Proceedings of Machine Learning Research vol 195:1–55, 2023 36th Annual Conference on Learning Theory

The Implicit Bias of Batch Normalization in Linear Models and Two-layer Linear Convolutional Neural Networks

Yuan Cao ABC@SAMPLE.COM
The University of Hong Kong

Difan Zou DZOU@CS.HKU.HK
The University of Hong Kong

Yuanzhi Li YUANZHIL@ANDREW.CMU.EDU
Carnegie Mellon University

Quanquan Gu QGU@CS.UCLA.EDU
University of California, Los Angeles

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We study the implicit bias of batch normalization trained by gradient descent. We show that when learning a linear model with batch normalization for binary classification, gradient descent converges to a *uniform margin classifier* on the training data with an $\exp(-\Omega(\log^2 t))$ convergence rate. This distinguishes linear models with batch normalization from those without batch normalization in terms of both the type of implicit bias and the convergence rate. We further extend our result to a class of two-layer, single-filter linear convolutional neural networks, and show that batch normalization has an implicit bias towards a *patch-wise uniform margin*. Based on two examples, we demonstrate that patch-wise uniform margin classifiers can outperform the maximum margin classifiers in certain learning problems. Our results contribute to a better theoretical understanding of batch normalization.

Keywords: batch normalization, implicit bias, uniform margin

1. Introduction

Batch normalization (BN) is a popular deep learning technique that normalizes network layers by re-centering/re-scaling within a batch of training data (Ioffe and Szegedy, 2015). It has been empirically demonstrated to be helpful for training and often leads to better generalization performance. A series of works have attempted to understand and explain the empirical success of BN from different perspectives. Santurkar et al. (2018); Bjorck et al. (2018); Arora et al. (2019b) showed that BN enables a more feasible learning rate in the training process, thus leading to faster convergence and possibly finding flat minima. A similar phenomenon has also been discovered in Luo et al. (2019), which studied BN by viewing it as an implicit weight decay regularizer. Cai et al. (2019); Kohler et al. (2019) conducted studies on the benefits of BN in training linear models with gradient descent: they quantitatively demonstrated the accelerated convergence of gradient descent with BN compared to that without BN. However, these works on the convergence analysis cannot fully reveal the role of BN in model training and explain its effect on the generalization performance.

To this end, a more important research direction is to study the *implicit bias* (Neyshabur et al., 2014) for BN, or more precisely, identify the special properties of the solution obtained by gradient descent with BN. When the BN is absent, the implicit bias of gradient descent for the linear model

© 2023 Y. Cao, D. Zou, Y. Li & Q. Gu.

(or deep linear models) is widely studied (Soudry et al., 2018; Ji and Telgarsky, 2018): when the data is linearly separable, gradient descent on the cross-entropy loss will converge to the *maximum margin solution*. However, when the BN is added, the implicit bias of gradient descent is far less understood. Even for the simplest linear classification model, there are nearly no such results for the gradient descent with BN except Lyu and Li (2019), which established an implicit bias guarantee for gradient descent on general homogeneous models (covering homogeneous neural networks with BN, if ruling out the zero point). However, they only proved that gradient descent will converge to a Karush–Kuhn–Tucker (KKT) point of the maximum margin problem, while it remains unclear whether the obtained solution is guaranteed to achieve maximum margin, or has other properties while satisfying the KKT conditions.

In this paper, we aim to systematically study the implicit bias of batch normalization in training linear models and a class of linear convolutional neural networks (CNNs) with gradient descent. Specifically, consider a training data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and assume that the training data inputs are centered without loss of generality. Then we consider the linear model and linear single-filter CNN model with batch normalization as follows

$$f(\mathbf{w}, \gamma, \mathbf{x}) = \gamma \cdot \frac{\langle \mathbf{w}, \mathbf{x} \rangle}{\sqrt{n^{-1} \sum_{i=1}^n \langle \mathbf{w}, \mathbf{x}_i \rangle^2}}, \; g(\mathbf{w}, \gamma, \mathbf{x}) = \sum_{p=1}^P \gamma \cdot \frac{\langle \mathbf{w}, \mathbf{x}^{(p)} \rangle}{\sqrt{n^{-1} P^{-1} \sum_{i=1}^n \sum_{p=1}^P \langle \mathbf{w}, \mathbf{x}_i^{(p)} \rangle^2}},$$

where $\mathbf{w} \in \mathbb{R}^d$ is the parameter vector, γ is the scale parameter in batch normalization, and $\mathbf{x}^{(p)}$, $\mathbf{x}_i^{(p)}$ denote the p-th patch in $\mathbf{x}$, $\mathbf{x}_i$ respectively. $\mathbf{w}, \gamma$ are both trainable parameters in the model. To train $f(\mathbf{w}, \gamma, \mathbf{x})$ and $g(\mathbf{w}, \gamma, \mathbf{x})$, we use gradient descent starting from $\mathbf{w}^{(0)}$, $\gamma^{(0)}$ to minimize the cross-entropy loss. The following informal theorem gives a simplified summary of our main results.

Theorem 1.1 (Simplification of Theorems 2.2 and 3.2) *Suppose that $\gamma^{(0)} = 1$, and that the initialization scale $\|\mathbf{w}^{(0)}\|_2$ and learning rate of gradient descent are sufficiently small. Then the following results hold:*

1. ***(Implicit bias of batch normalization in linear models)*** *When training $f(\mathbf{w}, \gamma, \mathbf{x})$, if the equation system $\langle \mathbf{w}, y_i \cdot \mathbf{x}_i \rangle = 1$, $i \in [n]$ has at least one solution, then the training loss converges to zero, and the iterate $\mathbf{w}^{(t)}$ of gradient descent satisfies that*

$$\frac{1}{n^2} \sum_{i,i'=1}^n \left[\mathrm{margin}\big(\mathbf{w}^{(t)}, (\mathbf{x}_{i'}, y_{i'})\big) - \mathrm{margin}\big(\mathbf{w}^{(t)}, (\mathbf{x}_i, y_i)\big)\right]^2 = \exp(-\Omega(\log^2 t)),$$

where $\mathrm{margin}\big(\mathbf{w}, (\mathbf{x}, y)\big) := y \cdot \frac{\langle \mathbf{w}, \mathbf{x} \rangle}{\sqrt{n^{-1} \sum_{i=1}^n \langle \mathbf{w}, \mathbf{x}_i \rangle^2}}$.

2. ***(Implicit bias of batch normalization in single-filter linear CNNs)*** *When training $g(\mathbf{w}, \gamma, \mathbf{x})$, if the equation system $\langle \mathbf{w}, y_i \cdot \mathbf{x}_i^{(p)} \rangle = 1$, $i \in [n]$, $p \in [P]$ has at least one solution, then the training loss converges to zero, and the iterate $\mathbf{w}^{(t)}$ of gradient descent satisfies that*

$$\frac{1}{n^2} \sum_{i,i'=1}^n \left[\mathrm{margin}_{\mathrm{patch}}\big(\mathbf{w}^{(t)}, (\mathbf{x}_{i'}^{(p')}, y_{i'})\big) - \mathrm{margin}_{\mathrm{patch}}\big(\mathbf{w}^{(t)}, (\mathbf{x}_i^{(p)}, y_i)\big)\right]^2 = \exp(-\Omega(\log^2 t)),$$

where $\mathrm{margin}_{\mathrm{patch}}\big(\mathbf{w}, (\mathbf{x}^{(p)}, y)\big) := y \cdot \frac{\langle \mathbf{w}, \mathbf{x}^{(p)} \rangle}{\sqrt{n^{-1} P^{-1} \sum_{i=1}^n \sum_{p=1}^P \langle \mathbf{w}, \mathbf{x}_i^{(p)} \rangle^2}}$.

Theorem 1.1 indicates that batch normalization in linear models and single-filter linear CNNs has an implicit bias towards a *uniform margin*, i.e., gradient descent eventually converges to such a predictor that all training data points are on its margin. Notably, for CNNs, the margins of the predictor are uniform not only over different training data inputs, but also over different patches within each data input.

The major contributions of this paper are as follows:

- By revealing the implicit bias of batch normalization towards a uniform margin, our result distinguishes linear models and CNNs with batch normalization from those without batch normalization. The sharp convergence rate given in our result also allows us to make comparisons with existing results of the implicit bias for linear models and linear networks. In contrast to the $1/\log(t)$ convergence rate of linear logistic regression towards the maximum margin solution (Soudry et al., 2018), the linear model with batch normalization converges towards a uniform margin solution with convergence rate $\exp(-\Omega(\log^2 t))$, which is significantly faster. Therefore, our result demonstrates that batch normalization can significantly increase the directional convergence speed of linear model weight vectors and linear CNN filters.
- Our results can also serve as important examples in the literature of implicit bias of homogeneous models. Although Lyu and Li (2019) showed that gradient descent converges to a KKT point of the maximum margin problem, the KKT points are usually not unique. Recently, it has been shown by Vardi et al. (2022) that for certain neural network architectures, there exist data sets such that the actual solution given by gradient descent is not the maximum margin solution. Our result also provides novel, concrete and practical examples – linear models and linear CNNs with batch normalization – whose implicit bias are not maximum margin, but uniform margin.
- We note that the uniform margin in linear models can also be achieved by performing linear regression minimizing the square loss. However, the patch-wise uniform margin result for single-filter linear CNNs is unique and cannot be achieved by regression models without batch normalization. To further demonstrate the benefit of such an implicit bias for generalization, we also construct two example learning problems, for which the patch-wise uniform margin classifier is guaranteed to outperform the maximum margin classifier in terms of test accuracy.
- From a technical perspective, our analysis gives many novel proof techniques that may be of independent interest. To derive tight bounds, we prove an inequality that is similar to the Chebyshev's sum inequality, but for pair-wise maxima and differences. We also establish an equivalence result connecting a quantity named margin discrepancy to the Euclidean metric and the metric induced by the data sample covariance matrix. At last, we develop an induction argument over five induction hypotheses, which play a key role in developing the sharp convergence rate towards the uniform margin.

2. Batch Normalization in Linear Models

Suppose that $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$ are n arbitrary training data points, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$ for $i \in [n]$. We consider using a linear model with batch normalization to fit these data points. During training, the prediction function is given as

$$f(\mathbf{w}, \gamma, \mathbf{x}) = \gamma \cdot \frac{\langle \mathbf{w}, \mathbf{x} \rangle - \langle \mathbf{w}, \overline{\mathbf{x}} \rangle}{\sqrt{n^{-1} \sum_{i=1}^{n} (\langle \mathbf{w}, \mathbf{x}_i \rangle - \langle \mathbf{w}, \overline{\mathbf{x}} \rangle)^2}},$$

where $\overline{\mathbf{x}} = n^{-1}\sum_{i=1}^n \mathbf{x}_i$ is the mean of all the training data inputs, $\mathbf{w} \in \mathbb{R}^d$ is the linear model parameter vector, and γ is the scale parameter in batch normalization. $\mathbf{w}, \gamma$ are both trainable parameters in the model. Clearly, the above definition exactly gives a linear model with batch normalization during the training of full-batch gradient descent. Note that we can assume the data are centered (i.e., $\overline{\mathbf{x}} = \mathbf{0}$) without loss of generality: if $\overline{\mathbf{x}} \neq \mathbf{0}$, we can simply consider new data inputs $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \overline{\mathbf{x}}$. Therefore, assuming $\overline{\mathbf{x}} = \mathbf{0}$, we have

$$f(\mathbf{w}, \gamma, \mathbf{x}) = \gamma \cdot \frac{\langle \mathbf{w}, \mathbf{x} \rangle}{\|\mathbf{w}\|_{\mathbf{\Sigma}}}, \quad \text{where} \quad \mathbf{\Sigma} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top.$$

Consider training $f(\mathbf{w}, \gamma, \mathbf{x})$ by minimizing the cross-entropy loss

$$L(\mathbf{w}, \gamma) = \frac{1}{n}\sum_{i=1}^n \ell[y_i \cdot f(\mathbf{w}, \gamma, \mathbf{x}_i)]$$

with gradient descent, where $\ell(z) = \log(1 + \exp(-z))$ is the logistic loss. Then starting from the initial $\mathbf{w}^{(0)}$ and $\gamma^{(0)}$, gradient descent with learning rate η takes the following update

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \frac{\eta \cdot \gamma^{(t)}}{n \cdot \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \sum_{i=1}^n \ell'[y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)] \cdot y_i \cdot \left(\mathbf{I} - \frac{\mathbf{\Sigma}\mathbf{w}^{(t)}\mathbf{w}^{(t)\top}}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2}\right)\mathbf{x}_i, \tag{2.1}$$

$$\gamma^{(t+1)} = \gamma^{(t)} - \frac{\eta}{n}\sum_{i=1}^n \ell'[y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)] \cdot y_i \cdot \frac{\langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}}. \tag{2.2}$$

Our goal is to show that $f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x})$ trained by (2.1) and (2.2) eventually achieves the same margin on all training data points. Note that $f(\mathbf{w}, \gamma, \mathbf{x})$ is 1-homogeneous in γ and 0-homogeneous in $\mathbf{w}$, and the prediction of $f(\mathbf{w}, \gamma, \mathbf{x})$ on whether a data input $\mathbf{x}$ belongs to class $+1$ or -1 does not depend on the magnitude of γ. Therefore we focus on the normalized margin $y_i \cdot \langle \mathbf{w}, \mathbf{x} \rangle / \|\mathbf{w}\|_{\mathbf{\Sigma}}$. Moreover, we also introduce the following quantity which we call margin discrepancy:

$$D(\mathbf{w}) := \frac{1}{n^2}\sum_{i,i'=1}^n \left(y_{i'} \cdot \frac{\langle \mathbf{w}, \mathbf{x}_{i'} \rangle}{\|\mathbf{w}\|_{\mathbf{\Sigma}}} - y_i \cdot \frac{\langle \mathbf{w}, \mathbf{x}_i \rangle}{\|\mathbf{w}\|_{\mathbf{\Sigma}}}\right)^2.$$

The margin discrepancy $D(\mathbf{w})$ measures how uniform the margin achieved by $f(\mathbf{w}, \gamma, \mathbf{x})$ is. When $D(\mathbf{w})$ is zero, $f(\mathbf{w}, \gamma, \mathbf{x})$ achieves the exact same margin on all the training data points. Therefore, we call $f(\mathbf{w}, \gamma, \mathbf{x})$ or the corresponding linear classifier $\mathbf{w}$ the *uniform margin classifier* if $D(\mathbf{w}) = 0$.

2.1. The Implicit Bias of Batch Normalization in Linear Models

In this subsection we present our main result on the implicit bias of batch normalization in linear models. We first state the following assumption.

Assumption 2.1 *The equation system $\langle \mathbf{w}, y_i \cdot \mathbf{x}_i \rangle = 1$, $i \in [n]$ has at least one solution.*

Assumption 2.1 is a very mild assumption which commonly holds under over-parameterization. For example, when $d \geq n$, Assumption 2.1 holds almost surely when $\mathbf{x}_i$ are sampled from a non-degenerate distribution. Assumption 2.1 can also hold in many low-dimensional learning problems

as well. Note that Assumption 2.1 is our *only* assumption on the data: we do not make any distribution assumption on the data, and we do not make any assumption on d and n either.

In order to present the most general implicit bias result for batch normalization that covers both the $d \geq n$ case and the $d < n$ case, we introduce some additional notations to unify the argument. We note that by the gradient descent update formula (2.1), $\mathbf{w}^{(t)}$ is only updated in the space $\mathcal{X} := \mathrm{span}\{\mathbf{x}_1, \ldots .\mathbf{x}_n\}$, and the component of $\mathbf{w}^{(t)}$ in $\mathcal{X}^{\perp}$ is unchanged during training (of course, if $\mathcal{X} = \mathbb{R}^d$, then there is no such component). This motivates us to define

$$\lambda_{\max} = \sup_{\mathbf{u} \in \mathcal{X} \setminus \{\mathbf{0}\}} \frac{\mathbf{u}^\top \boldsymbol{\Sigma} \mathbf{u}}{\|\mathbf{u}\|_2^2}, \qquad \lambda_{\min} = \inf_{\mathbf{u} \in \mathcal{X} \setminus \{\mathbf{0}\}} \frac{\mathbf{u}^\top \boldsymbol{\Sigma} \mathbf{u}}{\|\mathbf{u}\|_2^2}.$$

Moreover, we let $\mathbf{P}_{\mathcal{X}}$ be the projection matrix onto $\mathcal{X}$ (if $\mathcal{X} = \mathbb{R}^d$, then $\mathbf{P}_{\mathcal{X}} = \mathbf{I}$ is simply the identity matrix). We can briefly check the scaling of $\lambda_{\max}$ and $\lambda_{\min}$ with a specific example: suppose that $\mathbf{x}_i$, $i \in [n]$ are independently drawn from $N(\mathbf{0}, \mathbf{I})$. Then when $d/n \to c$ for some constant $c \neq 1$, with high probability, $\lambda_{\max}$ and $\lambda_{\max}$ are both of constant order (Vershynin, 2010).

Our main result for the implicit bias of batch normalization in linear models is given in the following theorem.

Theorem 2.2 *Suppose that Assumption 2.1 holds. There exist $M = 1/\mathrm{poly}(\lambda_{\min}^{-1}, \lambda_{\max}, \max_i \|\mathbf{x}_i\|_2)$, $\overline{\eta} = 1/\mathrm{poly}(\lambda_{\min}^{-1}, \lambda_{\max}, \max_i \|\mathbf{x}_i\|_2, \|\mathbf{P}_{\mathcal{X}} \mathbf{w}^{(0)}\|_2^{-1})$ and constants $C_1, C_2, C_3, C_4 > 0$, such that when $\gamma^{(0)} = 1$, $\|\mathbf{P}_{\mathcal{X}} \mathbf{w}^{(0)}\|_2 \leq M$, and $\eta \leq \overline{\eta}$, there exists $t_0 \leq \eta^{-1} \cdot \mathrm{poly}(\lambda_{\min}^{-1}, \lambda_{\max}, \max_i \|\mathbf{x}_i\|_2)$ and the following inequalities hold for all $t \geq t_0$:*

$$\frac{C_1}{\eta \cdot (t - t_0 + 1)} \leq L(\mathbf{w}^{(t)}, \gamma^{(t)}) \leq \frac{C_2}{\eta \cdot (t - t_0 + 1)},$$
$$D(\mathbf{w}^{(t)}) \leq \frac{C_3 \lambda_{\max}}{\lambda_{\min}} \cdot \exp\left[- \frac{C_4 \lambda_{\min}}{\lambda_{\max}^{3/2} \cdot \|\mathbf{P}_{\mathcal{X}} \mathbf{w}^{(0)}\|_2^2} \cdot \log^2((8/9)\eta \cdot (t - t_0) + 1) \right].$$

By the first result in Theorem 2.2, we see that the loss function converges to zero, which guarantees perfect fitting on the training dataset. Moreover, the upper and lower bounds together demonstrate that the $\Theta(1/t)$ convergence rate is sharp. The second result in Theorem 2.2 further shows that the margin discrepancy $D(\mathbf{w})$ converges to zero at a rate of $O(\exp(-\log^2 t))$, demonstrating that batch normalization in linear models has an implicit bias towards the uniform margin. Note that Theorem 2.2 holds under mild assumptions: we only require that (i) a uniform margin solution exists (which is obviously a necessary assumption), and (ii) the initialization scale $\|\mathbf{w}^{(0)}\|_2$ and the learning rate η are small enough (which are common in existing implicit bias results (Gunasekar et al., 2017; Li et al., 2018; Arora et al., 2019a)). Therefore, Theorem 2.2 gives a comprehensive characterization on the implicit bias of batch normalization in linear models.

Comparison with the implicit bias for linear models without BN. When the BN is absent, it has been widely known that gradient descent, performed on linear models with cross-entropy loss, converges to the maximum margin solution with a $O\big(1/\log(t)\big)$ rate (Soudry et al., 2018). In comparison, our result demonstrates that when batch normalization is added to the model, the implicit bias is changed from maximum margin to uniform margin and the (margin) convergence rate can be significantly improved to $\exp(-\Omega(\log^2 t))$, which is way faster than $O(1/\log(t))$.

Comparison with the implicit bias for homogeneous models. Note that when BN is added, the model function $f(\mathbf{w}, \gamma, \mathbf{x})$ is 1-homogeneous for any $\mathbf{w} \neq \mathbf{0}$, i.e., for any constant c, we have

$f(c\mathbf{w}, c\gamma, \mathbf{x}) = c \cdot f(\mathbf{w}, \gamma, \mathbf{x})$. Therefore, by the implicit bias result for general homogeneous models in Lyu and Li (2019), we can conclude that the uniform margin solution is a KKT point of the maximum margin problem. On the other hand, it is clear that the uniform margin solution may not be able to achieve the maximum margin on the training data. This implies that for general homogeneous models (or homogeneous models with BN, which is still a class of homogeneous models), it is possible that gradient descent will not converge to the maximum margin solution.

