rangle) \\
&= \frac{16\sqrt{\lambda_{\max}}}{\lambda_{\min}} \frac{\exp(3\alpha_0/2)}{\alpha_0} \max_{\tau \in [0, T_0)} \{\|\mathbf{w}_\tau\|\} \frac{1}{T_0 \eta} (\langle \mathbf{w}_{T_0}, \hat{\mathbf{w}} \rangle - \langle \mathbf{w}_0, \hat{\mathbf{w}} \rangle)
\end{aligned}$$

By Cauchy inequalities, we have

$$\begin{aligned}
\min_{\tau \in [0, T_0]} (\rho_\tau^\perp)^2 &\leq \frac{16\sqrt{\lambda_{\max}}}{\gamma\lambda_{\min}} \frac{\exp(3\alpha_0/2)}{\alpha_0} \max_{\tau \in [0, T_0]} \{\|\mathbf{w}_\tau\|\} \frac{1}{T_0\eta} (\|\mathbf{w}_0\| + \|\mathbf{w}_{T_0}\|) \\
&\leq \frac{32\sqrt{\lambda_{\max}}}{\gamma\lambda_{\min}} \frac{\exp(3\alpha_0/2)}{\alpha_0} \max_{\tau \in [0, T_0]} \{\|\mathbf{w}_\tau\|\} \frac{\|\mathbf{w}_{T_0}\|}{T_0\eta} \\
&\leq \frac{32\sqrt{\lambda_{\max}}}{\lambda_{\min}} \frac{\exp(3\alpha_0/2)}{\alpha_0} \frac{\|\mathbf{w}_{T_0}\|^2}{T_0\eta\gamma} \\
&\leq \frac{32}{\alpha_0} \frac{\lambda_{\max}}{\lambda_{\min}} \frac{\|\mathbf{w}_{T_0}\|^2}{T_0\eta\gamma} \quad \text{by } \alpha_0 \leq \frac{1}{3} \log(\lambda_{\max}) \\
&= C \frac{\|\mathbf{w}_{T_0}\|^2}{T_0\eta\gamma} \quad \text{for some constant } C,
\end{aligned}$$

where we denote $C = \frac{32}{\alpha_0} \frac{\lambda_{\max}}{\lambda_{\min}}$ for simplicity. Now, let us check $\frac{\|\mathbf{w}_{T_0}\|^2}{T_0\eta\gamma}$. Recall just derived bounds of $\|\mathbf{w}_{T_0}\|^2$ and $\|\mathbf{w}_1\|^2$, we have

$$\begin{aligned}
\frac{\|\mathbf{w}_{T_0}\|^2}{T_0\eta\gamma} &\leq \frac{1}{T_0\eta\gamma} \left(\|\mathbf{w}_1\|^2 + \frac{9}{4} \frac{\alpha_0^2 \lambda_{\max}}{\lambda_{\min}} \frac{1}{\|\mathbf{w}_1\|^2} T_0 \eta^2 \right) \\
&= \frac{\|\mathbf{w}_1\|^2}{T_0\eta\gamma} + \frac{9}{4} \frac{\alpha_0^2 \lambda_{\max}}{\lambda_{\min}} \frac{\eta}{\|\mathbf{w}_1\|^2 \gamma} \\
&\leq \frac{1}{T_0\eta\gamma} \left(\|\mathbf{w}_0\|^2 + \frac{\eta^2 \lambda_{\max}}{\lambda_{\min}} \frac{\alpha_0^2}{\|\mathbf{w}_0\|^2} \right) + \frac{9}{4} \frac{\alpha_0^2 \lambda_{\max}}{\lambda_{\min}} \frac{\eta}{\gamma} \left(\frac{\eta^2}{16} \frac{\lambda_{\min}}{\lambda_{\max}} \frac{1}{\|\mathbf{w}_0\|^2} \frac{\alpha_0^2}{e^{2\alpha_0}} (\rho_0^\perp, \Sigma)^2 \right)^{-1} \\
&= \|\mathbf{w}_0\|^2 \frac{1}{T_0\eta\gamma} + \frac{\lambda_{\max}}{\lambda_{\min}} \frac{\alpha_0^2}{\|\mathbf{w}_0\|^2} \frac{\eta}{T_0\gamma} + 36 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} \|\mathbf{w}_0\|^2 e^{2\alpha_0} \frac{1}{(\rho_0^\perp, \Sigma)^2} \frac{1}{\eta\gamma} \\
&\leq \|\mathbf{w}_0\|^2 \frac{1}{\eta\gamma} + \frac{\lambda_{\max}}{\lambda_{\min}} \frac{\alpha_0^2}{\|\mathbf{w}_0\|^2} \frac{\eta}{T_0\gamma} + 36 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} e^{2\alpha_0} \frac{1}{(\rho_0^\perp, \Sigma)^2} \frac{\|\mathbf{w}_0\|^2}{\eta\gamma} \\
&= \alpha_0^2 \frac{\lambda_{\max}}{\lambda_{\min}} \frac{\eta}{\|\mathbf{w}_0\|^2} \frac{1}{T_0\gamma} + \left(1 + 36 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} e^{2\alpha_0} \frac{1}{(\rho_0^\perp, \Sigma)^2} \right) \frac{\|\mathbf{w}_0\|^2}{\eta\gamma},
\end{aligned}$$