3. Batch Normalization in Two-layer Linear CNNs

In Section 2, we have shown that batch normalization in linear models has an implicit bias towards the uniform margin. This actually reveals the fact that linear predictors with batch normalization in logistic regression may behave more similarly to the linear regression predictor trained by minimizing the square loss. To further distinguish batch normalization from other methods, in this section we extend our results in Section 2 to a class of linear CNNs with a single convolution filter.

Suppose that $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are n training data points, and each $\mathbf{x}_i$ consists of P patches $\mathbf{x}_i = [\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \ldots, \mathbf{x}_i^{(P)}]$, where $\mathbf{x}_i^{(p)} \in \mathbb{R}^d$, $p \in [P]$. We train a linear CNN with a single filter $\mathbf{w}$ to fit these data points. During training, the CNN model with batch normalization is given as

$$g(\mathbf{w}, \gamma, \mathbf{x}) = \sum_{p=1}^{P} \gamma \cdot \frac{\langle \mathbf{w}, \mathbf{x}^{(p)} \rangle}{\|\mathbf{w}\|_{\mathbf{\Sigma}}} \quad \text{with} \quad \mathbf{\Sigma} = \frac{1}{nP} \sum_{i=1}^{n} \sum_{p=1}^{P} \mathbf{x}_i^{(p)} \mathbf{x}_i^{(p)\top},$$

where γ is the scale parameter in batch normalization. Here $\mathbf{w}, \gamma$ are both trainable parameters. Similar to the linear model case, the above definition can be obtained by centering the training data points. Moreover, note that here different patches of a data point $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(P)}$ can have arbitrary overlaps, and our theory applies even when some of the patches are identical. Note also that in our definition, batch normalization is applied not only over the batch of data, but also over patches, which matches the definition of batch normalization in CNN in Ioffe and Szegedy (2015).

We again consider training $g(\mathbf{w}, \gamma, \mathbf{x})$ by minimizing the cross-entropy loss with gradient descent starting from initialization $\mathbf{w}^{(0)}$, $\gamma^{(0)}$. As the counterpart of the margin discrepancy $D(\mathbf{w})$ defined in Section 2, here we define the patch-wise margin discrepancy as

$$D_{\text{patch}}(\mathbf{w}) = \frac{1}{n^2 P^2} \sum_{i,i'=1}^{n} \sum_{p,p'=1}^{P} \left(y_{i'} \cdot \frac{\langle \mathbf{w}, \mathbf{x}_{i'}^{(p')} \rangle}{\|\mathbf{w}\|_{\mathbf{\Sigma}}} - y_i \cdot \frac{\langle \mathbf{w}, \mathbf{x}_i^{(p)} \rangle}{\|\mathbf{w}\|_{\mathbf{\Sigma}}} \right)^2.$$

We call $g(\mathbf{w}, \gamma, \mathbf{x})$ or the corresponding linear classifier defined by $\mathbf{w}$ the *patch-wise uniform margin classifier* if $D_{\text{patch}}(\mathbf{w}) = 0$.

3.1. The Implicit Bias of Batch Normalization in Two-Layer Linear CNNs

Similar to Assumption 2.1, we make the following assumption to guarantee the existence of the patch-wise uniform margin solution.

Assumption 3.1 *The equation system* $\langle \mathbf{w}, y_i \cdot \mathbf{x}_i^{(p)} \rangle = 1$, $i \in [n]$, $p \in [P]$ *has at least one solution.*

We also extend the notations in Section 2.1 to the multi-patch setting as follows. We define $\mathcal{X} = \text{span}\{\mathbf{x}_i^{(p)}, i \in [n], p \in [P]\}$, and

$$\lambda_{\max} = \sup_{\mathbf{u} \in \mathcal{X} \setminus \{\mathbf{0}\}} \frac{\mathbf{u}^\top \boldsymbol{\Sigma} \mathbf{u}}{\|\mathbf{u}\|_2^2}, \qquad \lambda_{\min} = \inf_{\mathbf{u} \in \mathcal{X} \setminus \{\mathbf{0}\}} \frac{\mathbf{u}^\top \boldsymbol{\Sigma} \mathbf{u}}{\|\mathbf{u}\|_2^2}.$$

Moreover, we let $\mathbf{P}_{\mathcal{X}}$ be the projection matrix onto $\mathcal{X}$.

The following theorem states the implicit bias of batch normalization in two-layer linear CNNs.

Theorem 3.2 *Suppose that Assumption 3.1 holds. There exist $M = 1/\text{poly}(\lambda_{\min}^{-1}, \lambda_{\max}, \max_i \|\mathbf{x}_i\|_2, P)$, $\overline{\eta} = 1/\text{poly}(\lambda_{\min}^{-1}, \lambda_{\max}, \max_i \|\mathbf{x}_i\|_2, P, \|\mathbf{P}_{\mathcal{X}} \mathbf{w}^{(0)}\|_2^{-1})$ and constants $C_1, C_2, C_3, C_4 > 0$, such that when $\gamma^{(0)} = 1$, $\|\mathbf{P}_{\mathcal{X}} \mathbf{w}^{(0)}\|_2 \le M$, and $\eta \le \overline{\eta}$, there exists $t_0 \le \eta^{-1} \cdot \text{poly}(\lambda_{\min}^{-1}, \lambda_{\max}, \max_i \|\mathbf{x}_i\|_2, P)$ and the following inequalities hold for all $t \ge t_0$:*

$$\frac{C_1}{\eta P \cdot (t - t_0 + 1)} \le L(\mathbf{w}^{(t)}, \gamma^{(t)}) \le \frac{C_2}{\eta P \cdot (t - t_0 + 1)},$$

$$D_{\text{patch}}(\mathbf{w}^{(t)}) \le \frac{C_3 \lambda_{\max}}{\lambda_{\min}} \cdot \exp\left[- \frac{C_4 \lambda_{\min}}{\lambda_{\max}^{3/2} \cdot \|\mathbf{P}_{\mathcal{X}} \mathbf{w}^{(0)}\|_2^2} \cdot \log^2((8/9)\eta P \cdot (t - t_0) + 1) \right].$$

Theorem 3.2 is the counterpart of Theorem 2.2 for two-layer single-filter CNNs. It further reveals that fact that batch normalization encourages CNNs to achieve the same margin on all data patches. It is clear that the convergence rate is $\exp(-\Omega(\log^2 t))$, which is fast compared with the convergence towards the maximum classifier when training linear CNNs without batch normalization.

3.2. Examples Where Uniform Margin Outperforms Maximum Margin

Here we give two examples of learning problems, and show that when training two-layer, single-filter linear CNNs, the patch-wise uniform classifier given by batch normalization outperforms the maximum margin classifier given by the same neural network model without batch normalization.

We note that when making predictions on a test data input $\mathbf{x}_{\text{test}}$, the standard deviation calculated in batch normalization (i.e., the denominator $\|\mathbf{w}\|_{\boldsymbol{\Sigma}}$) is still based on the training data set and is therefore unchanged. Then as long as $\gamma > 0$, we have

$$\text{sign}[g(\mathbf{w}, \gamma, \mathbf{x}_{\text{test}})] = \text{sign}\left[\sum_{p=1}^{P} \gamma \cdot \frac{\langle \mathbf{w}, \mathbf{x}^{(p)} \rangle}{\|\mathbf{w}\|_{\boldsymbol{\Sigma}}} \right] = \text{sign}(\langle \mathbf{w}, \overline{\mathbf{x}}_{\text{test}} \rangle),$$

where we define $\overline{\mathbf{x}}_{\text{test}} := \sum_{p=1}^{P} \mathbf{x}_{\text{test}}^{(p)}$. We compare the performance of the patch-wise uniform margin classifier $\mathbf{w}_{\text{uniform}}$ with the maximum margin classifier, which is defined as

$$\mathbf{w}_{\max} = \text{argmin} \|\mathbf{w}\|_2^2, \text{ subject to } y_i \cdot \langle \mathbf{w}, \overline{\mathbf{x}}_i \rangle \ge 1, i \in [n]. \tag{3.1}$$

By Soudry et al. (2018), $\mathbf{w}_{\max}$ can be obtained by training a two-layer, single-filter CNN without batch normalization. We hence study the difference between $\mathbf{w}_{\text{uniform}}$ and $\mathbf{w}_{\max}$ in two examples. Below we present the first learning problem example and the learning guarantees.

Example 3.1 *Let $\mathbf{u} \in \mathbb{R}^d$ be a fixed vector. Then each data point $(\mathbf{x}, y)$ with*

$$\mathbf{x}_i = [\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \ldots, \mathbf{x}_i^{(P)}] \in \mathbb{R}^{d \times P}$$

and $y \in \{-1, 1\}$ is generated from the distribution $\mathcal{D}_1$ as follows:

- *The label y is generated as a Rademacher random variable.*
- *For $p \in [P]$, the input patch $\mathbf{x}^{(p)}$ is given as $y \cdot \mathbf{u} + \boldsymbol{\xi}^{(p)}$, where $\boldsymbol{\xi}^{(p)}$, $p \in [n]$ are independent noises generated from $N(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{u}\mathbf{u}^\top/\|\mathbf{u}\|_2^2))$.*

Theorem 3.3 *Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be the training data set consisting of n independent data points drawn from the distribution $\mathcal{D}$ in Example 3.1. Suppose that $d = 2n$, $P \geq 4$ and $\sigma \geq 20\|\mathbf{u}\|_2 \cdot P^{1/2} d^{1/2}$. Then with probability 1, the maximum margin classifier $\mathbf{w}_{\max}$ on S exists and is unique, and the patch-wise uniform margin classifier $\mathbf{w}_{\text{uniform}}$ on S exist, and is unique up to a scaling factor. Moreover, with probability at least $1 - \exp(-\Omega(d))$ with respect to the randomness in the training data, the following results hold:*

- $\mathbb{P}_{(\mathbf{x}_{\text{test}}, y_{\text{test}}) \sim \mathcal{D}_1}(y_{\text{test}} \cdot \langle \mathbf{w}_{\text{uniform}}, \overline{\mathbf{x}}_{\text{test}} \rangle < 0) = 0.$
- $\mathbb{P}_{(\mathbf{x}_{\text{test}}, y_{\text{test}}) \sim \mathcal{D}_1}(y_{\text{test}} \cdot \langle \mathbf{w}_{\max}, \overline{\mathbf{x}}_{\text{test}} \rangle < 0) = \Theta(1).$

Below we give the second learning problem example and the corresponding learning guarantees for the patch-wise uniform margin classifier and the maximum margin classifier.

Example 3.2 *Let $\mathbf{x} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}]$ be the data with two patches, where $\mathbf{x}^{(p)} \in \mathbb{R}^d$ is a d-dimensional vector. Let $\mathbf{u}$ and $\mathbf{v}$ be two fixed vectors and $\rho \in [0, 0.5)$. Then given the Rademacher label $y \in \{-1, 1\}$, the data input $\mathbf{x}$ is generated from the distribution $\mathcal{D}_2$ as follows:*

- *With probability $1 - \rho$, one data patch $\mathbf{x}^{(p)}$ is the strong signal $y\mathbf{u}$ and the other patch is the random Gaussian noise $\boldsymbol{\xi} \sim N\big(0, \sigma^2(\mathbf{I} - \mathbf{u}\mathbf{u}^\top/\|\mathbf{u}\|_2^2 - \mathbf{v}\mathbf{v}^\top/\|\mathbf{v}\|_2^2)\big)$.*
- *With probability ρ, one data patch $\mathbf{x}^{(p)}$ is the weak signal $y\mathbf{v}$ and the other patch is the combination of random noise $\boldsymbol{\xi} \sim N(0, \sigma^2\mathbf{I})$ and feature noise $\alpha \cdot \zeta\mathbf{u}$, where ζ is randomly drawn from $\{-1, 1\}$ equally and $\alpha \in (0, 1)$.*

The signals are set to be orthogonal to each other, i.e., $\langle \mathbf{u}, \mathbf{v} \rangle = 0$. Moreover, we set $d = n^2 \log(n)$, $\sigma_0 = d^{-1/2}$, $\rho = n^{-3/4}$, $\alpha = n^{-1/2}$, $\|\mathbf{u}\|_2 = 1$, and $\|\mathbf{v}\|_2 = \alpha^2$.

Theorem 3.4 *Suppose that the data is generated according to Example 3.2, then let $\mathbf{w}_{\text{uniform}}$ and $\mathbf{w}_{\max}$ be the uniform margin and maximum margin solution in the subspace $\mathcal{X}$ respectively. Then with probability at least $1 - \exp(-\Omega(n^{1/4}))$ with respect to the randomness in the training data, the following holds:*

- $\mathbb{P}_{(\mathbf{x}_{\text{test}}, y_{\text{test}}) \sim \mathcal{D}_2}(y \cdot \langle \mathbf{w}_{\text{uniform}}, \overline{\mathbf{x}}_{\text{test}} \rangle < 0) \leq \frac{1}{n^{10}}.$
- $\mathbb{P}_{(\mathbf{x}_{\text{test}}, y_{\text{test}}) \sim \mathcal{D}_2}(y \cdot \langle \mathbf{w}_{\max}, \overline{\mathbf{x}}_{\text{test}} \rangle < 0) \geq \frac{1}{4n^{3/4}}.$

We note that data models similar to Examples 3.1 and 3.2 have been considered in a series of works (Allen-Zhu and Li, 2020; Zou et al., 2021; Cao et al., 2022). According to Theorems 3.3 and 3.4, the patch-wise uniform margin classifier achieves better test accuracy than the maximum margin classifier in both examples. The intuition behind these examples is that by ensuring a patch-wise uniform margin, the classifier amplifies the effect of weak, stable features over strong, unstable features and noises. We remark that it is not difficult to contract more examples where patch-wise uniform margin classifier performs better. However, we can also similarly construct other examples where the maximum margin classifier gives better predictions. The goal of our discussion here is just to demonstrate that there exist such learning problems where patch-wise uniform margin classifiers perform well.

4. Overview of the Proof Technique

In this section, we explain our key proof techniques in the study of implicit bias of batch normalization, and discuss the key technical contributions in this paper. For clarity, we will mainly focus on the setting of batch normalization in linear models as is defined in Section 2.

We first introduce some simplification of notations. As we discussed in Subsection 2.1, the training of $\mathbf{w}^{(t)}$ always happen in the subspace $\mathcal{X} = \mathrm{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. And when $\mathcal{X} \subsetneq \mathbb{R}^d$, we need the projection matrix $\mathbf{P}_{\mathcal{X}}$ in our result. In fact, under the setting where $\mathcal{X} \subsetneq \mathbb{R}^d$, we need to apply such a projection whenever $\mathbf{w}$ occurs, and the notations can thus be quite complicated. To simplify notations, throughout our proof we use the slight abuse of notation

$$\|\mathbf{a}\|_2 := \|\mathbf{P}_{\mathcal{X}}\mathbf{a}\|_2, \qquad \langle \mathbf{a}, \mathbf{b}\rangle = \langle \mathbf{P}_{\mathcal{X}}\mathbf{a}, \mathbf{P}_{\mathcal{X}}\mathbf{b}\rangle$$

for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$. Then by the definition of $\lambda_{\max}, \lambda_{\min}$, we have $\lambda_{\min} \cdot \|\mathbf{a}\|_2^2 \le \|\mathbf{a}\|_{\mathbf{\Sigma}}^2 \le \lambda_{\max} \cdot \|\mathbf{a}\|_2^2$ for all $\mathbf{a} \in \mathbb{R}^d$. Moreover, let $\mathbf{w}^* \in \mathcal{X}$ be the unique vector satisfying

$$\mathbf{w}^* \in \mathcal{X}, \quad \langle \mathbf{w}^*, y_i \cdot \mathbf{x}_i\rangle = 1, \quad i \in [n], \tag{4.1}$$

and denote $\mathbf{z}_i = y_i \cdot \mathbf{x}_i$, $i \in [n]$. Then it is clear that $\langle \mathbf{w}^*, \mathbf{x}_i\rangle = 1$ for all $i \in [n]$. In our analysis, we frequently encounter the derivatives of the cross entropy loss on each data point. Therefore we also denote $\ell_i' = \ell'[y_i \cdot f(\mathbf{w}, \gamma, \mathbf{x}_i)]$, $\ell_i'^{(t)} = \ell'[y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)]$, $i \in [n]$.

4.1. Positive Correlation Between $\mathbf{w}^*$ and the Gradient Update

In order to study the implicit bias of batch normalization, the first challenge is to identify a proper target to which the predictor converges. In our analysis, this proper target, i.e. the uniform margin solution, is revealed by a key identity which is presented in the following lemma.

Lemma 4.1 *Under Assumption 2.1, for any $\mathbf{w} \in \mathbb{R}^d$, it holds that*

$$\begin{aligned}&\langle -\nabla_{\mathbf{w}} L(\mathbf{w}, \gamma), \mathbf{w}^*\rangle \\ &\qquad = \frac{\gamma}{2n^2\|\mathbf{w}\|_{\mathbf{\Sigma}}^3} \sum_{i,i'=1}^{n} |\ell_i'| \cdot |\ell_{i'}'| \cdot (\langle \mathbf{w}, \mathbf{z}_{i'}\rangle - \langle \mathbf{w}, \mathbf{z}_i\rangle) \cdot (|\ell_{i'}'|^{-1} \cdot \langle \mathbf{w}, \mathbf{z}_{i'}\rangle - |\ell_i'|^{-1} \cdot \langle \mathbf{w}, \mathbf{z}_i\rangle).\end{aligned}$$

Recall that $\mathbf{w}^*$ defined in (4.1) is a uniform margin solution. Lemma 4.1 thus gives an exact calculation on the component of the training loss gradient pointing towards a uniform margin classifier. More importantly, we see that the factors $|\ell_i'|^{-1}$, $|\ell_{i'}'|^{-1}$ are essentially also functions of $\langle \mathbf{w}, \mathbf{z}_i\rangle$ and $\langle \mathbf{w}, \mathbf{z}_{i'}\rangle$ respectively. If the margins of the predictor on a pair of data points $(\mathbf{x}_i, y_i)$ and $(\mathbf{x}_{i'}, y_{i'})$ are not equal, i.e., $\langle \mathbf{w}, \mathbf{z}_i\rangle \neq \langle \mathbf{w}, \mathbf{z}_{i'}\rangle$, then we can see from Lemma 4.1 that a gradient descent step on the current predictor will push the predictor towards the direction of $\mathbf{w}^*$ by a positive length. To more accurately characterize this property, we give the following lemma.

Lemma 4.2 *For all $t \ge 0$, it holds that*

$$\langle \mathbf{w}^{(t+1)}, \mathbf{w}^*\rangle \ge \langle \mathbf{w}^{(t)}, \mathbf{w}^*\rangle + \frac{\gamma^{(t)}\eta}{2\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \cdot \max\left\{\frac{\exp(-\gamma^{(t)})}{8}, \min_i |\ell_i'^{(t)}|\right\} \cdot D(\mathbf{w}^{(t)}),$$

Lemma 4.2 is established based on Lemma 4.1. It shows that as long as the margin discrepancy is not zero, the inner product $\langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle$ will increase during training. It is easy to see that the result in Lemma 4.2 essentially gives two inequalities:

$$\langle \mathbf{w}^{(t+1)}, \mathbf{w}^* \rangle \geq \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle + \frac{\gamma^{(t)}\eta}{2\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \cdot \frac{\exp(-\gamma^{(t)})}{8} \cdot D(\mathbf{w}^{(t)}), \tag{4.2}$$

$$\langle \mathbf{w}^{(t+1)}, \mathbf{w}^* \rangle \geq \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle + \frac{\gamma^{(t)}\eta}{2\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \cdot \min_i |\ell_i'^{(t)}| \cdot D(\mathbf{w}^{(t)}). \tag{4.3}$$

In fact, inequality (4.3) above with the factor $\min_i |\ell_i'^{(t)}|$ is relatively easy to derive – we essentially lower bound each $|\ell_i'^{(t)}|$ by the minimum over all of them. In comparison, inequality (4.2) with the factor $\exp(-\gamma^{(t)})$ is highly nontrivial: in the early stage of training where the predictor may have very different margins on different data, we can see that $y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i) = \gamma^{(t)} \cdot \langle \mathbf{w}^{(t)}, y_i \cdot \mathbf{x}_i \rangle \cdot (n^{-1} \sum_j \langle \mathbf{w}^{(t)}, \mathbf{x}_j \rangle^2)^{-1/2}$ can be as large as $\gamma^{(t)} \cdot \sqrt{n}$ when only one inner product among $\{\langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle\}_{i=1}^n$ is large and the other are all close to zero. Hence, $\min_i |\ell_i'^{(t)}|$ can be smaller than $\exp(-\gamma^{(t)} \cdot \sqrt{n})$ in the worst case. Therefore, (4.2) is tighter than (4.3) when the margins of the predictor on different data are not close. This tighter result is proved based on a technical inequality (see Lemma C.1), and deriving this result is one of the key technical contributions of this paper.