where the last inequality is since $T_0 \geq 1$. A large T_0 can make the second terms arbitrarily small and a large η can make the first and the last term arbitrarily small. Therefore, given any $\rho_{\min}^\perp > 0$, we have

$$\begin{aligned}
\min_{\tau \in [0, T_0]} (\rho_\tau^\perp)^2 &\leq (\rho_{\min}^\perp)^2 \Leftarrow C \cdot \frac{\|\mathbf{w}_{T_0}\|^2}{T_0\eta\gamma} \leq (\rho_{\min}^\perp)^2 \\
&\Leftarrow \frac{\eta}{\|\mathbf{w}_0\|^2} \geq C_1 \cdot \frac{1}{(\rho_{\min}^\perp)^2} \frac{1}{\gamma} \quad \text{and} \quad T_0 \geq C_2 \cdot \frac{\eta}{\|\mathbf{w}_0\|^2} \cdot \frac{1}{(\rho_{\min}^\perp)^2} \frac{1}{\gamma} \quad \text{for some constants } C_1, C_2
\end{aligned}$$

where C_1, C_2 is some constants related to the initial parameter and data:

$$C_1 := 2C \cdot \left(1 + 36 \frac{\lambda_{\max}^2}{\lambda_{\min}^2} e^{2\alpha_0} (\rho_0^\perp, \Sigma)^{-2} \right); \quad C_2 := 2C \cdot \alpha_0^2 \frac{\lambda_{\max}}{\lambda_{\min}}.$$

To obtain the bound for ρ_t^\perp / ρ_t , we set $(\rho_{\min}^\perp)^2 = \tan_{\min}^2 / (\gamma^2 \cdot (1 + \tan_{\min}^2))$. Using $1/\gamma^2 = (\rho_t^\perp)^2 + \rho_t^2$, if $\rho_t^\perp \leq \rho_{\min}^\perp$ as defined above, then $\rho_t^\perp / \rho_t \leq \tan_{\min}$. Therefore, we have

$$\begin{aligned} \min_{\tau \in [0, T_0]} (\rho_\tau^\perp / \rho_\tau)^2 &\leq \tan_{\min}^2 \\ \Leftrightarrow \frac{\eta}{\|\mathbf{w}_0\|^2} &\geq C_1 \cdot \left(1 + \frac{1}{\tan_{\min}^2}\right) \gamma \text{ and } T_0 \geq C_2 \cdot \left(1 + \frac{1}{\tan_{\min}^2}\right) \frac{\eta\gamma}{\|\mathbf{w}_0\|^2}. \end{aligned}$$

Finally, remember that all the sufficient conditions above hold under the condition of $\eta_\alpha \leq \frac{1}{2T_0} \alpha_0$. Therefore, we have the condition for η_α

$$\eta_\alpha \leq \frac{1}{2T_0} \alpha_0 = C_3 \cdot \frac{\tan_{\min}^2}{1 + \tan_{\min}^2} \frac{\|\mathbf{w}_0\|^2}{\eta\gamma}, \text{ where } C_3 := \frac{\alpha_0}{2C_2}$$

□

Theorem 6 Let ℓ be ℓ_{\log} and $\hat{\mathbf{w}}$ be the SVM solution as defined in Definition 1. Suppose Assumption 1 holds and $\lambda_{\max} \geq 1$. There exists constants $C_1, C_2, C_3, C_4, C_5, C_6, \tilde{C}, \Phi$ such that if $\alpha_0 \leq \frac{1}{3} \log(\lambda_{\max})$ and

$$\max \left(C_1 \frac{16\lambda_{\max}^2}{\lambda_{\min}^2}, C_4 \right) \cdot \gamma \leq \frac{\eta}{\|\mathbf{w}_0\|^2} \leq C_5 \cdot \gamma^{-1} \text{ and } \eta_\alpha \leq C_3 \cdot \frac{\lambda_{\min}^2}{16\lambda_{\max}^2} \cdot \frac{\|\mathbf{w}_0\|^2}{\eta\gamma},$$