4.2. Equivalent Metrics of Margin Discrepancy and Norm Bounds

We note that the result in Lemma 4.2 involves the inner products $\langle \mathbf{w}^{(t+1)}, \mathbf{w}^* \rangle$, $\langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle$ as well as the margin discrepancy $D(\mathbf{w}^{(t)})$. The inner products and the margin discrepancy are essentially different metrics on how uniform the margins are. To proceed, we need to unify these different metrics. We have the following lemma addressing this issue.

Lemma 4.3 *For any $\mathbf{w} \in \mathbb{R}^d$, it holds that*

$$\|\mathbf{w}\|_{\mathbf{\Sigma}}^2 \cdot D(\mathbf{w}) = \frac{1}{n^2} \sum_{i,i'=1}^n (\langle \mathbf{w}, \mathbf{z}_{i'} \rangle - \langle \mathbf{w}, \mathbf{z}_i \rangle)^2 = 2\|\mathbf{w} - \langle \mathbf{w}^*, \mathbf{w} \rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_{\mathbf{\Sigma}}^2,$$

$$\lambda_{\min} \cdot \left\| \mathbf{w} - \frac{\langle \mathbf{w}^*, \mathbf{w} \rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^* \right\|_2^2 \leq \|\mathbf{w} - \langle \mathbf{w}^*, \mathbf{w} \rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_{\mathbf{\Sigma}}^2 \leq \lambda_{\max} \cdot \left\| \mathbf{w} - \frac{\langle \mathbf{w}^*, \mathbf{w} \rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^* \right\|_2^2.$$

By Lemma 4.3, it is clear that the $\|\mathbf{w}\|_{\mathbf{\Sigma}} \cdot \sqrt{D(\mathbf{w})}$, $\|\mathbf{w} - \langle \mathbf{w}^*, \mathbf{w} \rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_{\mathbf{\Sigma}}$ and $\|\mathbf{w} - \|\mathbf{w}^*\|_2^{-2} \cdot \langle \mathbf{w}^*, \mathbf{w} \rangle \cdot \mathbf{w}^*\|_2$ are equivalent metrics on the distance between $\mathbf{w}$ and $\mathrm{span}\{\mathbf{w}^*\}$. Lemma 4.3 is fundamental throughout our proof as it unifies (i) the Euclidean geometry induced by linear model and gradient descent and (ii) the geometry defined by $\mathbf{\Sigma}$ induced by batch normalization. Eventually, Lemma 4.3 converts both metrics to the margin discrepancy, which is essential in our proof.

Even with the metric equivalence results, Lemma 4.2 alone is still not sufficient to demonstrate the convergence to a uniform margin: If the gradient consistently has a even larger component along a different direction, then the predictor, after normalization, will be pushed farther away from the uniform margin solution. Therefore, it is necessary to control the growth of $\mathbf{w}^{(t)}$ along the other directions. To do so, we give upper bounds on $\|\mathbf{w}^{(t)}\|_2$. Intuitively, the change of $\mathbf{w}$ in the radial direction will not change the objective function value as $f(\mathbf{w}, \gamma, \mathbf{x})$ is 0-homogeneous in $\mathbf{w}$. Therefore, the training loss gradient is orthogonal to $\mathbf{w}$, and when the learning rate is small, $\|\mathbf{w}^{(t)}\|_2$ will hardly change during training. The following lemma is established following this intuition.

Lemma 4.4 *For all $t \geq 0$, it holds that*

$$\|\mathbf{w}^{(t)}\|_2^2 \leq \|\mathbf{w}^{(t+1)}\|_2^2 \leq \|\mathbf{w}^{(t)}\|_2^2 + 4\eta^2 \cdot \frac{\gamma^{(t)2} \cdot \max_i \|\mathbf{x}_i\|_2^3}{\lambda_{\min}^2 \cdot \|\mathbf{w}^{(t)}\|_2^2}.$$

Moreover, if $\|\mathbf{w}^{(t)} - \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle \cdot \|\mathbf{w}^*\|_2^{-1} \cdot \mathbf{w}^*\|_2 \leq \|\mathbf{w}^{(0)}\|_2/2$, then

$$\|\mathbf{w}^{(t+1)}\|_2^2 \leq \|\mathbf{w}^{(t)}\|_2^2 + \eta^2 G \cdot H^{(t)} \cdot \left\| \mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)} \rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^* \right\|_2^2,$$

where $G = 64\lambda_{\min}^{-3} \cdot \max_i \|\mathbf{x}_i\|_2^6 \cdot \|\mathbf{w}^{(0)}\|_2^{-4}$, and $H^{(t)} = \max\{|\ell_1'^{(t)}|^2, \ldots, |\ell_n'^{(t)}|^2, \exp(-2\gamma^{(t)})\} \cdot \max\{\gamma^{(t)2}, \gamma^{(t)4}\}$.

Lemma 4.4 demonstrates that $\|\mathbf{w}^{(t)}\|_2$ is monotonically increasing during training, but the speed it increases is much slower than the speed $\langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle$ increases (when η is small), as is shown in Lemma 4.2. We note that Lemma 4.4 particularly gives a tighter inequality when $\mathbf{w}^{(t)}$ is close enough to $\mathrm{span}\{\mathbf{w}^*\}$. This inequality shows that in the later stage of training when the margins tend uniform and the loss derivatives $|\ell_i'^{(t)}|$, $i \in [n]$ on the training data tend small, $\|\mathbf{w}^{(t)}\|_2$ increases even slower. The importance of this result can be seen considering the case where all $|\ell_i'^{(t)}|$'s are equal (up to constant factors): in this case, combining the bounds in Lemmas 4.2, 4.3 and 4.4 can give an inequality of the form

$$\left\| \mathbf{w}^{(t+1)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t+1)} \rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^* \right\|_2^2 \leq (1 - A^{(t)}) \cdot \left\| \mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)} \rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^* \right\|_2^2 \tag{4.4}$$

for some $A^{(t)} > 0$ that depends on $|\ell_i'^{(t)}|$, $i \in [n]$. (4.4) is clearly the key to show the monotonicity and convergence of the $\|\mathbf{w} - \|\mathbf{w}^*\|_2^{-2} \cdot \langle \mathbf{w}^*, \mathbf{w} \rangle \cdot \mathbf{w}^*\|_2$, which eventually leads to a last-iterate bound of the margin discrepancy according to Lemma 4.3. However, the rigorous version of the inequality above needs to be proved within an induction, which we explain in the next subsection.

4.3. Final Convergence Guarantee With Sharp Convergence Rate

Lemmas 4.2, 4.3 and 4.4 give the key intermediate results in our proof. However, it is still technically challenging to show the convergence and give a sharp convergence rate. We remind the readers that during training, the loss function value at each data point $(\mathbf{x}_i, y_i)$ converges to zero, and so is the absolute value of the loss derivative $|\ell_i'^{(t)}|$. Based on the bounds in Lemma 4.2 and the informal result in (4.4), we see that if $|\ell_i'^{(t)}|$, $i \in [n]$ vanish too fast, then the margin discrepancy may not have sufficient time to converge. Therefore, in order to show the convergence of margin discrepancy, we need to accurately characterize the orders of $\gamma^{(t)}$ and $|\ell_i'^{(t)}|$.

Intuitively, characterizing the orders of $\gamma^{(t)}$ and $|\ell_i'^{(t)}|$ is easier when the margins are relatively uniform, because in this case $|\ell_i'^{(t)}|$, $i \in [n]$ are all roughly of the same order. Inspired by this, we implement a two-stage analysis, where the first stage provides a warm start for the second stage with relatively uniform margins. The following lemma presents the guarantees in the first stage.

Lemma 4.5 *Let* $\epsilon = (16\max_i \|\mathbf{z}_i\|_2)^{-1} \cdot \|\mathbf{w}^*\|_2^{-1} \cdot \min\big\{1/3, \|\mathbf{w}^{(0)}\|_2^{-1} \cdot \lambda_{\min}^{1/2}/(40\lambda_{\max}^{3/4})\big\}$. *There exist constants* $c_1, c_2, c_3 > 0$ *such that if*

$$\|\mathbf{w}^{(0)}\|_2 \le c_1 \cdot \min\big\{1, (\max_i \|\mathbf{z}_i\|_2)^{-1/2} \cdot \|\mathbf{w}^*\|_2^{-1}\big\} \cdot \lambda_{\max}^{-3/4} \cdot \lambda_{\min}^{1/2},$$
$$\eta \le c_2 \cdot \min\big\{1, \epsilon \cdot \|\mathbf{w}^{(0)}\|_2^2 \cdot \lambda_{\min} \cdot (\max_i \|\mathbf{x}_i\|_2)^{-3/2}, \epsilon^4 \cdot \lambda_{\min}^3 \cdot \lambda_{\max}^{-3/2} \cdot \|\mathbf{w}^*\|_2^{-1} \cdot (\max_i \|\mathbf{x}_i\|_2)^{-3}\big\},$$

then there exists $t_0 \le c_3\eta^{-1}\epsilon^{-2} \cdot \lambda_{\max}^{3/2} \cdot \lambda_{\min}^{-1} \cdot \|\mathbf{w}^{(0)}\|_2^2 \cdot \|\mathbf{w}^*\|_2$ *such that*

1. $1/2 \le \gamma^{(t)} \le 3/2$ *for all* $0 \le t \le t_0$.
2. $\|\mathbf{w}^{(0)}\|_2 \le |\mathbf{w}^{(t)}\|_2 \le (1+\epsilon/2) \cdot \|\mathbf{w}^{(0)}\|_2$ *for all* $0 \le t \le t_0$.
3. $\big\|\mathbf{w}^{(t_0)} - \langle \mathbf{w}^*, \mathbf{w}^{(t_0)}\rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*\big\|_2 \le \epsilon \cdot \|\mathbf{w}^{(0)}\|_2$.

In Lemma 4.5, we set up a target rate ϵ that only depends on the training data and the initialization $\mathbf{w}^{(0)}$, so that $\mathbf{w}^{(t_0)}$ satisfying the three conclusions Lemma 4.5 can serve as a good enough warm start for our analysis on the second stage starting from iteration t_0.

The study of the convergence of margin discrepancy for $t \ge t_0$ is the most technical part of our proof. The results are summarized in the following lemma.

Lemma 4.6 *Let* G *be defined as in Lemma 4.4, and* ϵ *be defined as in Lemma 4.5. Suppose that there exists* $t_0 \in \mathbb{N}_+$ *such that* $1/2 \le \gamma^{(t_0)} \le 3/2$, *and*

$$\|\mathbf{w}^{(0)}\|_2 \le |\mathbf{w}^{(t_0)}\|_2 \le (1+\epsilon/2) \cdot \|\mathbf{w}^{(0)}\|_2, \ \big\|\mathbf{w}^{(t_0)} - \langle \mathbf{w}^*, \mathbf{w}^{(t_0)}\rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*\big\|_2 \le \epsilon \cdot \|\mathbf{w}^{(0)}\|_2.$$

Then there exist constants c_1, c_2, c_3, c_4, *such that as long as* $\eta \le c_1 G^{-1} \min\{\epsilon^2, \lambda_{\min} \cdot \lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^{-2}\}$, *the following results hold for all* $t \ge t_0$:

1. $\big\|\mathbf{w}^{(t_0)} - \langle \mathbf{w}^*, \mathbf{w}^{(t_0)}\rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*\big\|_2 \ge \cdots \ge \big\|\mathbf{w}^{(t)} - \langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*\big\|_2$.
2. $\|\mathbf{w}^{(0)}\|_2 \le \|\mathbf{w}^{(t)}\|_2 \le (1+\epsilon) \cdot \|\mathbf{w}^{(0)}\|_2$.
3. $\log[(\eta/8) \cdot (t-t_0) + \exp(\gamma^{(t_0)})] \le \gamma^{(t)} \le \log[8\eta \cdot (t-t_0) + 2\exp(\gamma^{(t_0)})]$.
4. $\Big\|\mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\Big\|_2^2 \le \epsilon \cdot \|\mathbf{w}^{(0)}\|_2 \cdot \exp\Big[-\frac{c_2\lambda_{\min} \cdot \log^2((8/9)\eta \cdot (t-t_0)+1)}{\lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2}\Big]$.
5. $\max_i |\langle \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2, \mathbf{z}_i\rangle - \|\mathbf{w}^*\|_2^{-1}| \cdot \gamma^{(t)} \le \|\mathbf{w}^*\|_2^{-1}/4$.
6. *It holds that*

$$\frac{1}{40} \cdot \frac{1}{\eta \cdot (t-t_0)+1} \le \ell(y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)) \le \frac{12}{\eta \cdot (t-t_0)+1},$$
$$D(\mathbf{w}) \le \frac{c_3\lambda_{\max}}{\lambda_{\min}} \cdot \exp\Bigg[-\frac{c_4\lambda_{\min}}{\lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \log^2((8/9)\eta \cdot (t-t_0)+1)\Bigg].$$

A few remarks about Lemma 4.6 are in order: The first and the second results guarantee that the properties of the warm start $\mathbf{w}^{(t_0)}$ given in Lemma 4.5 are preserved during training. The third result on $\gamma^{(t)}$ controls the values of $|\ell_i'|$, $i \in [n]$ given uniform enough margins, and also implies the convergence rate of the training loss function. The fourth result gives the convergence rate of $\big\|\mathbf{w}^{(t)} - \langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*\big\|_2$, which is an equivalent metric of the margin discrepancy in ℓ_2

distance. The fifth result essentially follows by the convergence rates of $\gamma^{(t)}$ and $\left\|\mathbf{w}^{(t)} - \langle \mathbf{w}^*, \mathbf{w}^{(t)} \rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*\right\|_2$, and it further implies that $|\ell_i'|$, $i \in [n]$ only differ by constant factors. Finally, the sixth result reformulates the previous results and gives the conclusions of Theorem 2.2.

In our proof of Lemma 4.6, we first show the first five results based on an induction, where each of these results rely on the margin uniformity in the previous iterations. The last result is then proved based on the first five results. We also regard this proof as a key technical contribution of our work. Combining Lemmas 4.5 and 4.6 leads to Theorem 2.2, and the proof is thus complete.

5. Conclusion and Future Work

In this work we theoretically demonstrate that with batch normalization, gradient descent converges to a uniform margin solution when training linear models, and converges to a patch-wise uniform margin solution when training two-layer, single-filter linear CNNs. These results give a precise characterization of the implicit bias of batch normalization.

An important future work direction is to study the implicit bias of batch normalization in more complicated neural network models with multiple filters, multiple layers, and nonlinear activation functions. Our analysis may also find other applications in studying the implicit bias of other normalization techniques such as layer normalization.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. YL is supported by the National Science Foundation CCF-2145703. QG is supported in part by the National Science Foundation CAREER Award 1906169 and IIS-2008981, and the Sloan Research Fellowship. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019a.

Sanjeev Arora, Kaifeng Lyu, and Zhiyuan Li. Theoretical analysis of auto rate-tuning by batch normalization. In *7th International Conference on Learning Representations, ICLR 2019*, 2019b.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

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

Yongqiang Cai, Qianxiao Li, and Zuowei Shen. A quantitative analysis of the effect of batch normalization on gradient descent. In *International Conference on Machine Learning*, pages 882–890. PMLR, 2019.

Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2022.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

Yonatan Dukler, Quanquan Gu, and Guido Montúfar. Optimization theory for relu neural networks trained with normalization layers. In *International conference on machine learning*, pages 2751–2760. PMLR, 2020.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1827–1836, 2018a.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018b.

Elad Hoffer, Ron Banner, Itay Golan, and Daniel Soudry. Norm matters: efficient and accurate normalization schemes in deep networks. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

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

Jonas Kohler, Hadi Daneshmand, Aurelien Lucchi, Thomas Hofmann, Ming Zhou, and Klaus Neymeyr. Exponential convergence rates for batch normalization: The power of length-direction decoupling in non-convex optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 806–815. PMLR, 2019.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47, 2018.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021.

Ping Luo, Xinjiang Wang, Wenqi Shao, and Zhanglin Peng. Towards understanding regularization in batch normalization. In *International Conference on Learning Representations*, 2019.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Depen Morwani and Harish G Ramaswamy. Inductive bias of gradient descent for exponentially weight normalized smooth homogeneous neural nets. *arXiv preprint arXiv:2010.12909*, 2020.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428, 2019a.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3051–3059, 2019b.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

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

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Gal Vardi, Ohad Shamir, and Nati Srebro. On margin maximization in linear and relu networks. *Advances in Neural Information Processing Systems*, 35:37024–37036, 2022.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Xiaoxia Wu, Rachel Ward, and Léon Bottou. Wngrad: Learn the learning rate in gradient descent. *arXiv preprint arXiv:1803.02865*, 2018.

Xiaoxia Wu, Edgar Dobriban, Tongzheng Ren, Shanshan Wu, Zhiyuan Li, Suriya Gunasekar, Rachel Ward, and Qiang Liu. Implicit regularization and convergence for weight normalization. *Advances in Neural Information Processing Systems*, 33:2835–2847, 2020.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

Greg Yang, Jeffrey Pennington, Vinay Rao, Jascha Sohl-Dickstein, and Samuel S Schoenholz. A mean field theory of batch normalization. In *International Conference on Learning Representations*, 2019.

Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. Understanding the generalization of adam in learning neural networks with proper regularization. *arXiv preprint arXiv:2108.11371*, 2021.

Appendix A. Additional Related Work

Implicit Bias. In recent years, there have emerged a large number of works studying the implicit bias of various optimization algorithms for different models. We will only review a part of them that is mostly related to this paper.

Theoretical analysis of implicit bias has originated from linear models. For linear regression, Gunasekar et al. (2018b) showed that starting from the origin point, gradient descent converges to the minimum ℓ_2 norm solution. For linear classification problems, gradient descent is shown to converge to the maximum margin solution on separable data (Soudry et al., 2018; Nacson et al., 2019a; Ji and Telgarsky, 2019). Similar results have also been obtained for other optimization algorithms such as mirror descent (Gunasekar et al., 2018a) and stochastic gradient descent (Nacson et al., 2019b). The implicit bias has also been widely studied beyond linear models such as matrix factorization (Gunasekar et al., 2017; Li et al., 2021; Arora et al., 2019a), linear networks (Li et al., 2018; Ji and Telgarsky, 2018; Gunasekar et al., 2018b; Pesme et al., 2021), and more complicated nonlinear networks (Chizat and Bach, 2020), showing that (stochastic) gradient descent may converge to different types of solutions. However, none of them can be adapted to our setting.

Theory for Normalization Methods. Many normalization methods, including batch normalization, weight normalization (Salimans and Kingma, 2016), layer normalization (Ba et al., 2016), and group normalization (Wu and He, 2018), have been recently developed to improve generalization performance. From a theoretical perspective, a series of works have established the close connection between the normalization methods and the adaptive effective learning rate (Hoffer et al., 2018; Arora et al., 2019b; Morwani and Ramaswamy, 2020). Based on the auto-learning rate adjustment effect of weight normalization, Wu et al. (2018) developed an adaptive learning rate scheduler and proved a near-optimal convergence rate for SGD. Moreover, Wu et al. (2020) studied the implicit regularization effect of weight normalization for linear regression, showing that gradient descent can always converge close to the minimum ℓ_2 norm solution, even from the initialization that is far from zero. Dukler et al. (2020) further investigated the convergence of gradient descent for training a two-layer ReLU network with weight normalization. For multi-layer networks, Yang et al. (2019) developed a mean-field theory for batch normalization, showing that the gradient explosion brought by large network depth cannot be addressed by BN, if there is no residual connection. These works either concerned different normalization methods or investigated the behavior of BN in different aspects, which can be seen as orthogonal to our results.