it holds that during $t \in [0, T_0]$

- **Monotonic Decrease of ρ_t^\perp / ρ_t** : ρ_t^\perp / ρ_t keeps decreasing as long as $(\rho_t^\perp / \rho_t)^2 \geq 4/(\sqrt{\Phi^2 + 4} - \Phi)^2 - 1$ and $\rho_t > 0$;
- **Occurrence of Spike**: $\exists t_0$ such that $(\rho_{t_0}^\perp / \rho_{t_0})^2 \leq \tan_{\min}^2$, and at this time a spike occurs;
- **Peak of Rising Edge**: the peak of $(\rho_t^\perp / \rho_t)^2$ in the Rising Edge is at most $C_6 \cdot \frac{\eta^2}{\|\mathbf{w}_0\|^4} \cdot \frac{1}{\gamma^2} - 1$.

We mention T_0 and \tan_{\min} are defined as:

$$T_0 = C_2 \cdot \left(1 + \frac{1}{\tan_{\min}^2}\right) \cdot \frac{\eta\gamma}{\|\mathbf{w}_0\|^2}, \quad \tan_{\min}^2 = \frac{\gamma^2 \lambda_{\min}}{8\lambda_{\max}^2}.$$

And C_1, C_2, C_3 are defined in Lemma 12 and

$$\begin{aligned} \Phi &:= \frac{6\lambda_{\max}^2}{\lambda_{\min}^2} \alpha_0 \cdot \frac{\eta\gamma}{\|\mathbf{w}_0\|^2}; \quad \tilde{C} := \frac{3}{256} \frac{\lambda_{\min}}{\lambda_{\max}} \alpha_0; \quad C_4 := 2/\tilde{C}; \\ C_5 &:= \sqrt{\frac{\tilde{C}}{2} / \left(36C_2 \frac{\alpha_0^2 \lambda_{\max}^3}{\lambda_{\min}^3}\right)}, \quad C_6 := \left(6 \frac{\lambda_{\max}^{5/2}}{\lambda_{\min}^{5/2}} \alpha_0 e^{3\alpha_0/2} \left(\frac{3}{2} \alpha_0 + 1\right)^2\right)^2. \end{aligned}$$

Proof We observe that $\frac{\lambda_{\min}^2}{8\lambda_{\max}^2} \leq \tan_{\min}^2 \leq 1/4 < 1$ since $\lambda_{\min} \leq \gamma^2 \leq \lambda_{\max}$. Then, by the lower bound of $\frac{\eta}{\|\mathbf{w}_0\|^2}$, we have

$$\frac{\eta}{\|\mathbf{w}_0\|^2} \geq C_1 \frac{16\lambda_{\max}^2}{\lambda_{\min}^2} \gamma \geq \frac{2C_1}{\tan_{\min}^2} \gamma \geq C_1 \left(1 + \frac{1}{\tan_{\min}^2}\right) \gamma.$$

Next, we consider the bound for η_α

$$\eta_\alpha \leq C_3 \cdot \frac{\lambda_{\min}^2}{16\lambda_{\max}^2} \cdot \frac{\|\mathbf{w}_0\|^2}{\eta\gamma} \leq C_3 \cdot \frac{\tan_{\min}^2}{2} \cdot \frac{\|\mathbf{w}_0\|^2}{\eta\gamma} \leq C_3 \cdot \frac{\tan_{\min}^2}{1 + \tan_{\min}^2} \cdot \frac{\|\mathbf{w}_0\|^2}{\eta\gamma}$$

Therefore, by Lemma 12, we know there exists a $t_0 < T_0$ such that $(\rho_{t_0}^\perp / \rho_{t_0})^2 \leq \tan_{\min}^2 \leq 1/4$, and for any $t < T_0$, $\frac{1}{2}\alpha_0 \leq \alpha_t \leq \frac{3}{2}\alpha_0$. Based on the above results, we first prove (2), then prove (1) and (3).

Proof of (2) We have two cases to consider, $\rho_{t_0} < 0$ and $\rho_{t_0} > 0$. The former is the relatively simple case. If $\rho_{t_0} < 0$, we have $-\frac{1}{2} \leq \rho_{t_0}^\perp / \rho_{t_0} < 0$. By Lemma 7, it has to be that $\rho_{t_0}^\perp \leq \rho_{t_0+1}^\perp$, which means a spike occurs at t_0 -th iteration. In the left part of this proof, we consider $\rho_{t_0} > 0$, which is much more complicated.

By the second conclusion of Lemma 7, we need to verify the following condition.

$$\frac{\eta}{\|\mathbf{w}_{t_0}\|} \|\nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}_{t_0}, \alpha_{t_0})\| \geq \frac{2\rho_{t_0}\rho_{t_0}^\perp}{\rho_{t_0}^2 - (\rho_{t_0}^\perp)^2} \quad (\text{C21})$$

By Lemma 11, we have

$$\begin{aligned} (\text{C21}) &\Leftarrow \frac{1}{4} \lambda_{\min} \frac{\alpha_{t_0}}{e^{\alpha_{t_0}}} \frac{\eta}{\|\mathbf{w}_{t_0}\| \|\mathbf{w}_{t_0}\|_{\Sigma}} \rho_{t_0}^\perp \geq \frac{2\rho_{t_0}\rho_{t_0}^\perp}{\rho_{t_0}^2 - (\rho_{t_0}^\perp)^2} \\ &\Leftrightarrow \frac{1}{8} \lambda_{\min} \frac{\alpha_{t_0}}{e^{\alpha_{t_0}}} \left(1 - \left(\frac{\rho_{t_0}^\perp}{\rho_{t_0}}\right)^2\right) \geq \frac{\|\mathbf{w}_{t_0}\| \|\mathbf{w}_{t_0}\|_{\Sigma}}{\eta\rho_{t_0}}, \quad \text{rearrange expression} \end{aligned} \quad (\text{C22})$$

Since $1/2\alpha_0 \leq \alpha_t \leq 3/2\alpha_0 \forall t < T_0$ and $\rho_{t_0}^\perp / \rho_{t_0} \leq \frac{1}{2}$, we have

$$\begin{aligned} (\text{C22}) &\Leftarrow \frac{1}{16} \lambda_{\min} \frac{\alpha_0}{e^{3\alpha_0/2}} \left(1 - \left(\frac{\rho_{t_0}^\perp}{\rho_{t_0}}\right)^2\right) \geq \frac{\|\mathbf{w}_{t_0}\| \|\mathbf{w}_{t_0}\|_{\Sigma}}{\eta\rho_{t_0}} \\ &\Leftarrow \frac{3}{64} \lambda_{\min} \frac{\alpha_0}{e^{3\alpha_0/2}} \geq \frac{\|\mathbf{w}_{t_0}\| \|\mathbf{w}_{t_0}\|_{\Sigma}}{\eta\rho_{t_0}} \end{aligned} \quad (\text{C23})$$

Note that by Lemma 9, we have

$$\begin{aligned} \frac{\|\mathbf{w}_{t_0}\|_{\Sigma}}{\|\mathbf{w}_{t_0}\|} &\leq \gamma^2 \rho_{t_0} + 2\sqrt{2} \cdot \lambda_{\max} \cdot \gamma \frac{\|\mathbf{w}_{t_0}\|}{\|\mathbf{w}_{t_0}\|_{\Sigma}} \rho_{t_0}^\perp \\ &\leq \gamma^2 \rho_{t_0} + 2\sqrt{2} \cdot \frac{\lambda_{\max}}{\sqrt{\lambda_{\min}}} \cdot \gamma \rho_{t_0}^\perp \\ &\leq 2\gamma^2 \rho_{t_0} \quad \text{by } \rho_{t_0}^\perp / \rho_{t_0} \leq \frac{\gamma\sqrt{\lambda_{\min}}}{2\sqrt{2} \cdot \lambda_{\max}} \end{aligned}$$

Therefore, we have

$$(\text{C23}) \Leftarrow \frac{3}{256} \lambda_{\min} \frac{\alpha_0}{e^{3\alpha_0/2}} \geq \frac{\gamma^2}{\eta} \|\mathbf{w}_{t_0}\|^2 \quad (\text{C24})$$

Since $\|\mathbf{w}_{T_0}\| \geq \|\mathbf{w}_t\| \forall t < T_0$, we have

$$(C24) \Leftarrow \frac{3}{256} \lambda_{\min} \frac{\alpha_0}{e^{3\alpha_0/2}} \geq \frac{\gamma^2}{\eta} \|\mathbf{w}_{T_0}\|^2 \quad (C25)$$

For $\|\mathbf{w}_{T_0}\|^2$, we apply the first upper bound of gradient in Lemma 11 to obtain