Appendix B. Experiments

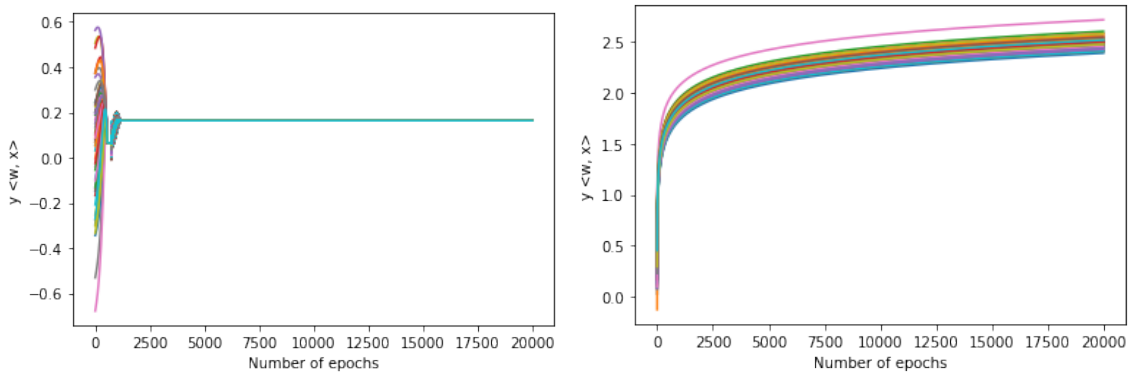
Here we present some preliminary simulation and real data experiment results to demonstrate that batch normalization encourages a uniform margin. The results are given in Figure 1 and Figure 2 respectively.

In Figure 1, we train linear models with/without batch normalization, and plot the values of $y_i \cdot \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle$, $i = 1, \ldots, n$ in Figure 1. From Figure 1, we can conclude that:

1. The obtained linear model with batch normalization indeed achieves a uniform margin.
2. Batch normalization plays a key role in determining the implicit regularization effect. Without batch normalization, the obtained linear model does not achieve a uniform margin.

Clearly, our simulation results match our theory well.

In Figure 2, we present experiment results for VGG-16 with/without batch normalization trained by stochastic gradient descent to classify cat and dog images in the CIFAR-10 data set. We focus on the last hidden layer of VGG-16, and for each neuron on this layer, we estimate the margin uniformity over (i) all activated data from class cat; (ii) all activated data from class dog; (iii) all activated data from both classes. The final result is then calculated by taking an average over all neurons. Note that in these experiments, the neural network model, the data and the training algorithm all mismatch the exact setting in Theorem 3.2. Nevertheless, the experiment results still corroborate our theory to a certain extent.



(a) Linear model with batch normalization (b) Linear model without batch normalization

Figure 1: The values of $y_i \cdot \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle$, $i = 1, \ldots, n$ during the training of linear models with and without batch normalization. Each curve in the figures corresponds to a specific $i \in \{1, \ldots, n\}$ and illustrates the dynamics of a specific inner product $y_i \cdot \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle$. (a) gives the result for the linear model with batch normalization; (b) gives the result for the linear model without batch normalization. For both settings, we set the sample size $n = 50$ and dimension $d = 1000$. For $i = 1, \ldots, 50$, we generate the data inputs $\mathbf{x}_i$ independently as standard Gaussian random vectors, and set y_i randomly as $+1$ or -1 with equal probability.

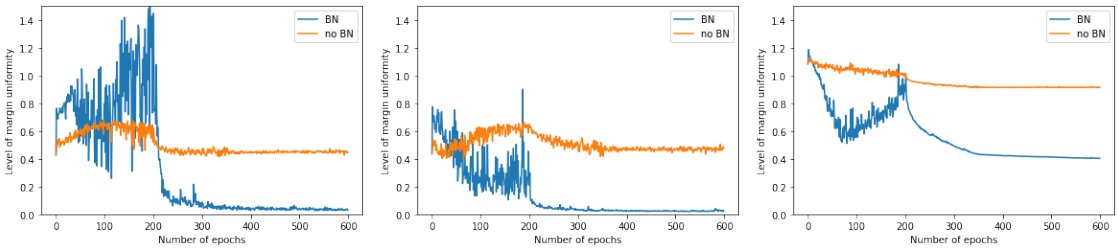
Appendix C. Proofs for Batch Normalization in Linear Models

In this section we present the proofs of the lemmas in Section 4. Combining these proofs with the discussion given in Section 4 would give the complete proof of Theorem 2.2.

C.1. Proof of Lemma 4.1

Proof [Proof of Lemma 4.1] By definition, we have

$$\nabla_{\mathbf{w}} L(\mathbf{w}, \gamma) = \frac{1}{n} \sum_{i=1}^{n} \ell'[y_i \cdot f(\mathbf{w}, \gamma, \mathbf{x}_i)] \cdot y_i \cdot \nabla_{\mathbf{w}} f(\mathbf{w}, \gamma, \mathbf{x}_i).$$



(*a*) Dog (*b*) Cat (*c*) Both

Figure 2: Margin uniformity of the last hidden layer in VGG-16 with/without batch normalization. The model is trained to classify cat and dog images in the CIFAR-10 data set. For each neuron on last hidden layer, treating it as a function $\mathrm{neuron}(\mathbf{x})$ of the data input $\mathbf{x}$, we use $\frac{\frac{1}{n^2}\sum_{i,i'\in\mathcal{I}}[y_{i'}\cdot\mathrm{neuron}(\mathbf{x}_{i'})-y_i\cdot\mathrm{neuron}(\mathbf{x}_i)]^2}{\frac{1}{n}\sum_{i\in\mathcal{I}}[\mathrm{neuron}(\mathbf{x}_i)]^2}$ to measure the margin uniformity, where we set $\mathcal{I}$ as the index set of (a) all activated data in the class cat, (b) all activated data in the class dog, and (c) all activated data in both classes. The plots are for the average of this quantity over all hidden neurons. Clearly, a smaller value implies a higher level of margin uniformity.

Then by the definition of the linear predictor with batch normalization, we have the following calculation using chain rule:

$$\begin{aligned}
\nabla_{\mathbf{w}} L(\mathbf{w},\gamma) &= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-1}\cdot\frac{1}{n}\sum_{i=1}^{n}\ell'[y_i\cdot f(\mathbf{w},\gamma,\mathbf{x}_i)]\cdot y_i\cdot\gamma\cdot\Big(\mathbf{I}-\|\mathbf{w}\|_{\mathbf{\Sigma}}^{-2}\cdot\mathbf{\Sigma}\mathbf{w}\mathbf{w}^\top\Big)\mathbf{x}_i\\
&= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3}\cdot\frac{\gamma}{n}\sum_{i=1}^{n}\ell_i'\cdot y_i\cdot\Big(\|\mathbf{w}\|_{\mathbf{\Sigma}}^{2}\cdot\mathbf{I}-\mathbf{\Sigma}\mathbf{w}\mathbf{w}^\top\Big)\mathbf{x}_i\\
&= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3}\cdot\frac{\gamma}{n}\sum_{i=1}^{n}\ell_i'\cdot\left(\frac{1}{n}\sum_{i'=1}^{n}\langle\mathbf{w},\mathbf{z}_{i'}\rangle^2\cdot\mathbf{I}-\frac{1}{n}\sum_{i'=1}^{n}\langle\mathbf{w},\mathbf{z}_{i'}\rangle\cdot\mathbf{z}_{i'}\mathbf{w}^\top\right)\mathbf{z}_i\\
&= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3}\cdot\frac{\gamma}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\ell_i'\cdot\langle\mathbf{w},\mathbf{z}_{i'}\rangle^2\cdot\mathbf{z}_i-\|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3}\cdot\frac{\gamma}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\ell_i'\cdot\langle\mathbf{w},\mathbf{z}_{i'}\rangle\cdot\langle\mathbf{w},\mathbf{z}_i\rangle\cdot\mathbf{z}_{i'},
\end{aligned}$$

where we remind readers that $\mathbf{z}_i = y_i\cdot\mathbf{x}_i$, $i\in[n]$. By Assumption 2.1, taking inner product with $\mathbf{w}^*$ on both sides above then gives

$$\langle\nabla_{\mathbf{w}}L(\mathbf{w},\gamma),\mathbf{w}^*\rangle = \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3}\cdot\frac{\gamma}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\ell_i'\cdot\langle\mathbf{w},\mathbf{z}_{i'}\rangle^2-\|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3}\cdot\frac{\gamma}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\ell_i'\cdot\langle\mathbf{w},\mathbf{z}_{i'}\rangle\cdot\langle\mathbf{w},\mathbf{z}_i\rangle.$$

Further denote $u_i = \langle\mathbf{w},\mathbf{z}_i\rangle$ for $i\in[n]$. Then we have

$$\langle\nabla_{\mathbf{w}}L(\mathbf{w},\gamma),\mathbf{w}^*\rangle = \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3}\cdot\frac{\gamma}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\ell_i'\cdot u_{i'}^2-\|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3}\cdot\frac{\gamma}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\ell_i'\cdot u_{i'}u_i. \tag{C.1}$$

Switching the index notations i, i' in the above equation also gives

$$\langle\nabla_{\mathbf{w}}L(\mathbf{w},\gamma),\mathbf{w}^*\rangle = \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3}\cdot\frac{\gamma}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\ell_{i'}'\cdot u_i^2-\|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3}\cdot\frac{\gamma}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\ell_{i'}'\cdot u_{i'}u_i. \tag{C.2}$$

We can add (C.1) and (C.2) together to obtain

$$2 \cdot \langle \nabla_{\mathbf{w}} L(\mathbf{w}, \gamma), \mathbf{w}^* \rangle = \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2} \sum_{i=1}^{n} \sum_{i'=1}^{n} (\ell_i' \cdot u_{i'}^2 - \ell_i' \cdot u_{i'} u_i + \ell_{i'}' \cdot u_i^2 - \ell_{i'}' \cdot u_{i'} u_i)$$
$$= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2} \sum_{i=1}^{n} \sum_{i'=1}^{n} (u_{i'} - u_i)(\ell_i' \cdot u_{i'} - \ell_{i'}' \cdot u_i).$$

Note that by definition we have $\ell_i' < 0$. Therefore,

$$-\langle \nabla_{\mathbf{w}} L(\mathbf{w}, \gamma), \mathbf{w}^* \rangle = \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{2n^2} \sum_{i=1}^{n} \sum_{i'=1}^{n} |\ell_i'| \cdot |\ell_{i'}'| \cdot (u_{i'} - u_i)(|\ell_{i'}'|^{-1} \cdot u_{i'} - |\ell_i'|^{-1} \cdot u_i).$$

This completes the proof. ■

C.2. Proof of Lemma 4.2

Proof [Proof of Lemma 4.2] By Lemma 4.1 and the gradient descent update rule, we have

$$\langle \mathbf{w}^{(t+1)}, \mathbf{w}^* \rangle - \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle = \eta \cdot \langle -\nabla_{\mathbf{w}} L(\mathbf{w}^{(t)}, \gamma^{(t)}), \mathbf{w}^* \rangle$$
$$= \frac{\gamma^{(t)} \eta}{2n^2 \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \sum_{i,i'=1}^{n} |\ell_i'^{(t)}| \cdot |\ell_{i'}'^{(t)}| \cdot (\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle) \cdot (|\ell_{i'}'^{(t)}|^{-1} \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - |\ell_i'^{(t)}|^{-1} \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle) \tag{C.3}$$

for all $t \geq 0$. Note that $\ell(z) = \log[1 + \exp(-z)]$, $\ell'(z) = -1/[1 + \exp(z)]$ and

$$|\ell_i'^{(t)}|^{-1} = -\{\ell'[y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)]\}^{-1} = 1 + \exp(\gamma^{(t)} \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle / \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}).$$

Denote $F_i = \gamma^{(t)} \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle / \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}$ for $i \in [n]$. Then we have

$$\begin{aligned}
\frac{(|\ell_{i'}'^{(t)}|^{-1} \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - |\ell_i'^{(t)}|^{-1} \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle)}{\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle} &= \frac{(|\ell_{i'}'^{(t)}|^{-1} \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - |\ell_i'^{(t)}|^{-1} \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle) \cdot \gamma^{(t)} / \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}}{(\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle) \cdot \gamma^{(t)} / \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \\
&= \frac{F_{i'} \cdot [1 + \exp(F_{i'})] - F_i \cdot [1 + \exp(F_i)]}{F_{i'} - F_i} \\
&= 1 + \frac{F_{i'} \cdot \exp(F_{i'}) - F_i \cdot \exp(F_i)}{F_{i'} - F_i} \\
&\geq 1 + \exp(\max\{F_{i'}, F_i\}) \\
&= \max\{|\ell_i'^{(t)}|^{-1}, |\ell_{i'}'^{(t)}|^{-1}\},
\end{aligned} \tag{C.4}$$

where the inequality follows by the fact that $[a \exp(a) - b \exp(b)]/(a - b) \geq \exp(\max\{a, b\})$ for all $a \neq b$. Plugging (C.4) into (C.3) gives

$$\begin{aligned}
&\langle \mathbf{w}^{(t+1)}, \mathbf{w}^* \rangle - \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle \\
&\quad \geq \frac{\gamma^{(t)} \eta}{2n^2 \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \sum_{i,i'=1}^{n} |\ell_i'^{(t)}| \cdot |\ell_{i'}'^{(t)}| \cdot (\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle)^2 \cdot \max\{|\ell_i'^{(t)}|^{-1}, |\ell_{i'}'^{(t)}|^{-1}\} \\
&\quad = \frac{\gamma^{(t)} \eta}{2n^2 \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \sum_{i,i'=1}^{n} \max\{|\ell_i'^{(t)}|, |\ell_{i'}'^{(t)}|\} \cdot (\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle)^2.
\end{aligned} \tag{C.5}$$

Now by Lemma C.1, we have

$$\begin{aligned}&\langle \mathbf{w}^{(t+1)}, \mathbf{w}^*\rangle - \langle \mathbf{w}^{(t)}, \mathbf{w}^*\rangle \\ &\qquad \geq \frac{\gamma^{(t)}\eta}{8n^2\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \cdot \left(\frac{1}{n}\sum_{i=1}^n |\ell_i'^{(t)}|\right) \cdot \sum_{i,i'=1}^n (\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}\rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle)^2. \end{aligned} \tag{C.6}$$

Moreover, by definition we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n |\ell_i'^{(t)}| &= \frac{1}{n}\sum_{i=1}^n \left[1 + \exp\left(\frac{\gamma^{(t)} \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_j\rangle}{\sqrt{\frac{1}{n}\sum_{j=1}^n \langle \mathbf{w}^{(t)}, \mathbf{z}_j\rangle^2}}\right)\right]^{-1} \\ &\geq \frac{1}{n}\sum_{i=1}^n \left[1 + \exp\left(\frac{\gamma^{(t)} \cdot |\langle \mathbf{w}^{(t)}, \mathbf{z}_j\rangle|}{\sqrt{\frac{1}{n}\sum_{j=1}^n \langle \mathbf{w}^{(t)}, \mathbf{z}_j\rangle^2}}\right)\right]^{-1},\end{aligned}$$

where the inequality follows by the fact that $[1 + \exp(z)]^{-1}$ is a decreasing function. Further note that $[1 + \exp(z)]^{-1}$ is convex over $z \geq 0$. Therefore by Jensen's inequality, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n |\ell_i'^{(t)}| &\geq \frac{1}{n}\sum_{i=1}^n \left[1 + \exp\left(\frac{\gamma^{(t)} \cdot |\langle \mathbf{w}^{(t)}, \mathbf{z}_j\rangle|}{\sqrt{\frac{1}{n}\sum_{j=1}^n \langle \mathbf{w}^{(t)}, \mathbf{z}_j\rangle^2}}\right)\right]^{-1} \\ &\geq \left[1 + \exp\left(\frac{1}{n}\sum_{i=1}^n \frac{\gamma^{(t)} \cdot |\langle \mathbf{w}^{(t)}, \mathbf{z}_j\rangle|}{\sqrt{\frac{1}{n}\sum_{j=1}^n \langle \mathbf{w}^{(t)}, \mathbf{z}_j\rangle^2}}\right)\right]^{-1} \\ &\geq [1 + \exp(\gamma^{(t)})]^{-1} \\ &\geq \exp(-\gamma^{(t)})/2.\end{aligned} \tag{C.7}$$

Plugging (C.7) into (C.6) then gives

$$\begin{aligned}\langle \mathbf{w}^{(t+1)}, \mathbf{w}^*\rangle - \langle \mathbf{w}^{(t)}, \mathbf{w}^*\rangle &\geq \frac{\gamma^{(t)}\eta}{16n^2\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \cdot \exp(-\gamma^{(t)}) \cdot \sum_{i,i'=1}^n (\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}\rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle)^2 \\ &= \frac{\gamma^{(t)}\eta}{16\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \cdot \exp(-\gamma^{(t)}) \cdot D(\mathbf{w}^{(t)})\end{aligned} \tag{C.8}$$

Moreover, we can simply utilize (C.5) to obtain

$$\langle \mathbf{w}^{(t+1)}, \mathbf{w}^*\rangle - \langle \mathbf{w}^{(t)}, \mathbf{w}^*\rangle \geq \frac{\gamma^{(t)}\eta}{2n^2\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \cdot \min_i |\ell_i'^{(t)}| \cdot \sum_{i,i'=1}^n (\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}\rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle)^2 \tag{C.9}$$

$$= \frac{\gamma^{(t)}\eta}{2\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \cdot \min_i |\ell_i'^{(t)}| \cdot D(\mathbf{w}^{(t)}) \tag{C.10}$$

Combining (C.8) and (C.10) finishes the proof. ■

C.3. Proof of Lemma 4.3

Proof [Proof of Lemma 4.3] We note that the following identity holds:

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i,i'=1}^{n}(\langle \mathbf{w},\mathbf{z}_{i'}\rangle - \langle \mathbf{w},\mathbf{z}_i\rangle)^2 \\
&\quad= \frac{1}{n^2}\sum_{i,i'=1}^{n}(\langle \mathbf{w},\mathbf{z}_{i'}\rangle^2 - 2\cdot\langle \mathbf{w},\mathbf{z}_{i'}\rangle\cdot\langle \mathbf{w},\mathbf{z}_i\rangle + \langle \mathbf{w},\mathbf{z}_i\rangle^2) \\
&\quad= \frac{2}{n}\sum_{i}^{n}\langle \mathbf{w},\mathbf{z}_i\rangle^2 - 2\cdot\left(\frac{1}{n}\sum_{i}^{n}\langle \mathbf{w},\mathbf{z}_i\rangle\right)^2 \\
&\quad= \frac{2}{n}\sum_{i}^{n}\langle \mathbf{w},\mathbf{z}_i\rangle^2 - 4\cdot\left(\frac{1}{n}\sum_{i}^{n}\langle \mathbf{w},\mathbf{z}_i\rangle\right)^2 + 2\cdot\left(\frac{1}{n}\sum_{i}^{n}\langle \mathbf{w},\mathbf{z}_i\rangle\right)^2 \\
&\quad= \frac{2}{n}\sum_{i}^{n}\langle \mathbf{w},\mathbf{z}_i\rangle^2 - \frac{4}{n}\sum_{i}^{n}\langle \mathbf{w},\mathbf{z}_i\rangle\cdot\left(\frac{1}{n}\sum_{i}^{n}\langle \mathbf{w},\mathbf{z}_i\rangle\right) + \frac{2}{n}\sum_{i}^{n}\left(\frac{1}{n}\sum_{i}^{n}\langle \mathbf{w},\mathbf{z}_i\rangle\right)^2 \\
&\quad= \frac{2}{n}\sum_{i}^{n}\left(\langle \mathbf{w},\mathbf{z}_i\rangle - \frac{1}{n}\sum_{i'}^{n}\langle \mathbf{w},\mathbf{z}_{i'}\rangle\right)^2 \\
&\quad= \frac{2}{n}\left\|\mathbf{w}^\top\mathbf{Z}\left(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)\right\|_2^2 \\
&\quad= \frac{2}{n}\mathbf{w}^\top\mathbf{Z}\left(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)\mathbf{Z}^\top\mathbf{w}.
\end{aligned}$$

It is easy to see that the null space of $\mathbf{I}-\mathbf{1}\mathbf{1}^\top/n$ is $\mathrm{span}\{\mathbf{1}\}$, and the non-zero eigenvalues of $\mathbf{I}-\mathbf{1}\mathbf{1}^\top/n$ are all 1's. Moreover, we note that the projection of the vector $\mathbf{Z}^\top\mathbf{w}$ onto the space $\mathrm{span}\{\mathbf{1}\}^\perp$ is

$$\mathbf{Z}^\top\mathbf{w} - \mathbf{1}\mathbf{1}^\top\mathbf{Z}^\top\mathbf{w}/n = \mathbf{Z}^\top\mathbf{w} - \mathbf{Z}\mathbf{w}^*\mathbf{w}^{*\top}\mathbf{Z}^\top\mathbf{Z}^\top\mathbf{w}/n = \mathbf{Z}^\top\mathbf{w} - \mathbf{Z}\mathbf{w}^*\cdot\langle \mathbf{w}^{*\top},\mathbf{w}\rangle_{\mathbf{\Sigma}},$$

where we utilize the property that $\mathbf{Z}\mathbf{w}^* = \mathbf{1}$ to obtain the first equality. Therefore, we have

$$\begin{aligned}
\frac{1}{n^2}\sum_{i,i'=1}^{n}(\langle \mathbf{w},\mathbf{z}_{i'}\rangle - \langle \mathbf{w},\mathbf{z}_i\rangle)^2 &= \frac{2}{n}(\mathbf{w}-\langle \mathbf{w}^*,\mathbf{w}\rangle_{\mathbf{\Sigma}}\cdot\mathbf{w}^*)^\top\mathbf{Z}\mathbf{Z}^\top(\mathbf{w}-\langle \mathbf{w}^*,\mathbf{w}\rangle_{\mathbf{\Sigma}}\cdot\mathbf{w}^*) \\
&= 2\cdot\|\mathbf{w}-\langle \mathbf{w}^*,\mathbf{w}\rangle_{\mathbf{\Sigma}}\cdot\mathbf{w}^*\|_{\mathbf{\Sigma}}^2.
\end{aligned}$$