$$\begin{aligned} \|\mathbf{w}_{T_0}\|^2 &= \|\mathbf{w}_0\|^2 + \eta^2 \sum_{\tau=0}^{T_0-1} \|\nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}_\tau, \alpha_\tau)\|^2 \\ &\leq \|\mathbf{w}_0\|^2 + \frac{\eta^2 \lambda_{\max}}{\lambda_{\min}} \sum_{\tau=0}^{T_0-1} \frac{\alpha_\tau^2}{\|\mathbf{w}_\tau\|^2} \\ &\leq \|\mathbf{w}_0\|^2 + T_0 \eta^2 \frac{\lambda_{\max}}{\lambda_{\min}} \max_{\tau \in [0, T_0)} \left\{ \alpha_\tau^2 \right\} \frac{1}{\|\mathbf{w}_0\|^2} \quad \text{since } \|\mathbf{w}_0\| \leq \|\mathbf{w}_\tau\| \forall \tau \geq 0 \\ &\leq \|\mathbf{w}_0\|^2 + \frac{9}{4} \frac{\alpha_0^2 \lambda_{\max}}{\lambda_{\min}} \frac{1}{\|\mathbf{w}_0\|^2} T_0 \eta^2 \quad \text{by } \alpha_\tau \leq \frac{3}{2} \alpha_0 \forall \tau < T_0 \end{aligned}$$

Therefore, it holds that

$$\begin{aligned} (C25) &\Leftarrow \frac{3}{256} \lambda_{\min} \frac{\alpha_0}{e^{3\alpha_0/2}} \geq \frac{\gamma^2}{\eta} \left(\|\mathbf{w}_0\|^2 + \frac{9}{4} \frac{\alpha_0^2 \lambda_{\max}}{\lambda_{\min}} \frac{1}{\|\mathbf{w}_0\|^2} T_0 \eta^2 \right) \\ &\Leftrightarrow \frac{3}{256} \lambda_{\min} \frac{\alpha_0}{e^{3\alpha_0/2}} \geq \gamma^2 \frac{\|\mathbf{w}_0\|^2}{\eta} + \frac{9}{4} \frac{\alpha_0^2 \lambda_{\max}}{\lambda_{\min}} \frac{1}{\|\mathbf{w}_0\|^2} T_0 \eta^2 \\ &\Leftrightarrow \frac{3}{256} \lambda_{\min} \frac{\alpha_0}{e^{3\alpha_0/2}} \geq \gamma^2 \frac{\|\mathbf{w}_0\|^2}{\eta} + \frac{9}{4} C_2 \frac{\alpha_0^2 \lambda_{\max}}{\lambda_{\min}} \left(1 + \frac{1}{\tan_{\min}^2} \right) \frac{\eta^2 \gamma^3}{\|\mathbf{w}_0\|^4} \\ &\quad \text{recall } T_0 = C_2 \cdot \left(1 + \frac{1}{\tan_{\min}^2} \right) \cdot \frac{\eta \gamma}{\|\mathbf{w}_0\|^2} \\ &\Leftarrow \frac{3}{256} \lambda_{\min} \frac{\alpha_0}{e^{3\alpha_0/2}} \geq \gamma^2 \frac{\|\mathbf{w}_0\|^2}{\eta} + \frac{9}{4} C_2 \frac{\alpha_0^2 \lambda_{\max}}{\lambda_{\min}} \left(1 + \frac{8\lambda_{\max}^2}{\lambda_{\min}^2} \right) \frac{\eta^2 \gamma^3}{\|\mathbf{w}_0\|^4} \\ &\quad \text{recall } \tan_{\min}^2 = \frac{\gamma^2 \lambda_{\min}}{8\lambda_{\max}^2} \geq \frac{\lambda_{\min}^2}{8\lambda_{\max}^2} \\ &\Leftarrow \frac{3}{256} \lambda_{\min} \frac{\alpha_0}{e^{3\alpha_0/2}} \geq \gamma^2 \frac{\|\mathbf{w}_0\|^2}{\eta} + 36C_2 \frac{\alpha_0^2 \lambda_{\max}^3}{\lambda_{\min}^3} \frac{\eta^2 \gamma^3}{\|\mathbf{w}_0\|^4} \\ &\Leftarrow \frac{3}{256} \frac{\lambda_{\min}}{\sqrt{\lambda_{\max}}} \alpha_0 \geq \gamma^2 \frac{\|\mathbf{w}_0\|^2}{\eta} + 36C_2 \frac{\alpha_0^2 \lambda_{\max}^3}{\lambda_{\min}^3} \frac{\eta^2 \gamma^3}{\|\mathbf{w}_0\|^4} \quad \text{by } \alpha_0 \leq \frac{1}{3} \log(\lambda_{\max}) \\ &\Leftrightarrow \frac{3}{256} \frac{\lambda_{\min}}{\lambda_{\max}} \frac{\sqrt{\lambda_{\max}}}{\gamma} \alpha_0 \geq \gamma \frac{\|\mathbf{w}_0\|^2}{\eta} + 36C_2 \frac{\alpha_0^2 \lambda_{\max}^3}{\lambda_{\min}^3} \frac{\eta^2 \gamma^2}{\|\mathbf{w}_0\|^4} \\ &\Leftarrow \frac{3}{256} \frac{\lambda_{\min}}{\lambda_{\max}} \alpha_0 \geq \gamma \frac{\|\mathbf{w}_0\|^2}{\eta} + 36C_2 \frac{\alpha_0^2 \lambda_{\max}^3}{\lambda_{\min}^3} \frac{\eta^2 \gamma^2}{\|\mathbf{w}_0\|^4} \quad \text{by } \gamma \leq \sqrt{\lambda_{\max}} \end{aligned} \quad (C26)$$

Then, we have

$$(C21) \Leftarrow \dots \Leftarrow (C26) \Leftarrow C_4 \cdot \gamma \leq \frac{\eta}{\|\mathbf{w}_0\|^2} \leq C_5 \cdot \gamma^{-1}, \quad \text{for some constants } C_4 \text{ and } C_5$$

where C_4 and C_5 are defined as

$$\tilde{C} := \frac{3}{256} \frac{\lambda_{\min}}{\lambda_{\max}} \alpha_0; \quad C_4 := 2/\tilde{C}; \quad C_5 := \sqrt{\frac{\tilde{C}}{2} / \left(36C_2 \frac{\alpha_0^2 \lambda_{\max}^3}{\lambda_{\min}^3} \right)}.$$

Proof of (1) Next, we prove, during $t \in [0, T_0)$, ρ_t^\perp keeps decreasing until is smaller than $\left(1 - \left(\sqrt{\Phi^2 + 4} - \Phi\right)^2 / 4\right) \cdot \frac{1}{\gamma^2}$. To see if ρ_t^\perp is decreasing, we need to verify the convergence condition in Lemma 7 during $t \in [0, T_0)$:

$$\frac{\eta \rho_t}{\|\mathbf{w}_t\|} \|\nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}_t, \alpha_t)\|^2 \leq -2 \langle \hat{\mathbf{w}}, \nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}_t, \alpha_t) \rangle \quad (\text{C27})$$

By Lemma 10 and 11, we have

$$\begin{aligned} -\langle \hat{\mathbf{w}}, \nabla_{\mathbf{w}} \mathcal{R}_t \rangle &\geq \frac{\lambda_{\min}}{8\sqrt{\lambda_{\max}}} \frac{\alpha_t e^{-\alpha_t}}{\|\mathbf{w}_t\|} \left(\rho_t^\perp\right)^2; \\ \|\nabla_{\mathbf{w}} \mathcal{R}_t\| &\leq \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \frac{\alpha_t}{\|\mathbf{w}_t\|}. \end{aligned}$$

By above bounds, we have

$$\begin{aligned} (\text{C27}) &\Leftarrow \frac{\lambda_{\max}}{\lambda_{\min}} \frac{\eta \rho_t}{\|\mathbf{w}_t\|} \frac{\alpha_t^2}{\|\mathbf{w}_t\|^2} \leq \frac{\lambda_{\min}}{4} \frac{1}{\|\mathbf{w}_t\|_{\Sigma}} \frac{\alpha_t}{e^{\alpha_t}} \left(\rho_t^\perp\right)^2 \\ &\Leftarrow \frac{\lambda_{\max}}{\lambda_{\min}} \frac{\eta \rho_t}{\|\mathbf{w}_t\|} \frac{\alpha_t^2}{\|\mathbf{w}_t\|^2} \leq \frac{\lambda_{\min}}{4\sqrt{\lambda_{\max}}} \frac{1}{\|\mathbf{w}_t\|} \frac{\alpha_t}{e^{\alpha_t}} \left(\rho_t^\perp\right)^2 \\ &\Leftrightarrow \frac{4\lambda_{\max}^{3/2}}{\lambda_{\min}^2} \frac{\eta}{\|\mathbf{w}_t\|^2} \alpha_t e^{\alpha_t} \rho_t \leq \left(\rho_t^\perp\right)^2 \\ &\Leftarrow \frac{6\lambda_{\max}^{3/2}}{\lambda_{\min}^2} \frac{\eta}{\|\mathbf{w}_t\|^2} \alpha_0 e^{3\alpha_0/2} \rho_t \leq \left(\rho_t^\perp\right)^2 \quad \text{since } \frac{1}{2}\alpha_0 \leq \alpha_0 \leq \frac{3}{2}\alpha_0 \forall t < T_0 \\ &\Leftarrow \frac{6\lambda_{\max}^{3/2}}{\lambda_{\min}^2} \frac{\eta}{\|\mathbf{w}_0\|^2} \alpha_0 e^{3\alpha_0/2} \rho_t \leq \left(\rho_t^\perp\right)^2 \quad \text{since } \|\mathbf{w}_0\| \leq \|\mathbf{w}_t\| \forall t \geq 0 \\ &\Leftarrow \frac{6\lambda_{\max}^2}{\lambda_{\min}^2} \frac{\eta}{\|\mathbf{w}_0\|^2} \alpha_0 \rho_t \leq \left(\rho_t^\perp\right)^2 \quad \text{by } \alpha_0 \leq \frac{1}{3} \log(\lambda_{\max}) \end{aligned} \quad (\text{C28})$$