This finishes the proof of the first result. For the second result, we first note that by definition, $\|\mathbf{w}^*\|_{\mathbf{\Sigma}} = 1$, and therefore for any $\mathbf{w}\in\mathbb{R}^d$, $\langle \mathbf{w}^*,\mathbf{w}\rangle_{\mathbf{\Sigma}}\cdot\mathbf{w}^*$ is the projection of $\mathbf{w}$ onto $\mathrm{span}\{\mathbf{w}^*\}$ under the inner product $\langle\cdot,\cdot\rangle_{\mathbf{\Sigma}}$. Therefore we have

$$\|\mathbf{w}-\langle \mathbf{w}^*,\mathbf{w}\rangle_{\mathbf{\Sigma}}\cdot\mathbf{w}^*\|_{\mathbf{\Sigma}}^2 \le \|\mathbf{w}-c\cdot\mathbf{w}^*\|_{\mathbf{\Sigma}}^2$$

for all $c\in\mathbb{R}$, and hence

$$\|\mathbf{w}-\langle \mathbf{w}^*,\mathbf{w}\rangle_{\mathbf{\Sigma}}\cdot\mathbf{w}^*\|_{\mathbf{\Sigma}}^2 \le \left\|\mathbf{w}-\frac{\langle \mathbf{w}^*,\mathbf{w}\rangle}{\|\mathbf{w}^*\|_2^2}\cdot\mathbf{w}^*\right\|_{\mathbf{\Sigma}}^2 \le \lambda_{\max}\cdot\left\|\mathbf{w}-\frac{\langle \mathbf{w}^*,\mathbf{w}\rangle}{\|\mathbf{w}^*\|_2^2}\cdot\mathbf{w}^*\right\|_2^2.$$

Similarly, we note that $\|\mathbf{w}^*\|_2^{-2} \cdot \langle \mathbf{w}^*, \mathbf{w} \rangle \cdot \mathbf{w}^*$ is the projection of $\mathbf{w}$ onto $\mathrm{span}\{\mathbf{w}^*\}$ under the Euclidean inner product $\langle \cdot, \cdot \rangle$. Therefore we have

$$\lambda_{\min} \cdot \left\| \mathbf{w} - \frac{\langle \mathbf{w}^*, \mathbf{w} \rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^* \right\|_2^2 \leq \lambda_{\min} \cdot \|\mathbf{w} - \langle \mathbf{w}^*, \mathbf{w} \rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_2^2 \leq \|\mathbf{w} - \langle \mathbf{w}^*, \mathbf{w} \rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_{\mathbf{\Sigma}}^2.$$

This completes the proof. ■

C.4. Proof of Lemma 4.4

We first present the following technical lemma.

Lemma C.1 *Let $\{a_i\}_{i=1,\ldots,n}$ and $\{b_i\}_{i=1,\ldots,n}$ be two sequences that satisfy*

$$a_1 \leq a_2 \leq \cdots \leq a_n; \quad b_1 \geq b_2 \geq \cdots \geq b_n \geq 0.$$

Then it holds that

$$\sum_{i,i'=1}^{n} \max\{b_i, b_{i'}\} \cdot (a_i - a_{i'})^2 \geq \frac{\sum_{i=1}^{n} b_i}{4n} \cdot \sum_{i,i'=1}^{n} (a_i - a_{i'})^2.$$

Based on Lemma C.1, the proof of Lemma 4.4 is as follows.

Proof [Proof of Lemma 4.4] By the gradient descent update rule, we have

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \cdot \nabla_{\mathbf{w}} L(\mathbf{w}^{(t)}, \gamma^{(t)}). \tag{C.11}$$

Note that we have the calculation

$$\nabla_{\mathbf{w}} L(\mathbf{w}^{(t)}, \gamma^{(t)}) = \frac{1}{n \cdot \|\mathbf{w}\|_{\mathbf{\Sigma}}} \sum_{i=1}^{n} \ell'[y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)] \cdot y_i \cdot \gamma^{(t)} \cdot \left(\mathbf{I} - \frac{\mathbf{\Sigma} \mathbf{w}^{(t)} \mathbf{w}^{(t)\top}}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \right) \mathbf{x}_i.$$

It is easy to see that $\mathbf{w}^{(t)}$ is orthogonal to $\nabla_{\mathbf{w}} L(\mathbf{w}^{(t)}, \gamma^{(t)})$. Therefore, taking $\|\cdot\|_2^2$ on both sides of (C.11) gives

$$\|\mathbf{w}^{(t+1)}\|_2^2 = \|\mathbf{w}^{(t)}\|_2^2 + \eta^2 \cdot \|\nabla_{\mathbf{w}} L(\mathbf{w}^{(t)}, \gamma^{(t)})\|_2^2. \tag{C.12}$$

Therefore we directly conclude that $\|\mathbf{w}^{(t)}\|_2^2 \le \|\mathbf{w}^{(t+1)}\|_2^2$. Besides, plugging in the calculation of $\nabla_{\mathbf{w}} L(\mathbf{w}^{(t)}, \gamma^{(t)})$ also gives

$$
\begin{aligned}
\|\mathbf{w}^{(t+1)}\|_2^2 &= \|\mathbf{w}^{(t)}\|_2^2 + \eta^2 \cdot \left\| \frac{1}{n \cdot \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \sum_{i=1}^n \ell'[y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)] \cdot y_i \cdot \gamma^{(t)} \cdot \left(\mathbf{I} - \frac{\mathbf{\Sigma}\mathbf{w}^{(t)}\mathbf{w}^{(t)\top}}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \right) \mathbf{x}_i \right\|_2^2 \\
&\le \|\mathbf{w}^{(t)}\|_2^2 + \eta^2 \cdot \frac{\gamma^{(t)2}}{n \cdot \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \sum_{i=1}^n \left\| \left(\mathbf{I} - \frac{\mathbf{\Sigma}\mathbf{w}^{(t)}\mathbf{w}^{(t)\top}}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \right) \mathbf{x}_i \right\|_2^2 \\
&\le \|\mathbf{w}^{(t)}\|_2^2 + \eta^2 \cdot \frac{\gamma^{(t)2} \cdot \max_i \|\mathbf{x}_i\|_2^2}{n \cdot \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \cdot \sum_{i=1}^n \left(1 + \left\| \frac{\mathbf{\Sigma}\mathbf{w}^{(t)}\mathbf{w}^{(t)\top}}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \right\|_2 \right)^2 \\
&= \|\mathbf{w}^{(t)}\|_2^2 + \eta^2 \cdot \frac{\gamma^{(t)2} \cdot \max_i \|\mathbf{x}_i\|_2^2}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \cdot \left(1 + \frac{\|\mathbf{\Sigma}\mathbf{w}^{(t)}\|_2 \cdot \|\mathbf{w}^{(t)}\|_2}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \right)^2 \\
&\le \|\mathbf{w}^{(t)}\|_2^2 + 4\eta^2 \cdot \frac{\gamma^{(t)} \cdot \max_i \|\mathbf{x}_i\|_2^2}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \cdot \left(\frac{\|\mathbf{\Sigma}\mathbf{w}^{(t)}\|_2 \cdot \|\mathbf{w}^{(t)}\|_2}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \right)^2 \\
&= \|\mathbf{w}^{(t)}\|_2^2 + 4\eta^2 \cdot \frac{\gamma^{(t)2} \cdot \max_i \|\mathbf{x}_i\|_2^2}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^6} \cdot \|\mathbf{\Sigma}\mathbf{w}^{(t)}\|_2^2 \cdot \|\mathbf{w}^{(t)}\|_2^2,
\end{aligned}
$$

where the first inequality follows by Jensen's inequality, and the last inequality follows by the fact that $\|\mathbf{w}\|_{\mathbf{\Sigma}}^2 = \langle \mathbf{w}, \mathbf{\Sigma}\mathbf{w} \rangle \le \|\mathbf{\Sigma}\mathbf{w}\|_2 \cdot \|\mathbf{w}\|_2$. Further plugging in the definition of $\mathbf{\Sigma}$ gives

$$
\begin{aligned}
\|\mathbf{w}^{(t+1)}\|_2^2 &\le \|\mathbf{w}^{(t)}\|_2^2 + 4\eta^2 \cdot \frac{\gamma^{(t)2} \cdot \max_i \|\mathbf{x}_i\|_2^2}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^6} \cdot \|\mathbf{w}^{(t)}\|_2^2 \cdot \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \mathbf{w}^{(t)} \right\|_2^2 \\
&\le \|\mathbf{w}^{(t)}\|_2^2 + 4\eta^2 \cdot \frac{\gamma^{(t)2} \cdot \max_i \|\mathbf{x}_i\|_2^3}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^6} \cdot \|\mathbf{w}^{(t)}\|_2^2 \cdot \left(\frac{1}{n} \sum_{i=1}^n |\langle \mathbf{x}_i, \mathbf{w}^{(t)} \rangle| \right)^2 \\
&\le \|\mathbf{w}^{(t)}\|_2^2 + 4\eta^2 \cdot \frac{\gamma^{(t)2} \cdot \max_i \|\mathbf{x}_i\|_2^3}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^6} \cdot \|\mathbf{w}^{(t)}\|_2^2 \cdot \frac{1}{n} \sum_{i=1}^n \langle \mathbf{x}_i, \mathbf{w}^{(t)} \rangle^2 \\
&= \|\mathbf{w}^{(t)}\|_2^2 + 4\eta^2 \cdot \frac{\gamma^{(t)2} \cdot \max_i \|\mathbf{x}_i\|_2^3}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^4} \cdot \|\mathbf{w}^{(t)}\|_2^2 \\
&\le \|\mathbf{w}^{(t)}\|_2^2 + 4\eta^2 \cdot \frac{\gamma^{(t)2} \cdot \max_i \|\mathbf{x}_i\|_2^3}{\lambda_{\min}^2 \cdot \|\mathbf{w}^{(t)}\|_2^4} \cdot \|\mathbf{w}^{(t)}\|_2^2 \\
&= \|\mathbf{w}^{(t)}\|_2^2 + 4\eta^2 \cdot \frac{\gamma^{(t)2} \cdot \max_i \|\mathbf{x}_i\|_2^3}{\lambda_{\min}^2 \cdot \|\mathbf{w}^{(t)}\|_2^2}.
\end{aligned}
$$

This finishes the proof of the first result.

To prove the second result in the lemma, we first denote $\hat{\mathbf{w}}_t^* = \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*$, and define

$$
\mathcal{B}^{(t)} := \{ \mathbf{w} \in \mathbb{R}^d : \|\mathbf{w} - \mathbf{w}^{(t)}\|_2 \le \|\mathbf{w}^{(t)}\|_2/2 \}.
$$

Then the condition $\|\mathbf{w}^{(t)} - \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle \cdot \|\mathbf{w}^*\|_2^{-1} \cdot \mathbf{w}^*\|_2 \le \|\mathbf{w}^{(0)}\|_2/2$ and the result in the first part that $\|\mathbf{w}^{(t)}\|_2^2 \le \|\mathbf{w}^{(t+1)}\|_2^2$ for all $t \ge 0$ imply that

$$
\|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2 \le \|\mathbf{w}^{(0)}\|_2/2 \le \|\mathbf{w}^{(t)}\|_2/2,
$$

and therefore

$$\hat{\mathbf{w}}_t^* \in \mathcal{B}^{(t)}. \tag{C.13}$$

It is also clear that under this condition we have $\langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle > 0$.

We proceed to derive an upper bound of $\|\nabla_{\mathbf{w}} L(\mathbf{w}^{(t)}, \gamma^{(t)})\|_2$. By the definition of $\mathbf{w}^*, \hat{\mathbf{w}}_t^*$, the positive homogeneity of f in $\mathbf{w}$ and the fact that $\langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle > 0$, it is easy to see that

$$y_i \cdot f(\hat{\mathbf{w}}_t^*, \gamma^{(t)}, \mathbf{x}_i) = y_i \cdot f(\langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*, \gamma^{(t)}, \mathbf{x}_i) = y_i \cdot f(\mathbf{w}^*, \gamma^{(t)}, \mathbf{x}_i) = \gamma^{(t)}$$

for all $i \in [n]$. Therefore, denoting $\ell_*'^{(t)} = \ell'(\gamma^{(t)})$, then we have $\ell_*'^{(t)} = \ell'[y_i \cdot f(\langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*, \gamma^{(t)}, \mathbf{x}_i)] = \ell'[y_i \cdot f(\hat{\mathbf{w}}_t^*, \gamma^{(t)}, \mathbf{x}_i)]$ for all $i \in [n]$. Moreover, we have

$$\begin{aligned}
\|\nabla_{\mathbf{w}} L(\mathbf{w}^{(t)}, \gamma^{(t)})\|_2 &\leq \bigg\| \frac{\gamma^{(t)}}{n \cdot \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \sum_{i=1}^n \ell_*'^{(t)} \cdot y_i \cdot \bigg(\mathbf{I} - \frac{\mathbf{\Sigma} \mathbf{w}^{(t)} \mathbf{w}^{(t)\top}}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \bigg) \mathbf{x}_i \bigg\|_2 \\
&\quad + \bigg\| \frac{\gamma^{(t)}}{n \cdot \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \sum_{i=1}^n (\ell_i'^{(t)} - \ell_*'^{(t)}) \cdot y_i \cdot \bigg(\mathbf{I} - \frac{\mathbf{\Sigma} \mathbf{w}^{(t)} \mathbf{w}^{(t)\top}}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \bigg) \mathbf{x}_i \bigg\|_2 \\
&\leq \bigg\| \frac{\gamma^{(t)}}{n \cdot \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \sum_{i=1}^n \ell_*'^{(t)} \cdot y_i \cdot \bigg(\mathbf{I} - \frac{\mathbf{\Sigma} \mathbf{w}^{(t)} \mathbf{w}^{(t)\top}}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \bigg) \mathbf{x}_i \bigg\|_2 \\
&\quad + \frac{\gamma^{(t)}}{n \cdot \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \sum_{i=1}^n |\ell_i'^{(t)} - \ell_*'^{(t)}| \cdot \bigg\| \bigg(\mathbf{I} - \frac{\mathbf{\Sigma} \mathbf{w}^{(t)} \mathbf{w}^{(t)\top}}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \bigg) \mathbf{x}_i \bigg\|_2 \\
&\leq \bigg\| \frac{\gamma^{(t)}}{n \cdot \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \sum_{i=1}^n \ell_*'^{(t)} \cdot y_i \cdot \bigg(\mathbf{I} - \frac{\mathbf{\Sigma} \mathbf{w}^{(t)} \mathbf{w}^{(t)\top}}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \bigg) \mathbf{x}_i \bigg\|_2 \\
&\quad + \frac{\gamma^{(t)}}{n \cdot \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \sum_{i=1}^n |\ell_i'^{(t)} - \ell_*'^{(t)}| \cdot \|\mathbf{x}_i\|_2 \\
&= \underbrace{\frac{\gamma^{(t)} \cdot |\ell_*'^{(t)}|}{n \cdot \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \cdot \bigg\| \sum_{i=1}^n \bigg(\mathbf{I} - \frac{\mathbf{\Sigma} \mathbf{w}^{(t)} \mathbf{w}^{(t)\top}}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \bigg) \mathbf{z}_i \bigg\|_2}_{I_1} \\
&\quad + \underbrace{\frac{\gamma^{(t)}}{n \cdot \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \sum_{i=1}^n |\ell_i'^{(t)} - \ell_*'^{(t)}| \cdot \|\mathbf{x}_i\|_2}_{I_2} .
\end{aligned} \tag{C.14}$$

In the following, we bound I_1 and I_2 separately. For I_1, we have

$$\begin{aligned}
\sum_{i=1}^n \bigg(\mathbf{I} - \frac{\mathbf{\Sigma}\mathbf{w}^{(t)}\mathbf{w}^{(t)\top}}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2}\bigg)\mathbf{z}_i &= \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^{-2} \cdot \sum_{i=1}^n \Big(\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2 \cdot \mathbf{I} - \mathbf{\Sigma}\mathbf{w}^{(t)}\mathbf{w}^{(t)\top}\Big)\mathbf{z}_i \\
&= \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^{-2} \cdot \sum_{i=1}^n \bigg(\frac{1}{n}\sum_{i'=1}^n \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}\rangle^2 \cdot \mathbf{I} - \frac{1}{n}\sum_{i'=1}^n \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}\rangle \cdot \mathbf{z}_{i'}\mathbf{w}^{(t)\top}\bigg)\mathbf{z}_i \\
&= \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^{-2} \cdot \frac{1}{n}\sum_{i,i'=1}^n \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}\rangle^2 \cdot \mathbf{z}_i - \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^{-2} \cdot \frac{1}{n}\sum_{i,i'=1}^n \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}\rangle \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle \cdot \mathbf{z}_{i'} \\
&= \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^{-2} \cdot \frac{1}{n}\sum_{i,i'=1}^n \langle \mathbf{w}^{(t)}, \mathbf{z}_{i}\rangle^2 \cdot \mathbf{z}_{i'} - \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^{-2} \cdot \frac{1}{n}\sum_{i,i'=1}^n \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}\rangle \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle \cdot \mathbf{z}_{i'} \\
&= \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^{-2} \cdot \frac{1}{n}\sum_{i,i'=1}^n \big(\langle \mathbf{w}^{(t)}, \mathbf{z}_{i}\rangle^2 - \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}\rangle \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle\big) \cdot \mathbf{z}_{i'}.
\end{aligned} \tag{C.15}$$

Moreover, by the definition of $\hat{\mathbf{w}}_t^*$, it is clear that

$$\langle \hat{\mathbf{w}}_t^*, \mathbf{z}_i\rangle^2 - \langle \hat{\mathbf{w}}_t^*, \mathbf{z}_{i'}\rangle \cdot \langle \hat{\mathbf{w}}_t^*, \mathbf{z}_i\rangle = 0.$$

Therefore we have

$$\begin{aligned}
&\langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle^2 - \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}\rangle \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle \\
&\quad = \langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle^2 - \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}\rangle \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle - \langle \hat{\mathbf{w}}_t^*, \mathbf{z}_i\rangle^2 + \langle \hat{\mathbf{w}}_t^*, \mathbf{z}_{i'}\rangle \cdot \langle \hat{\mathbf{w}}_t^*, \mathbf{z}_i\rangle \\
&\quad = \big(\langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle^2 - \langle \hat{\mathbf{w}}_t^*, \mathbf{z}_i\rangle^2\big) - \big(\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}\rangle \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle - \langle \hat{\mathbf{w}}_t^*, \mathbf{z}_{i'}\rangle \cdot \langle \hat{\mathbf{w}}_t^*, \mathbf{z}_i\rangle\big) \\
&\quad = \big(\langle \mathbf{w}^{(t)} + \hat{\mathbf{w}}_t^*, \mathbf{z}_i\rangle\big) \cdot \big(\langle \mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*, \mathbf{z}_i\rangle\big) - \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}\rangle \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{z}_i\rangle + \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{z}_{i'}\rangle \cdot \langle \hat{\mathbf{w}}_t^*, \mathbf{z}_i\rangle.
\end{aligned}$$

Taking absolute value on both sides and applying triangle inequality gives

$$\begin{aligned}
&|\langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle^2 - \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}\rangle \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle| \\
&\quad \le |\langle \mathbf{w}^{(t)} + \hat{\mathbf{w}}_t^*, \mathbf{z}_i\rangle| \cdot |\langle \mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*, \mathbf{z}_i\rangle| + |\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}\rangle| \cdot |\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{z}_i\rangle| + |\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{z}_{i'}\rangle| \cdot |\langle \hat{\mathbf{w}}_t^*, \mathbf{z}_i\rangle| \\
&\quad \le (\|\mathbf{w}^{(t)}\|_2 + \|\hat{\mathbf{w}}_t^*\|_2) \cdot \|\mathbf{z}_i\|_2 \cdot \|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2 \cdot \|\mathbf{z}_i\|_2 + \|\mathbf{w}^{(t)}\|_2 \cdot \|\mathbf{z}_{i'}\|_2 \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \cdot \|\mathbf{z}_i\|_2 \\
&\qquad + \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \cdot \|\mathbf{z}_{i'}\|_2 \cdot \|\hat{\mathbf{w}}_t^*\|_2 \cdot \|\mathbf{z}_i\|_2 \\
&\quad \le 2(\|\mathbf{w}^{(t)}\|_2 + \|\hat{\mathbf{w}}_t^*\|_2) \cdot \max_i \|\mathbf{z}_i\|_2^2 \cdot \|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2 \\
&\quad \le 4\|\mathbf{w}^{(t)}\|_2 \cdot \max_i \|\mathbf{z}_i\|_2^2 \cdot \|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2 \\
&\quad = 4\|\mathbf{w}^{(t)}\|_2 \cdot \max_i \|\mathbf{x}_i\|_2^2 \cdot \|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2.
\end{aligned} \tag{C.16}$$