Then we denote

$$\Phi := \frac{6\lambda_{\max}^2}{\lambda_{\min}^2} \alpha_0 \cdot \frac{\eta \gamma}{\|\mathbf{w}_0\|^2},$$

to simplify the notation. Further, we have

$$(\text{C28}) \Leftrightarrow \frac{\Phi}{\gamma} \rho_t \leq \left(\rho_t^\perp\right)^2 \Leftrightarrow \rho_t^2 + \frac{\Phi}{\gamma} \rho_t - \frac{1}{\gamma^2} \leq 0$$

Since $\rho_t > 0$, we solve the range of ρ_t :

$$\rho_t \leq \frac{\sqrt{\Phi^2 + 4} - \Phi}{2} \cdot \frac{1}{\gamma} \Leftrightarrow \left(\frac{\rho_t^\perp}{\rho_t}\right)^2 \geq \frac{4}{\left(\sqrt{\Phi^2 + 4} - \Phi\right)^2} - 1,$$

where the range of $\frac{\rho_t^\perp}{\rho_t}$ is by $1/\gamma^2 = \left(\rho_t^\perp\right)^2 + \rho_t^2$. Therefore, we have

$$(\text{C27}) \Leftarrow (\text{C28}) \Leftrightarrow \left(\frac{\rho_t^\perp}{\rho_t}\right)^2 \geq \frac{4}{\left(\sqrt{\Phi^2 + 4} - \Phi\right)^2} - 1$$

Proof of (3) We investigate the peak of the spike, that is the largest ρ_t^\perp during *Rising Edge*, by checking convergence condition of Lemma 7. Result (1), in fact, provides a loose upper

bound of peaks during $t \in [0, T_0)$. To obtain a more precise upper bound of spike peak, we need upper bound $\|\nabla_{\mathbf{w}} \mathcal{R}_t\|$ by ρ_t^\perp , which is the second upper of Lemma 11:

$$\|\nabla_{\mathbf{w}} \mathcal{R}_t\| \leq \lambda_{\max} \frac{\alpha_t}{\|\mathbf{w}_t\|_{\Sigma}} (\alpha_t + 1) \rho_t^\perp$$

Therefore, we have

$$\begin{aligned}
(C27) &\Leftarrow 4 \frac{\lambda_{\max}^{5/2}}{\lambda_{\min}} \frac{\alpha_t e^{\alpha_t}}{\|\mathbf{w}_t\|_{\Sigma}^2} (\alpha_t + 1)^2 \eta \rho_t \leq 1 \\
&\Leftarrow 4 \frac{\lambda_{\max}^{5/2}}{\lambda_{\min}^2} \frac{\alpha_t e^{\alpha_t}}{\|\mathbf{w}_t\|^2} (\alpha_t + 1)^2 \eta \rho_t \leq 1 \\
&\Leftarrow 4 \frac{\lambda_{\max}^{5/2}}{\lambda_{\min}^2} \frac{\alpha_t e^{\alpha_t}}{\|\mathbf{w}_0\|^2} (\alpha_t + 1)^2 \eta \rho_t \leq 1 \quad \text{since } \|\mathbf{w}_t\| \geq \|\mathbf{w}_0\| \quad \forall t > 0 \\
&\Leftarrow 6 \frac{\lambda_{\max}^{5/2}}{\lambda_{\min}^2} \alpha_0 e^{3\alpha_0/2} \left(\frac{3}{2}\alpha_0 + 1\right)^2 \frac{\eta \rho_t}{\|\mathbf{w}_0\|^2} \leq 1 \quad \text{by } \frac{1}{2}\alpha_0 \leq \alpha_t \leq \frac{3}{2}\alpha_0 \quad \forall t \in [0, T_0) \\
&\Leftrightarrow 6 \frac{\lambda_{\max}^{5/2}}{\lambda_{\min}^{5/2}} \alpha_0 e^{3\alpha_0/2} \left(\frac{3}{2}\alpha_0 + 1\right)^2 \frac{\eta \gamma}{\|\mathbf{w}_0\|^2} \rho_t \leq 1 \quad \text{since } \gamma \geq \sqrt{\lambda_{\min}} \\
&\Leftrightarrow \rho_t^2 \leq \frac{\|\mathbf{w}_0\|^4}{\gamma^2 \eta^2} C_6^{-1} \quad \text{for some constant } C_6 \\
&\Leftrightarrow \frac{1}{\gamma^2} - C_6^{-1} \frac{\|\mathbf{w}_0\|^4}{\gamma^2 \eta^2} \leq (\rho_t^\perp)^2
\end{aligned}$$

where C_6 is

$$C_6 := \left(6 \frac{\lambda_{\max}^{5/2}}{\lambda_{\min}^{5/2}} \alpha_0 e^{3\alpha_0/2} \left(\frac{3}{2}\alpha_0 + 1\right)^2 \right)^2$$

□

References

- Ahn K, Zhang J, Sra S (2022) Understanding the unstable convergence of gradient descent. In: International Conference on Machine Learning, PMLR, pp 247–257
- Ahn K, Bubeck S, Chewi S, et al (2023) Learning threshold neurons via edge of stability. Advances in Neural Information Processing Systems 36:19540–19569
- Andriushchenko M, Varre AV, Pillaud-Vivien L, et al (2023) Sgd with large step sizes learns sparse features. In: International Conference on Machine Learning, PMLR, pp 903–925
- Arora S, Li Z, Lyu K (2018) Theoretical analysis of auto rate-tuning by batch normalization. arXiv preprint arXiv:181203981
- Ba JL (2016) Layer normalization. arXiv preprint arXiv:160706450
- Cao Y, Zou D, Li Y, et al (2023) The implicit bias of batch normalization in linear models and two-layer linear convolutional neural networks. In: The Thirty Sixth Annual Conference on Learning Theory, PMLR, pp 5699–5753

- Chen L, Bruna J (2022) On gradient descent convergence beyond the edge of stability. arXiv preprint arXiv:220604172 3
- Chowdhery A, Narang S, Devlin J, et al (2023) Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* 24(240):1–113
- Cohen JM, Kaur S, Li Y, et al (2021) Gradient descent on neural networks typically occurs at the edge of stability. arXiv preprint arXiv:210300065
- Damian A, Nichani E, Lee JD (2022) Self-stabilization: The implicit bias of gradient descent at the edge of stability. arXiv preprint arXiv:220915594
- Even M, Pesme S, Gunasekar S, et al (2023) (s) gd over diagonal linear networks: Implicit bias, large stepsizes and edge of stability. *Advances in Neural Information Processing Systems* 36:29406–29448
- Gunasekar S, Lee JD, Soudry D, et al (2018) Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems* 31
- Hoffer E, Banner R, Golan I, et al (2018) Norm matters: efficient and accurate normalization schemes in deep networks. *Advances in Neural Information Processing Systems* 31
- Ioffe S (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:150203167
- Kreisler I, Nacson MS, Soudry D, et al (2023) Gradient descent monotonically decreases the sharpness of gradient flow solutions in scalar networks and beyond. In: *International Conference on Machine Learning*, PMLR, pp 17684–17744
- Lu M, Wu B, Yang X, et al (2023) Benign oscillation of stochastic gradient descent with large learning rates. arXiv preprint arXiv:231017074
- Ma C, Kunin D, Wu L, et al (2022) Beyond the quadratic approximation: the multi-scale structure of neural network loss landscapes. arXiv preprint arXiv:220411326
- Molybog I, Albert P, Chen M, et al (2023) A theory on adam instability in large-scale machine learning. arXiv preprint arXiv:230409871
- Morwani D, Ramaswamy HG (2022) Inductive bias of gradient descent for weight normalized smooth homogeneous neural nets. In: *International Conference on Algorithmic Learning Theory*, PMLR, pp 827–880
- Salimans T, Kingma DP (2016) Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems* 29

- Soudry D, Hoffer E, Nacson MS, et al (2018) The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research* 19(70):1–57
- Takase S, Kiyono S, Kobayashi S, et al (2023) Spike no more: Stabilizing the pre-training of large language models. *arXiv preprint arXiv:231216903*
- Wang Z, Li Z, Li J (2022) Analyzing sharpness along gd trajectory: Progressive sharpening and edge of stability. *Advances in Neural Information Processing Systems* 35:9983–9994
- Wu J, Bartlett PL, Telgarsky M, et al (2024a) Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency. *arXiv preprint arXiv:240215926*
- Wu J, Braverman V, Lee JD (2024b) Implicit bias of gradient descent for logistic regression at the edge of stability. *Advances in Neural Information Processing Systems* 36
- Zeng A, Liu X, Du Z, et al (2022) Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:221002414*
- Zhu X, Wang Z, Wang X, et al (2022) Understanding edge-of-stability training dynamics with a minimalist example. *arXiv preprint arXiv:221003294*