Plugging (C.16) into (C.15) gives

$$
\begin{aligned}
I_1 &= \frac{\gamma^{(t)} \cdot |\ell_*'^{(t)}|}{n \cdot \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \cdot \bigg\| \sum_{i=1}^n \bigg(\mathbf{I} - \frac{\mathbf{\Sigma}\mathbf{w}^{(t)}\mathbf{w}^{(t)\top}}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \bigg) \mathbf{z}_i \bigg\|_2 \\
&\leq \frac{4\gamma^{(t)} \cdot |\ell_*'^{(t)}|}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \cdot \frac{\|\mathbf{w}^{(t)}\|_2}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \cdot \max_i \|\mathbf{x}_i\|_2^3 \cdot \|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2 \\
&\leq \frac{4\gamma^{(t)} \cdot \exp(-\gamma^{(t)})}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \cdot \frac{\|\mathbf{w}^{(t)}\|_2}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \cdot \max_i \|\mathbf{x}_i\|_2^3 \cdot \|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2 \\
&\leq \frac{4\gamma^{(t)} \cdot \exp(-\gamma^{(t)})}{\lambda_{\min}^{3/2} \cdot \|\mathbf{w}^{(t)}\|_2^2} \cdot \max_i \|\mathbf{x}_i\|_2^3 \cdot \|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2 \\
&\leq \frac{4\gamma^{(t)} \cdot \exp(-\gamma^{(t)})}{\lambda_{\min}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \max_i \|\mathbf{x}_i\|_2^3 \cdot \|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2 \qquad \text{(C.17)}
\end{aligned}
$$

where the second inequality follows by the fact that $-\exp(-\gamma^{(t)}) \leq \ell_*'^{(t)} = \ell'(\gamma^{(t)}) < 0$. Regarding the bound for I_2, by the mean value theorem, there exists z between $y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)$ and $y_i \cdot f(\hat{\mathbf{w}}_t^*, \gamma^{(t)}, \mathbf{x}_i) = \gamma^{(t)}$ such that

$$
\begin{aligned}
|\ell_i'^{(t)} - \ell_*'^{(t)}| &= |\ell''(z) \cdot y_i \cdot [f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i) - f(\hat{\mathbf{w}}_t^*, \gamma^{(t)}, \mathbf{x}_i)]| \\
&\leq |\ell''(z)| \cdot |f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i) - f(\hat{\mathbf{w}}_t^*, \gamma^{(t)}, \mathbf{x}_i)| \\
&\leq \max\{|\ell_1'^{(t)}|, \ldots, |\ell_n'^{(t)}|, \exp(-\gamma^{(t)})\} \cdot |f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i) - f(\hat{\mathbf{w}}_t^*, \gamma^{(t)}, \mathbf{x}_i)|. \qquad \text{(C.18)}
\end{aligned}
$$

where the first inequality follows by the property of cross-entropy loss that $0 \leq \ell''(z) \leq -\ell'(z)$, and the second inequality follows from the fact that z is between $y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)$ and $y_i \cdot f(\hat{\mathbf{w}}_t^*, \gamma^{(t)}, \mathbf{x}_i) = \gamma^{(t)}$. Moreover, by (C.13) we have $\hat{\mathbf{w}}_t^* \in \mathcal{B}^{(t)}$ and $\|\mathbf{w}\|_2 \geq \|\mathbf{w}^{(t)}\|_2/2$ for all $\mathbf{w} \in \mathcal{B}^{(t)} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w} - \mathbf{w}^{(t)}\|_2 \leq \|\mathbf{w}^{(t)}\|_2/2\}$. For all $\mathbf{w} \in \mathcal{B}^{(t)}$, we have

$$
\begin{aligned}
\|\nabla_{\mathbf{w}} f(\mathbf{w}, \gamma^{(t)}, \mathbf{x}_i)\|_2 &= \bigg\| \frac{\gamma^{(t)}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}} \cdot \bigg(\mathbf{I} - \frac{\mathbf{\Sigma}\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|_{\mathbf{\Sigma}}^2} \bigg) \mathbf{x}_i \bigg\|_2 \\
&\leq \frac{\gamma^{(t)}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}} \cdot \|\mathbf{x}_i\|_2 \cdot \bigg(1 + \bigg\| \frac{\mathbf{\Sigma}\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|_{\mathbf{\Sigma}}^2} \bigg\|_2 \bigg) \\
&= \frac{\gamma^{(t)}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}} \cdot \|\mathbf{x}_i\|_2 \cdot \bigg(1 + \frac{\|\mathbf{\Sigma}\mathbf{w}\|_2 \cdot \|\mathbf{w}\|_2}{\|\mathbf{w}\|_{\mathbf{\Sigma}}^2} \bigg) \\
&\leq \frac{2\gamma^{(t)}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}} \cdot \|\mathbf{x}_i\|_2 \cdot \frac{\|\mathbf{\Sigma}\mathbf{w}\|_2 \cdot \|\mathbf{w}\|_2}{\|\mathbf{w}\|_{\mathbf{\Sigma}}^2},
\end{aligned}
$$

where the last inequality follows by the fact that $\|\mathbf{w}\|_{\mathbf{\Sigma}}^2 = \langle \mathbf{w}, \mathbf{\Sigma}\mathbf{w}\rangle \le \|\mathbf{\Sigma}\mathbf{w}\|_2 \cdot \|\mathbf{w}\|_2$. Further plugging in the definition of $\mathbf{\Sigma}$ gives

$$
\begin{aligned}
\|\nabla_{\mathbf{w}} f(\mathbf{w}, \gamma^{(t)}, \mathbf{x}_i)\|_2 &\le \frac{2\gamma^{(t)}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}^3} \cdot \|\mathbf{x}_i\|_2 \cdot \|\mathbf{w}\|_2 \cdot \left\| \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j \cdot \langle \mathbf{w}, \mathbf{x}_j \rangle \right\|_2 \\
&\le \frac{2\gamma^{(t)}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}^3} \cdot \max_i \|\mathbf{x}_i\|_2^2 \cdot \|\mathbf{w}\|_2 \cdot \frac{1}{n} \sum_{i=1}^n |\langle \mathbf{w}, \mathbf{x}_i \rangle| \\
&\le \frac{2\gamma^{(t)}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}^3} \cdot \max_i \|\mathbf{x}_i\|_2^2 \cdot \|\mathbf{w}\|_2 \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n |\langle \mathbf{w}, \mathbf{x}_i \rangle|^2} \\
&= \frac{2\gamma^{(t)}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}^2} \cdot \max_i \|\mathbf{x}_i\|_2^2 \cdot \|\mathbf{w}\|_2 \\
&\le \frac{2\gamma^{(t)}}{\lambda_{\min} \cdot \|\mathbf{w}\|_2^2} \cdot \max_i \|\mathbf{x}_i\|_2^2 \cdot \|\mathbf{w}\|_2 \\
&\le \frac{4\gamma^{(t)}}{\lambda_{\min} \cdot \|\mathbf{w}^{(t)}\|_2} \cdot \max_i \|\mathbf{x}_i\|_2^2,
\end{aligned}
$$

where the third inequality follows by Jensen's inequality, and the last inequality follows by $\|\mathbf{w}\|_2 \ge \|\mathbf{w}^{(t)}\|_2/2$ for all $\mathbf{w} \in \mathcal{B}^{(t)}$. Therefore, $g(\mathbf{w}) := f(\mathbf{w}, \gamma^{(t)}, \mathbf{x}_i)$ is $(4\lambda_{\min}^{-1} \cdot \|\mathbf{w}^{(t)}\|_2^{-1} \cdot \gamma^{(t)} \cdot \max_i \|\mathbf{x}_i\|_2^2)$-Lipschitz continuous over $\mathcal{B}^{(t)}$, and

$$
|f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i) - f(\hat{\mathbf{w}}_t^*, \gamma^{(t)}, \mathbf{x}_i)| \le \frac{4\gamma^{(t)}}{\lambda_{\min} \cdot \|\mathbf{w}^{(t)}\|_2} \cdot \max_i \|\mathbf{x}_i\|_2^2 \cdot \|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2. \tag{C.19}
$$

Plugging (C.19) into (C.18) gives

$$
|\ell_i'^{(t)} - \ell_*'^{(t)}| \le \max\{|\ell_1'^{(t)}|, \ldots, |\ell_n'^{(t)}|, \exp(-\gamma^{(t)})\} \cdot \frac{4\gamma^{(t)}}{\lambda_{\min} \cdot \|\mathbf{w}^{(t)}\|_2} \cdot \max_i \|\mathbf{x}_i\|_2^2 \cdot \|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2
$$

for all $i \in [n]$ and $t \ge 0$. Then by the definition of I_2, we have

$$
\begin{aligned}
I_2 &= \frac{\gamma^{(t)}}{n \cdot \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \sum_{i=1}^n |\ell_i'^{(t)} - \ell_*'^{(t)}| \cdot \|\mathbf{x}_i\|_2 \\
&\le \frac{\gamma^{(t)}}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \cdot \max_i \|\mathbf{x}_i\|_2^3 \cdot \max\{|\ell_1'^{(t)}|, \ldots, |\ell_n'^{(t)}|, \exp(-\gamma^{(t)})\} \cdot \frac{4\gamma^{(t)}}{\lambda_{\min} \cdot \|\mathbf{w}^{(t)}\|_2} \cdot \|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2 \\
&\le \frac{4\gamma^{(t)2}}{\lambda_{\min}^{3/2} \cdot \|\mathbf{w}^{(t)}\|_2^2} \cdot \max_i \|\mathbf{x}_i\|_2^3 \cdot \max\{|\ell_1'^{(t)}|, \ldots, |\ell_n'^{(t)}|, \exp(-\gamma^{(t)})\} \cdot \|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2 \\
&\le \frac{4\gamma^{(t)2}}{\lambda_{\min}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \max_i \|\mathbf{x}_i\|_2^3 \cdot \max\{|\ell_1'^{(t)}|, \ldots, |\ell_n'^{(t)}|, \exp(-\gamma^{(t)})\} \cdot \|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2
\end{aligned} \tag{C.20}
$$

Plugging (C.14), (C.17) and (C.20) into (C.12) gives

$$\begin{aligned}\|\mathbf{w}^{(t+1)}\|_2^2 \leq \|\mathbf{w}^{(t)}\|_2^2 &+ 2\eta^2 \cdot \left[\frac{4\gamma^{(t)} \cdot \exp(-\gamma^{(t)})}{\lambda_{\min}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \max_i \|\mathbf{x}_i\|_2^3 \cdot \|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2\right]^2 \\ &+ 2\eta^2 \cdot \left[\frac{4\gamma^{(t)2}}{\lambda_{\min}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \max_i \|\mathbf{x}_i\|_2^3 \cdot \max\{|\ell_1'^{(t)}|, \ldots, |\ell_n'^{(t)}|, \exp(-\gamma^{(t)})\} \cdot \|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2\right]^2 \\ &\leq \|\mathbf{w}^{(t)}\|_2^2 + \eta^2 G \cdot \max\{\gamma^{(t)2}, \gamma^{(t)4}\} \cdot \max\{|\ell_1'^{(t)}|^2, \ldots, |\ell_n'^{(t)}|^2, \exp(-2\gamma^{(t)})\} \cdot \|\mathbf{w}^{(t)} - \hat{\mathbf{w}}_t^*\|_2^2,\end{aligned}$$

where $G = 64\lambda_{\min}^{-3} \cdot \max_i \|\mathbf{x}_i\|_2^6 \cdot \|\mathbf{w}^{(0)}\|_2^{-4}$. This finishes the proof. ■

C.5. Proof of Lemma 4.5

In preparation of the proof of Lemma 4.5, we first present the following lemma.

Lemma C.2 *Suppose that a sequence $a^{(t)}$, $t \geq 0$ follows the iterative formula*

$$a^{(t+1)} = a^{(t)} + c \cdot \exp(-a^{(t)})$$

for some $c > 0$. Then it holds that

$$\log(c \cdot t + \exp(a^{(0)})) \leq a^{(t)} \leq c\exp(-a^{(0)}) + \log(c \cdot t + \exp(a^{(0)}))$$

for all $t \geq 0$.

The proof of Lemma 4.5 is given as follows.

Proof [Proof of Lemma 4.5] Set $T_0 = 1 + 400\eta^{-1}\epsilon^{-2} \cdot \lambda_{\max}^{3/2} \cdot \lambda_{\min}^{-1} \cdot \|\mathbf{w}^{(0)}\|_2^2 \cdot \|\mathbf{w}^*\|_2$. By gradient descent update rule, we have

$$\begin{aligned}|\gamma^{(t+1)} - \gamma^{(t)}| &= \left|\frac{\eta}{n}\sum_{i=1}^n \ell_l'^{(t)} \cdot \frac{\langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle}{\sqrt{n^{-1} \cdot \sum_{j=1}^n \langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle^2}}\right| \\ &\leq \eta \cdot \frac{n^{-1} \cdot \sum_{i=1}^n |\langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle|}{\sqrt{n^{-1} \cdot \sum_{j=1}^n \langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle^2}} \\ &\leq \eta,\end{aligned}$$

where the first inequality follows by the fact that $|\ell'(z)| < 1$ for all $z \in \mathbb{R}$, and the second inequality follows by Jensen's inequality. Therefore, we have

$$|\gamma^{(t)} - 1| = |\gamma^{(t)} - \gamma^{(0)}| \leq \eta + 400\epsilon^{-2} \cdot \lambda_{\max}^{3/2} \cdot \lambda_{\min}^{-1} \cdot \|\mathbf{w}^{(0)}\|_2^2 \cdot \|\mathbf{w}^*\|_2.$$

Further plugging in the definition of ϵ gives

$$\begin{aligned}|\gamma^{(t)} - 1| &\leq \eta + \frac{6400 \max_i \|\mathbf{z}_i\|_2 \cdot \|\mathbf{w}^*\|_2^2 \cdot \lambda_{\max}^{3/2} \cdot \lambda_{\min}^{-1} \cdot \|\mathbf{w}^{(0)}\|_2^2}{\min\left\{1/3, \|\mathbf{w}^{(0)}\|_2^{-1} \cdot \lambda_{\min}^{1/2}/(40\lambda_{\max}^{3/4})\right\}} \\ &= \eta + 19200 \max_i \|\mathbf{z}_i\|_2 \cdot \|\mathbf{w}^*\|_2^2 \cdot \lambda_{\max}^{3/2} \cdot \lambda_{\min}^{-1} \cdot \|\mathbf{w}^{(0)}\|_2^2 \\ &\leq 1/2 \qquad \text{(C.21)}\end{aligned}$$

for all $t \in [T_0]$, where the first equality follows by the assumption that $\|\mathbf{w}^{(0)}\|_2 \leq \lambda_{\min}^{1/2}/(20\lambda_{\max}^{3/4})$, and the second inequality follows by the assumption that $\eta \leq 1/4$ and $\|\mathbf{w}^{(0)}\|_2 \leq (\max_i \|\mathbf{z}_i\|_2)^{-1/2} \cdot \|\mathbf{w}^*\|_2^{-1} \cdot \lambda_{\max}^{-3/4} \cdot \lambda_{\min}^{1/2}/280$.

By Lemma 4.4, we have

$$\|\mathbf{w}^{(t)}\|_2^2 \leq \|\mathbf{w}^{(t+1)}\|_2^2 \leq \|\mathbf{w}^{(t)}\|_2^2 + 4\eta^2 \cdot \frac{\gamma^{(t)2} \cdot \max_i \|\mathbf{x}_i\|_2^3}{\lambda_{\min}^2 \cdot \|\mathbf{w}^{(t)}\|_2^2}.$$

By the monotonicity of $\|\mathbf{w}^{(t)}\|_2$ and the result in (C.21) that $\gamma^{(t)} \leq 3/2$ for all $t \in [T_0]$, we have

$$\|\mathbf{w}^{(t+1)}\|_2^2 \leq \|\mathbf{w}^{(t)}\|_2^2 + 9\eta^2 \cdot \frac{\max_i \|\mathbf{x}_i\|_2^3}{\lambda_{\min}^2 \cdot \|\mathbf{w}^{(0)}\|_2^2}$$

for all $t \in [T_0]$. Taking a telescoping sum then gives

$$\|\mathbf{w}^{(t)}\|_2^2 \leq \|\mathbf{w}^{(0)}\|_2^2 + \frac{9\eta^2 t \cdot \max_i \|\mathbf{x}_i\|_2^3}{\lambda_{\min}^2 \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot t \leq \|\mathbf{w}^{(0)}\|_2^2 + \frac{9\eta^2 T_0 \cdot \max_i \|\mathbf{x}_i\|_2^3}{\lambda_{\min}^2 \cdot \|\mathbf{w}^{(0)}\|_2^2}$$

for all $t \in [T_0]$. Plugging in the definition of T_0 gives

$$\begin{aligned}
\|\mathbf{w}^{(t)}\|_2^2 &\leq \|\mathbf{w}^{(0)}\|_2^2 + \frac{9\eta^2 \cdot \max_i \|\mathbf{x}_i\|_2^3}{\lambda_{\min}^2 \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot (1 + 400\eta^{-1}\epsilon^{-2} \cdot \lambda_{\max}^{3/2} \cdot \lambda_{\min}^{-1} \cdot \|\mathbf{w}^{(0)}\|_2^2 \cdot \|\mathbf{w}^*\|_2) \\
&= \|\mathbf{w}^{(0)}\|_2^2 + \frac{9\eta^2 \cdot \max_i \|\mathbf{x}_i\|_2^3}{\lambda_{\min}^2 \cdot \|\mathbf{w}^{(0)}\|_2^2} + \frac{3600\eta \cdot \max_i \|\mathbf{x}_i\|_2^3}{\lambda_{\min}^3} \cdot \epsilon^{-2} \cdot \lambda_{\max}^{3/2} \cdot \|\mathbf{w}^*\|_2) \\
&\leq (1 + \epsilon^2/4) \cdot \|\mathbf{w}^{(0)}\|_2^2
\end{aligned}$$

for all $t \in [T_0]$, where the last inequality follows by the assumption that $\eta \leq \min\{\epsilon \cdot \|\mathbf{w}^{(0)}\|_2^2 \cdot \lambda_{\min} \cdot (\max_i \|\mathbf{x}_i\|_2)^{-3/2}/9, \epsilon^4 \cdot \lambda_{\min}^3 \cdot \lambda_{\max}^{-3/2} \cdot \|\mathbf{w}^*\|_2^{-1} \cdot (\max_i \|\mathbf{x}_i\|_2)^{-3}/28800\}$. Therefore

$$\|\mathbf{w}^{(t)}\|_2 \leq \sqrt{1 + \epsilon^2/4} \cdot \|\mathbf{w}^{(0)}\|_2 \leq (1 + \epsilon/2) \cdot \|\mathbf{w}^{(0)}\|_2 \tag{C.22}$$

for all $t \in [T_0]$.

By Lemma 4.2, we have

$$\langle \mathbf{w}^{(t+1)}, \mathbf{w}^* \rangle \geq \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle + \frac{\gamma^{(t)}\eta}{16\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \cdot \exp(-\gamma^{(t)}) \cdot \|\mathbf{w} - \langle \mathbf{w}^*, \mathbf{w} \rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_{\mathbf{\Sigma}}^2. \tag{C.23}$$

By the result that $\gamma^{(t)} \leq 3/2$ for all $t \in T_0$, we have $\exp(-\gamma^{(t)}) \geq \exp(-\gamma^{(3/2)}) \geq 1/5$ for all $t \in [T_0]$. Therefore by (C.23), we have

$$\begin{aligned}
\langle \mathbf{w}^{(t+1)}, \mathbf{w}^* \rangle &\geq \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle + \frac{\gamma^{(t)}\eta}{80\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \cdot \|\mathbf{w} - \langle \mathbf{w}^*, \mathbf{w} \rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_{\mathbf{\Sigma}}^2 \\
&\geq \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle + \frac{\gamma^{(t)}\eta}{80\|\mathbf{w}^{(0)}\|_{\mathbf{\Sigma}}^3} \cdot \|\mathbf{w} - \langle \mathbf{w}^*, \mathbf{w} \rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_{\mathbf{\Sigma}}^2 \\
&\geq \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle + \frac{\eta}{160\|\mathbf{w}^{(0)}\|_{\mathbf{\Sigma}}^3} \cdot \|\mathbf{w} - \langle \mathbf{w}^*, \mathbf{w} \rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_{\mathbf{\Sigma}}^2
\end{aligned}$$

for all $t \in [T_0]$, where the second inequality follows by Lemma 4.4, and the third inequality follows by the proved result that $\gamma^{(t)} \geq 1/2$ for all $t \in [T_0]$. Telescoping over $t = 0, \ldots, T_0 - 1$ then gives

$$\begin{aligned}\min_{t\in[T_0-1]} \|\mathbf{w} - \langle \mathbf{w}^*, \mathbf{w}\rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_{\mathbf{\Sigma}}^2 &\leq \frac{1}{T_0 - 1} \sum_{t=0}^{T_0-1} \|\mathbf{w} - \langle \mathbf{w}^*, \mathbf{w}\rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_{\mathbf{\Sigma}}^2 \\ &\leq \frac{160\|\mathbf{w}^{(0)}\|_{\mathbf{\Sigma}}^3}{(T_0 - 1)\eta} \cdot (\langle \mathbf{w}^{(T_0)}, \mathbf{w}^*\rangle - \langle \mathbf{w}^{(0)}, \mathbf{w}^*\rangle) \\ &\leq \frac{160\|\mathbf{w}^{(0)}\|_{\mathbf{\Sigma}}^3}{(T_0 - 1)\eta} \cdot (\|\mathbf{w}^{(T_0)}\|_2 + \|\mathbf{w}^{(0)}\|_2) \cdot \|\mathbf{w}^*\|_2.\end{aligned}$$

By the proved result that $\|\mathbf{w}^{(T_0)}\|_2 \leq (1 + \epsilon/2) \cdot \|\mathbf{w}^{(0)}\|_2 \leq 1.5\|\mathbf{w}^{(0)}\|_2$ (by definition, $\epsilon \leq \max_i \langle \mathbf{w}^*, \mathbf{z}_i\rangle^{-1}/48 = 1/48$), we then obtain

$$\begin{aligned}\min_{t\in[T_0-1]} \|\mathbf{w} - \langle \mathbf{w}^*, \mathbf{w}\rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_{\mathbf{\Sigma}}^2 &\leq \frac{400\|\mathbf{w}^{(0)}\|_{\mathbf{\Sigma}}^3}{(T_0 - 1)\eta} \cdot \|\mathbf{w}^{(0)}\|_2 \cdot \|\mathbf{w}^*\|_2 \\ &\leq \frac{400\lambda_{\max}^{3/2}}{(T_0 - 1)\eta} \cdot \|\mathbf{w}^{(0)}\|_2^4 \cdot \|\mathbf{w}^*\|_2 \\ &\leq \epsilon^2 \cdot \lambda_{\min} \cdot \|\mathbf{w}^{(0)}\|_2^2,\end{aligned}$$

where the last inequality follows by the definition that $T_0 = 1 + 400\eta^{-1}\epsilon^{-2} \cdot \lambda_{\max}^{3/2} \cdot \lambda_{\min}^{-1} \cdot \|\mathbf{w}^{(0)}\|_2^2 \cdot \|\mathbf{w}^*\|_2$. Therefore, we see that there exists $t_0 \in [T_0 - 1]$ such that

$$\begin{aligned}\left\|\mathbf{w}^{(t_0)} - \langle \mathbf{w}^*, \mathbf{w}^{(t_0)}\rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*\right\|_2 &\leq \|\mathbf{w} - \langle \mathbf{w}^*, \mathbf{w}\rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_2 \\ &\leq \lambda_{\min}^{-1/2} \cdot \|\mathbf{w} - \langle \mathbf{w}^*, \mathbf{w}\rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_{\mathbf{\Sigma}} \\ &\leq \epsilon \cdot \|\mathbf{w}^{(0)}\|_2,\end{aligned}$$

where the first inequality follows by the fact that $\langle \mathbf{w}^*, \mathbf{w}^{(t_0)}\rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*$ is the ℓ_2-projection of $\mathbf{w}^{(t_0)}$ on $\mathrm{span}\{\mathbf{w}^*\}$. Together with (C.21) and (C.22), we conclude that there exists $t_0 \in [T_0 - 1]$ such that all three results of Lemma 4.5 hold. ■

C.6. Proof of Lemma 4.6

Proof [Proof of Lemma 4.6] We prove the first five results together by induction, and then prove the sixth result. Clearly, all the results hold at $t = t_0$ by the assumptions. Now suppose that there exists $t_1 \geq t_0$ such that the results hold for $t = t_0, \ldots, t_1$, i.e., for $t = t_0, \ldots, t_1$ it holds that

(i) $\left\|\mathbf{w}^{(t_0)} - \langle \mathbf{w}^*, \mathbf{w}^{(t_0)}\rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*\right\|_2, \ldots, \left\|\mathbf{w}^{(t)} - \langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*\right\|_2$ is a decreasing sequence.

(ii) $\|\mathbf{w}^{(0)}\|_2 \leq \|\mathbf{w}^{(t)}\|_2 \leq (1 + \epsilon) \cdot \|\mathbf{w}^{(0)}\|_2$.

(iii) $\gamma^{(t)}$ has the following upper and lower bounds:

$$\begin{aligned}\gamma^{(t)} &\leq \log[8\eta \cdot (t - t_0) + 2\exp(\gamma^{(t_0)})], \\ \gamma^{(t)} &\geq \log[(\eta/8) \cdot (t - t_0) + \exp(\gamma^{(t_0)})].\end{aligned}$$

(iv) It holds that

$$\begin{aligned}
&\big\|\mathbf{w}^{(t)} - \langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*\big\|_2 \\
&\qquad \le \epsilon \cdot \|\mathbf{w}^{(0)}\|_2 \cdot \exp\bigg[-\frac{\lambda_{\min}}{1024\lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \log^2((8/9)\eta \cdot (t-t_0)+1)\bigg], \\
&n^{-2} \sum_{i,i'=1}^{n} (\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}\rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle)^2 \\
&\qquad \le \lambda_{\max} \cdot \epsilon^2 \cdot \|\mathbf{w}^{(0)}\|_2^2 \cdot \exp\bigg[-\frac{\lambda_{\min}}{512\lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \log^2((8/9)\eta \cdot (t-t_0)+1)\bigg].
\end{aligned}$$

(v) $\max_i |\langle \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2, \mathbf{z}_i\rangle - \|\mathbf{w}^*\|_2^{-1}| \cdot \gamma^{(t)} \le \|\mathbf{w}^*\|_2^{-1}/4$.

Then we aim to show that the above conclusions also hold at iteration $t_1 + 1$.

Preliminary results based on the induction hypotheses. By definition, it is easy to see that

$$\epsilon \le \|\mathbf{w}^*\|_2^{-1} \cdot (\max_i \|\mathbf{z}_i\|_2)^{-1}/24 \le (\min_i |\langle \mathbf{w}^*, \mathbf{z}_i\rangle|)^{-1}/24 = 1/24 < 1.$$

By induction hypothesis (i), we have

$$\big\|\mathbf{w}^{(t)} - \langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*\big\|_2 \le \big\|\mathbf{w}^{(t_0)} - \langle \mathbf{w}^*, \mathbf{w}^{(t_0)}\rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*\big\|_2 \le \epsilon \cdot \|\mathbf{w}^{(0)}\|_2. \tag{C.24}$$

Taking the square of both sides and dividing by $\|\mathbf{w}^{(t)}\|_2^2$ gives

$$1 - \langle \mathbf{w}^*/\|\mathbf{w}^*\|_2, \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2\rangle^2 \le \epsilon^2 \cdot \|\mathbf{w}^{(0)}\|_2^2/\|\mathbf{w}^{(t)}\|_2^2 \le \epsilon^2,$$

where the last inequality follows by Lemma 4.4 on the monotonicity of $\|\mathbf{w}^{(t)}\|_2$. Therefore, we have

$$(1-\epsilon) \le \sqrt{1-\epsilon^2} \le \langle \mathbf{w}^*/\|\mathbf{w}^*\|_2, \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2\rangle \le 1. \tag{C.25}$$

Moreover, by (C.24), we have

$$|\langle \mathbf{w}^{(t)}, \mathbf{z}_i\rangle - \langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \langle \mathbf{w}^*, \mathbf{z}_i\rangle| \le \max_i \|\mathbf{z}_i\|_2 \cdot \|\mathbf{w}^{(0)}\|_2 \cdot \epsilon$$

for all $i \in [n]$. Dividing by $\|\mathbf{w}^{(t)}\|_2$ on both sides above gives

$$\begin{aligned}
&|\langle \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2, \mathbf{z}_i\rangle - \langle \mathbf{w}^*/\|\mathbf{w}^*\|_2, \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2\rangle \cdot \langle \mathbf{w}^*/\|\mathbf{w}^*\|_2, \mathbf{z}_i\rangle| \\
&\qquad \le \max_i \|\mathbf{z}_i\|_2 \cdot \|\mathbf{w}^{(0)}\|_2/\|\mathbf{w}^{(t)}\|_2 \cdot \epsilon \le \max_i \|\mathbf{z}_i\|_2 \cdot \epsilon.
\end{aligned}$$

Recall that $\mathbf{w}^*$ is chosen such that $\langle \mathbf{w}^*, \mathbf{z}_i \rangle = 1$ for all $i \in [n]$. Therefore, rearranging terms and applying (C.25) then gives

$$\begin{aligned}
\langle \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2, \mathbf{z}_i \rangle &\geq \langle \mathbf{w}^*/\|\mathbf{w}^*\|_2, \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2 \rangle \cdot \langle \mathbf{w}^*/\|\mathbf{w}^*\|_2, \mathbf{z}_i \rangle - \max_i \|\mathbf{z}_i\|_2 \cdot \epsilon \\
&\geq \|\mathbf{w}^*\|_2^{-1} \cdot \langle \mathbf{w}^*, \mathbf{z}_i \rangle - 2 \max_i \|\mathbf{z}_i\|_2 \cdot \epsilon, \\
&= \|\mathbf{w}^*\|_2^{-1} - 2 \max_i \|\mathbf{z}_i\|_2 \cdot \|\mathbf{w}^{(0)}\|_2 \cdot \epsilon \\
\langle \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2, \mathbf{z}_i \rangle &\leq \langle \mathbf{w}^*/\|\mathbf{w}^*\|_2, \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2 \rangle \cdot \langle \mathbf{w}^*/\|\mathbf{w}^*\|_2, \mathbf{z}_i \rangle + \max_i \|\mathbf{z}_i\|_2 \cdot \epsilon \\
&\leq \|\mathbf{w}^*\|_2^{-1} \cdot \langle \mathbf{w}^*, \mathbf{z}_i \rangle + \max_i \|\mathbf{z}_i\|_2 \cdot \epsilon \\
&= \|\mathbf{w}^*\|_2^{-1} + \max_i \|\mathbf{z}_i\|_2 \cdot \epsilon.
\end{aligned}$$

Therefore, we have

$$|\langle \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2, \mathbf{z}_i \rangle - \|\mathbf{w}^*\|_2^{-1}| \leq 2 \max_i \|\mathbf{z}_i\|_2 \cdot \epsilon \leq \|\mathbf{w}^*\|_2^{-1}/4 \tag{C.26}$$

for all $i \in [n]$ and $t = t_0, \ldots, t_1$, where the second inequality follows by the definition of ϵ. Now denote $\alpha = \|\mathbf{w}^*\|_2^{-1}$ and $c^{(t)} = \max_i |\langle \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2, \mathbf{z}_i \rangle - \alpha|$. Then we have

$$\begin{aligned}
y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i) &= \frac{\gamma^{(t)} \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle}{\sqrt{n^{-1} \cdot \sum_{i'=1}^n \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle^2}} = \frac{\gamma^{(t)} \cdot \langle \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2, \mathbf{z}_i \rangle}{\sqrt{n^{-1} \cdot \sum_{i'=1}^n \langle \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2, \mathbf{z}_{i'} \rangle^2}} \\
&\leq \gamma^{(t)} \cdot \frac{\alpha + c^{(t)}}{\alpha - c^{(t)}} \leq \gamma^{(t)} \cdot (1 + 3c^{(t)}/\alpha) \leq \gamma^{(t)} + 3/4,
\end{aligned} \tag{C.27}$$

where the second inequality follows by (C.26) that $\max_i |\langle \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2, \mathbf{z}_i \rangle - \alpha| \leq \alpha/4$, and the last inequality follows by induction hypothesis (v). Similarly,

$$\begin{aligned}
y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i) &= \frac{\gamma^{(t)} \cdot \langle \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2, \mathbf{z}_i \rangle}{\sqrt{n^{-2} \cdot \sum_{i'=1}^n \langle \mathbf{w}^{(t)}/\|\mathbf{w}^{(t)}\|_2, \mathbf{z}_{i'} \rangle^2}} \\
&\geq \gamma^{(t)} \cdot \frac{\alpha - c^{(\tau)}}{\alpha + c^{(\tau)}} \geq \gamma^{(t)} \cdot (1 - c^{(\tau)}/\alpha) \geq \gamma^{(t)} - 1/4.
\end{aligned} \tag{C.28}$$

Note that

$$-\ell'(y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)) = \frac{1}{1 + \exp[y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)]}$$

for all $i \in [n]$ and all $t \in [t_0, t_1]$. Therefore by (C.27) and (C.28), we have

$$\exp(-\gamma^{(t)} - 3/4)/2 \leq -\ell'(y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)) \leq \exp(-\gamma^{(t)} + 3/4) \tag{C.29}$$

for all $i \in [n]$ and all $t \in [t_0, t_1]$.

Proof of induction hypothesis (i) at iteration $t_1 + 1$. By Lemma 4.2, for any $t = t_0, \ldots, t_1$, we have

$$
\begin{aligned}
&-\frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2} \cdot \langle \mathbf{w}^{(t+1)}, \mathbf{w}^*\rangle \\
&\le -\frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2} \cdot \langle \mathbf{w}^{(t)}, \mathbf{w}^*\rangle - \frac{\eta \cdot \gamma^{(t)} \cdot \langle \mathbf{w}^{(t)}, \mathbf{w}^*\rangle}{2\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3 \cdot \|\mathbf{w}^*\|_2} \cdot \min_i |\ell_i'^{(t)}| \cdot \|\mathbf{w}^{(t)} - \langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_{\mathbf{\Sigma}}^2 \\
&\le -\frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2} \cdot \langle \mathbf{w}^{(t)}, \mathbf{w}^*\rangle - \frac{\eta \cdot \gamma^{(t)} \cdot \lambda_{\min} \cdot \langle \mathbf{w}^{(t)}, \mathbf{w}^*\rangle}{2\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3 \cdot \|\mathbf{w}^*\|_2} \cdot \min_i |\ell_i'^{(t)}| \cdot \|\mathbf{w}^{(t)} - \langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_2^2 \\
&\le -\frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2} \cdot \langle \mathbf{w}^{(t)}, \mathbf{w}^*\rangle - \frac{\eta \cdot \gamma^{(t)} \cdot \lambda_{\min} \cdot \langle \mathbf{w}^{(t)}, \mathbf{w}^*\rangle}{2\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3 \cdot \|\mathbf{w}^*\|_2} \cdot \min_i |\ell_i'^{(t)}| \cdot \left\|\mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\right\|_2^2 \\
&\le -\frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2} \cdot \langle \mathbf{w}^{(t)}, \mathbf{w}^*\rangle - \frac{\eta \cdot \gamma^{(t)} \cdot \lambda_{\min} \cdot \langle \mathbf{w}^{(t)}, \mathbf{w}^*\rangle}{2\lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(t)}\|_2^3 \cdot \|\mathbf{w}^*\|_2} \cdot \min_i |\ell_i'^{(t)}| \cdot \left\|\mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\right\|_2^2 \\
&\le -\frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2} \cdot \langle \mathbf{w}^{(t)}, \mathbf{w}^*\rangle - \frac{\eta \cdot \gamma^{(t)} \cdot \lambda_{\min}}{8\lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(t)}\|_2^2} \cdot \exp(-\gamma^{(t)}) \cdot \left\|\mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\right\|_2^2 \\
&\le -\frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2} \cdot \langle \mathbf{w}^{(t)}, \mathbf{w}^*\rangle - \frac{\eta \cdot \gamma^{(t)} \cdot \lambda_{\min}}{16\lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \exp(-\gamma^{(t)}) \cdot \left\|\mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\right\|_2^2,
\end{aligned}
$$

where the third inequality follows by the fact that $\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*$ is the projection of $\mathbf{w}^{(t)}$ on $\mathrm{span}\{\mathbf{w}^*\}$ and $\|\mathbf{w}^{(t)} - \langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle \cdot \|\mathbf{w}^*\|_2^{-2} \cdot \mathbf{w}^*\|_2^2 \le \|\mathbf{w}^{(t)} - c \cdot \mathbf{w}^*\|_2^2$ for all $c \in \mathbb{R}$, the fifth inequality follows by (C.25) and (C.29), and the last inequality follows by (C.25) and $(1+\epsilon) \le \sqrt{2}$. Adding $\|\mathbf{w}^{(t)}\|_2^2 + \langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle^2 \cdot \|\mathbf{w}^*\|_2^{-2}/4$ to both sides above gives

$$
\begin{aligned}
&\left\|\mathbf{w}^{(t+1)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\right\|_2^2 + \|\mathbf{w}^{(t)}\|_2^2 - \|\mathbf{w}^{(t+1)}\|_2^2 \\
&\le \left\|\mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\right\|_2^2 - \frac{\eta \cdot \gamma^{(t)} \cdot \lambda_{\min}}{16\lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \exp(-\gamma^{(t)}) \cdot \left\|\mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\right\|_2^2.
\end{aligned} \tag{C.30}
$$

Now by Lemma 4.4, we have

$$
\begin{aligned}
&\|\mathbf{w}^{(t+1)}\|_2^2 - \|\mathbf{w}^{(t)}\|_2^2 \\
&\qquad \le \eta^2 G \cdot \max\{\gamma^{(t)2}, \gamma^{(t)4}\} \cdot \max\{|\ell_1'^{(t)}|^2, \ldots, |\ell_n'^{(t)}|^2, \exp(-2\gamma^{(t)})\} \cdot \left\|\mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\right\|_2^2,
\end{aligned}
$$

where $G = 64\lambda_{\min}^{-3} \cdot \max_i \|\mathbf{x}_i\|_2^6 \cdot \|\mathbf{w}^{(0)}\|_2^{-4}$. Then according to (C.29), we have

$$
\|\mathbf{w}^{(t+1)}\|_2^2 \le \|\mathbf{w}^{(t)}\|_2^2 + \eta^2 G \cdot \max\{\gamma^{(t)2}, \gamma^{(t)4}\} \cdot \exp(-2\gamma^{(t)}) \cdot \exp(3/2) \cdot \left\|\mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\right\|_2^2,
$$

Note that we have $0 < \max\{z, z^3\} \cdot \exp(-z) < 2$ for all $z > 0$. Therefore we have

$$\begin{aligned}\|\mathbf{w}^{(t+1)}\|_2^2 &\leq \|\mathbf{w}^{(t)}\|_2^2 + 2\eta^2 G \cdot \exp(-\gamma^{(t)}) \cdot \exp(3/2) \cdot \bigg\|\mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\bigg\|_2^2 \\ &\leq \frac{\eta \cdot \gamma^{(t)} \cdot \lambda_{\min}}{32\lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \exp(-\gamma^{(t)}) \cdot \bigg\|\mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\bigg\|_2^2, \end{aligned} \tag{C.31}$$

where the last inequality follows by the assumption that $\eta \leq G^{-1} \cdot \exp(-3/2) \cdot \frac{\lambda_{\min}}{64\lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2}$. Moreover, note again that $\frac{\langle \mathbf{w}^*, \mathbf{w}^{(t+1)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*$ is the projection of $\mathbf{w}^{(t+1)}$ onto the subspace $\mathrm{span}\{\mathbf{w}^*\}$, which implies that

$$\bigg\|\mathbf{w}^{(t+1)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t+1)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\bigg\|_2^2 \leq \bigg\|\mathbf{w}^{(t+1)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\bigg\|_2^2. \tag{C.32}$$

Plugging (C.31) and (C.32) into (C.30) gives

$$\begin{aligned}&\bigg\|\mathbf{w}^{(t+1)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t+1)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\bigg\|_2^2 \\ &\quad\leq \bigg[1 - \frac{\eta \cdot \gamma^{(t)} \cdot \lambda_{\min}}{32\lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \exp(-\gamma^{(t)})\bigg] \cdot \bigg\|\mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\bigg\|_2^2 \end{aligned} \tag{C.33}$$

for all $t \in [t_0, t_1]$. This implies that

$$\bigg\|\mathbf{w}^{(t+1)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t+1)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\bigg\|_2^2 \leq \bigg\|\mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\bigg\|_2^2$$

for all $t \in [t_0, t_1]$, which completes the proof of induction hypothesis (i) at iteration $t_1 + 1$.

Proof of induction hypothesis (ii) at iteration $t_1 + 1$. By Lemma 4.4 and (C.29), for any $t = t_0, \ldots, t_1$, we have

$$\begin{aligned}\|\mathbf{w}^{(t+1)}\|_2^2 &\leq \|\mathbf{w}^{(t)}\|_2^2 + \eta^2 G \cdot \max\{\gamma^{(t)2}, \gamma^{(t)4}\} \cdot \exp(-2\gamma^{(t)} + 3/2) \cdot \bigg\|\mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\bigg\|_2^2 \\ &\leq \|\mathbf{w}^{(t)}\|_2^2 + 400\eta^2 G \cdot \exp(-1.5\gamma^{(t)}) \cdot \bigg\|\mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\bigg\|_2^2 \\ &\leq \|\mathbf{w}^{(t)}\|_2^2 + 400\eta^2 G \cdot \exp(-1.5\gamma^{(t)}) \cdot \bigg\|\mathbf{w}^{(t_0)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t_0)}\rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^*\bigg\|_2^2 \\ &= \|\mathbf{w}^{(t)}\|_2^2 + \frac{400\eta^2 G}{\exp(1.5\gamma^{(t)})} \cdot \bigg\|\frac{\mathbf{w}^{(t_0)}}{\|\mathbf{w}^{(t_0)}\|_2} - \bigg\langle \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}, \frac{\mathbf{w}^{(t_0)}}{\|\mathbf{w}^{(t_0)}\|_2}\bigg\rangle \cdot \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}\bigg\|_2^2 \cdot \|\mathbf{w}^{(t_0)}\|_2^2 \\ &= \|\mathbf{w}^{(t)}\|_2^2 + \frac{400\eta^2 G}{\exp(1.5\gamma^{(t)})} \cdot \bigg(\bigg\|\frac{\mathbf{w}^{(t_0)}}{\|\mathbf{w}^{(t_0)}\|_2}\bigg\|_2^2 - \bigg\langle \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}, \frac{\mathbf{w}^{(t_0)}}{\|\mathbf{w}^{(t_0)}\|_2}\bigg\rangle^2\bigg) \cdot \|\mathbf{w}^{(t_0)}\|_2^2 \\ &\leq \|\mathbf{w}^{(t)}\|_2^2 + \frac{400\eta^2 G}{\exp(1.5\gamma^{(t)})} \cdot \|\mathbf{w}^{(t_0)}\|_2^2 \\ &\leq \|\mathbf{w}^{(t)}\|_2^2 + \frac{800\eta^2 G}{\exp(1.5\gamma^{(t)})} \cdot \|\mathbf{w}^{(0)}\|_2^2, \end{aligned}$$

where $G = 64\lambda_{\min}^{-3} \cdot \max_i \|\mathbf{x}_i\|_2^6 \cdot \|\mathbf{w}^{(0)}\|_2^{-4}$ as defined in Lemma 4.4, and the second inequality follows by the fact that $0 < \max\{z^2, z^4\} \cdot \exp(-0.5z) < 80$ for all $z > 0$, and the third inequality follows by induction hypothesis (i). Therefore, by induction hypothesis (iii), we have

$$\begin{aligned}
\|\mathbf{w}^{(t_1+1)}\|_2^2 &\leq \|\mathbf{w}^{(0)}\|_2^2 + 800\eta^2 G \cdot \|\mathbf{w}^{(0)}\|_2^2 \cdot \sum_{t=t_0}^{t_1} \frac{1}{\exp(1.5\gamma^{(t)})} \\
&\leq \|\mathbf{w}^{(0)}\|_2^2 + 800\eta^2 G \cdot \|\mathbf{w}^{(0)}\|_2^2 \cdot \sum_{t=t_0}^{t_1} \frac{1}{\exp\{1.5\log[(\eta/8)\cdot(t-t_0)+\exp(\gamma^{(t_0)})]\}} \\
&= \|\mathbf{w}^{(0)}\|_2^2 + 800\eta^2 G \cdot \|\mathbf{w}^{(0)}\|_2^2 \cdot \sum_{t=t_0}^{t_1} \frac{1}{[(\eta/8)\cdot(t-t_0)+\exp(\gamma^{(t_0)})]^{1.5}}. \qquad \text{(C.34)}
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\sum_{t=t_0}^{t_1} \frac{1}{[(\eta/8)\cdot(t-t_0)+\exp(\gamma^{(t_0)})]^{1.5}} &= \sum_{t=0}^{t_1-t_0} \frac{1}{[(\eta/8)\cdot t+\exp(\gamma^{(t_0)})]^{1.5}} \\
&\leq \sum_{t=0}^{\infty} \frac{1}{[(\eta/8)\cdot t+\exp(\gamma^{(t_0)})]^{1.5}} \\
&= \frac{1}{\exp(1.5\gamma^{(t_0)})} + \sum_{t=1}^{\infty} \frac{1}{[(\eta/8)\cdot t+\exp(\gamma^{(t_0)})]^{1.5}} \\
&\leq \frac{1}{\exp(1.5\gamma^{(t_0)})} + \int_1^{\infty} \frac{1}{[(\eta/8)\cdot t+\exp(\gamma^{(t_0)})]^{1.5}} \mathrm{d}t \\
&= \frac{1}{\exp(1.5\gamma^{(t_0)})} + \frac{2}{(\eta/8)\cdot\exp(\gamma^{(t_0)}/2)} \\
&\leq 1 + 16\eta^{-1}.
\end{aligned}$$

Plugging the bound above into (C.34) gives

$$\begin{aligned}
|\mathbf{w}^{(t_1+1)}\|_2^2 &\leq \|\mathbf{w}^{(0)}\|_2^2 + 800\eta^2 G \cdot \|\mathbf{w}^{(0)}\|_2^2 \cdot (1+16\eta^{-1}) \\
&\leq \|\mathbf{w}^{(0)}\|_2^2 + 16000\eta G \cdot \|\mathbf{w}^{(0)}\|_2^2 \\
&\leq (1+\epsilon^2) \cdot \|\mathbf{w}^{(0)}\|_2^2,
\end{aligned}$$

where the second inequality follows by the definition of H_1 and the assumption that $\eta \leq 1$, and the last inequality follows by the assumption that $\eta \leq G^{-1} \cdot \epsilon^2/16000$. Therefore we have

$$|\mathbf{w}^{(t_1+1)}\|_2 \leq \sqrt{1+\epsilon^2} \cdot \|\mathbf{w}^{(0)}\|_2 \leq (1+\epsilon) \cdot \|\mathbf{w}^{(0)}\|_2,$$

which finishes the proof of induction hypothesis (ii) at iteration $t_1 + 1$.

Proof of induction hypothesis (iii) at iteration $t_1 + 1$. The gradient descent update rule for $\gamma^{(t)}$ gives

$$\begin{aligned}\gamma^{(t+1)} &= \gamma^{(t)} - \eta \cdot \frac{1}{n}\sum_{i=1}^{n} \ell'(y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)) \cdot \frac{\langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} \\ &= \gamma^{(t)} - \eta \cdot \frac{1}{n}\sum_{i=1}^{n} \ell'(y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)) \cdot \frac{\langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle}{\sqrt{\frac{1}{n}\sum_{j=1}^{n} \langle \mathbf{w}^{(t)}, \mathbf{z}_j \rangle^2}}.\end{aligned}$$

By (C.26) and (C.29), we then have

$$\begin{aligned}\gamma^{(t+1)} &\le \gamma^{(t)} - \eta \cdot \frac{1}{n}\sum_{i=1}^{n} \ell'(y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)) \cdot \frac{\alpha + \alpha/4}{\sqrt{\frac{1}{n}\sum_{j=1}^{n} (\alpha - \alpha/4)^2}} \\ &\le \gamma^{(t)} + \eta \cdot \exp(-\gamma^{(t)} + 3/4) \cdot \frac{\alpha + \alpha/4}{\sqrt{\frac{1}{n}\sum_{j=1}^{n} (\alpha - \alpha/4)^2}} \\ &\le \gamma^{(t)} + 4\eta \cdot \exp(-\gamma^{(t)}), \qquad \text{(C.35)} \\ \gamma^{(t+1)} &\ge \gamma^{(t)} - \eta \cdot \frac{1}{n}\sum_{i=1}^{n} \ell'(y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)) \cdot \frac{\alpha - \alpha/4}{\sqrt{\frac{1}{n}\sum_{j=1}^{n} (\alpha + \alpha/4)^2}} \\ &\ge \gamma^{(t)} + \eta \cdot \exp(-\gamma^{(t)} - 3/4)/2 \cdot \frac{\alpha - \alpha/4}{\sqrt{\frac{1}{n}\sum_{j=1}^{n} (\alpha + \alpha/4)^2}} \\ &\ge \gamma^{(t)} + \frac{\eta}{8} \cdot \exp(-\gamma^{(t)}) \qquad \text{(C.36)}\end{aligned}$$

for all $t = t_0, \ldots, t_1$. The comparison theorem for discrete dynamical systems then gives $\gamma^{(t_1+1)} \le \overline{\gamma}^{(t_1+1)}$, where $\overline{\gamma}^{(t)}$ is given by the iterative formula

$$\overline{\gamma}^{(t+1)} = \overline{\gamma}^{(t)} + 4\eta \cdot \exp(-\overline{\gamma}^{(t)}), \quad \overline{\gamma}^{(0)} = \gamma^{(0)}.$$

Applying Lemma C.2 then gives

$$\gamma^{(t_1+1)} \le \overline{\gamma}^{(t_1+1)} \le 4\eta \cdot \exp(-\gamma^{(t_0)}) + \log[4\eta \cdot (t_1 + 1 - t_0) + \exp(\gamma^{(t_0)})].$$

By the assumption that $\eta \le 1/8 \le \log(2)/4$, we have

$$\gamma^{(t_1+1)} \le \log[8\eta \cdot (t_1 + 1 - t_0) + 2\exp(\gamma^{(t_0)})].$$

Similarly, by Lemma C.2 and (C.36), we also have

$$\gamma^{(t_1+1)} \ge \log[(\eta/8) \cdot (t_1 + 1 - t_0) + \exp(\gamma^{(t_0)})].$$

This finishes the proof of induction hypothesis (iii) at iteration $t_1 + 1$.

Proof of induction hypothesis (iv) at iteration $t_1 + 1$. By (C.33), we have

$$\begin{aligned}
&\left\|\mathbf{w}^{(t_1+1)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t_1+1)}\rangle}{\|\mathbf{w}^*\|_2^2}\cdot \mathbf{w}^*\right\|_2^2 \\
&\quad\le \left[1 - \frac{\eta\cdot\gamma^{(t_1)}\cdot\lambda_{\min}}{32\lambda_{\max}^{3/2}\cdot\|\mathbf{w}^{(0)}\|_2^2}\cdot\exp(-\gamma^{(t_1)})\right]\cdot\left\|\mathbf{w}^{(t_1)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t_1)}\rangle}{\|\mathbf{w}^*\|_2^2}\cdot \mathbf{w}^*\right\|_2^2 \\
&\quad\le \prod_{\tau=t_0}^{t_1}\left[1 - \frac{\eta\cdot\gamma^{(\tau)}\cdot\lambda_{\min}}{32\lambda_{\max}^{3/2}\cdot\|\mathbf{w}^{(0)}\|_2^2}\cdot\exp(-\gamma^{(\tau)})\right]\cdot\left\|\mathbf{w}^{(t_0)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t_0)}\rangle}{\|\mathbf{w}^*\|_2^2}\cdot \mathbf{w}^*\right\|_2^2 \\
&\quad\le \prod_{\tau=t_0}^{t_1}\left[1 - \frac{\eta\cdot\gamma^{(\tau)}\cdot\lambda_{\min}}{32\lambda_{\max}^{3/2}\cdot\|\mathbf{w}^{(0)}\|_2^2}\cdot\exp(-\gamma^{(\tau)})\right]\cdot\epsilon^2\cdot\|\mathbf{w}^{(0)}\|_2^2
\end{aligned}$$

Then by the fact that $1 - x \le \exp(-x)$ for all $x \in \mathbb{R}$, we have

$$\begin{aligned}
\left\|\mathbf{w}^{(t_1+1)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t_1+1)}\rangle}{\|\mathbf{w}^*\|_2^2}\cdot \mathbf{w}^*\right\|_2^2 &\le \prod_{\tau=t_0}^{t_1}\exp\left[-\frac{\eta\cdot\gamma^{(\tau)}\cdot\lambda_{\min}}{64\lambda_{\max}^{3/2}\cdot\|\mathbf{w}^{(0)}\|_2^2}\cdot\exp(-\gamma^{(\tau)})\right]\cdot\epsilon\|\mathbf{w}^{(0)}\|_2 \\
&= \exp\left[-\sum_{\tau=t_0}^{t_1}\frac{\eta\cdot\gamma^{(\tau)}\cdot\lambda_{\min}}{64\lambda_{\max}^{3/2}\cdot\|\mathbf{w}^{(0)}\|_2^2}\cdot\exp(-\gamma^{(\tau)})\right]\cdot\epsilon\|\mathbf{w}^{(0)}\|_2.
\end{aligned} \tag{C.37}$$

We then study the term $\gamma^{(\tau)}\cdot\exp(-\gamma^{(\tau)})$. Note that by induction hypothesis (iii), for all $\tau = t_0, \ldots, t_1$, we have

$$\gamma^{(\tau)} \ge \log[(\eta/8)\cdot(\tau - t_0) + \exp(\gamma^{(t_0)})] \ge \gamma^{(t_0)} \ge 1/2,$$
$$\gamma^{(\tau)} \le \log[8\eta\cdot(t - t_0) + 2\exp(\gamma^{(t_0)})] \le \log[8\eta\cdot(t - t_0) + 2\exp(1.5)] \le \log[8\eta\cdot(\tau - t_0) + 9].$$

Over the interval $[1/2, \log[H_0\eta\cdot(\tau - t_0) + 9]]$, the function $g(z) = z\cdot\exp(-z)$ is strictly increasing for $1/2 \le z \le 1$ and strictly decreasing for $1 \le z \le \log[H_0\eta\cdot(\tau - t_0) + 9]$. Note that $\log[H_0\eta\cdot(\tau - t_0) + 9] > \log(9) > 2$, and $g\{\log[H_0\eta\cdot(\tau - t_0) + 9]\} < g(2) < g(1/2)$. Therefore we have

$$\begin{aligned}
\gamma^{(\tau)}\cdot\exp(-\gamma^{(\tau)}) &\ge \min_{1/2\le z\le\log[8\eta\cdot(\tau-t_0)+9]} g(z) \\
&\ge \log[8\eta\cdot(\tau - t_0) + 9]\cdot\exp\{-\log[8\eta\cdot(\tau - t_0) + 9]\} \\
&= \frac{\log[8\eta\cdot(\tau - t_0) + 9]}{8\eta\cdot(\tau - t_0) + 9}.
\end{aligned}$$

Plugging the bound above into (C.37) gives

$$\begin{aligned}
&\left\|\mathbf{w}^{(t_1+1)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t_1+1)}\rangle}{\|\mathbf{w}^*\|_2^2}\cdot \mathbf{w}^*\right\|_2^2 \\
&\quad\le \exp\left[-\frac{\eta\cdot\lambda_{\min}}{128\lambda_{\max}^{3/2}\cdot\|\mathbf{w}^{(0)}\|_2^2}\cdot\sum_{\tau=t_0}^{t_1}\frac{\log[8\eta\cdot(\tau - t_0) + 9]}{8\eta\cdot(\tau - t_0) + 9}\right]\cdot\epsilon\cdot\|\mathbf{w}^{(0)}\|_2 \\
&\quad\le \exp\left[-\frac{\eta\cdot\lambda_{\min}}{128\lambda_{\max}^{3/2}\cdot\|\mathbf{w}^{(0)}\|_2^2}\cdot\int_0^{t_1+1-t_0}\frac{\log(8\eta\cdot\tilde{\tau} + 9)}{8\eta\cdot\tilde{\tau} + 9}\mathrm{d}\tilde{\tau}\right]\cdot\epsilon\cdot\|\mathbf{w}^{(0)}\|_2,
\end{aligned}$$

where the second inequality follows by $8\eta \cdot \tilde{\tau} + 9 > e$, and $\log(z)/z$ is monotonically decreasing for $z > e$. Further calculating the integral, we obtain

$$\begin{aligned}
&\left\| \mathbf{w}^{(t_1+1)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t_1+1)} \rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^* \right\|_2 \\
&\quad \leq \exp\left[- \frac{\lambda_{\min}}{1024 \lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \log^2(Q) \Big|_{Q=9}^{8\eta \cdot (t_1+1-t_0)+9} \right] \cdot \epsilon \cdot \|\mathbf{w}^{(0)}\|_2 \\
&\quad = \exp\left[- \frac{\lambda_{\min}}{1024 \lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot [\log^2(8\eta \cdot (t_1+1-t_0)+9) - \log^2(9)] \right] \cdot \epsilon \cdot \|\mathbf{w}^{(0)}\|_2 \\
&\quad \leq \exp\left[- \frac{\lambda_{\min}}{1024 \lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot [\log(8\eta \cdot (t_1+1-t_0)+9) - \log(9)]^2 \right] \cdot \epsilon \cdot \|\mathbf{w}^{(0)}\|_2 \\
&\quad = \exp\left[- \frac{\lambda_{\min}}{1024 \lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \log^2((8/9)\eta \cdot (t_1+1-t_0)+1) \right] \cdot \epsilon \cdot \|\mathbf{w}^{(0)}\|_2, \qquad \text{(C.38)}
\end{aligned}$$

where the second inequality follows by the fact that $a^2 - b^2 \geq (a-b)^2$ for $a > b > 0$. This finishes the proof of induction hypothesis (iv) at iteration $t_1 + 1$.

Proof of induction hypothesis (v) at iteration $t_1 + 1$. By (C.38), we have

$$\left\| \mathbf{w}^{(t_1+1)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t_1+1)} \rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^* \right\|_2 \leq \mathcal{E}^{(t_1+1)} \cdot \|\mathbf{w}^{(0)}\|_2,$$

where

$$\mathcal{E}^{(t_1+1)} := \exp\left[- \frac{\lambda_{\min}}{1024 \lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \log^2((8/9)\eta \cdot (t_1+1-t_0)+1) \right] \cdot \epsilon \qquad \text{(C.39)}$$

Taking the square of both sides and dividing by $\|\mathbf{w}^{(t_1+1)}\|_2^2$ gives

$$1 - \langle \mathbf{w}^*/\|\mathbf{w}^*\|_2, \mathbf{w}^{(t_1+1)}/\|\mathbf{w}^{(t_1+1)}\|_2 \rangle^2 \leq (\mathcal{E}^{(t_1+1)})^2 \cdot \|\mathbf{w}^{(0)}\|_2^2 / \|\mathbf{w}^{(t_1+1)}\|_2^2 \leq (\mathcal{E}^{(t_1+1)})^2,$$

where the last inequality follows by Lemma 4.4 on the monotonicity of $\|\mathbf{w}^{(t)}\|_2$. Therefore, we have

$$(1 - \mathcal{E}^{(t_1+1)}) \leq \sqrt{1 - (\mathcal{E}^{(t_1+1)})^2} \leq \langle \mathbf{w}^*/\|\mathbf{w}^*\|_2, \mathbf{w}^{(t_1+1)}/\|\mathbf{w}^{(t_1+1)}\|_2 \rangle \leq 1. \qquad \text{(C.40)}$$

Moreover, by (C.38), we have

$$\left| \langle \mathbf{w}^{(t_1+1)}, \mathbf{z}_i \rangle - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t_1+1)} \rangle}{\|\mathbf{w}^*\|_2^2} \cdot \langle \mathbf{w}^*, \mathbf{z}_i \rangle \right| \leq \max_i \|\mathbf{z}_i\|_2 \cdot \mathcal{E}^{(t_1+1)} \cdot \|\mathbf{w}^{(0)}\|_2$$

for all $i \in [n]$. Dividing by $\|\mathbf{w}^{(t_1+1)}\|_2$ on both sides above gives

$$\left| \left\langle \frac{\mathbf{w}^{(t_1+1)}}{\|\mathbf{w}^{(t_1+1)}\|_2}, \mathbf{z}_i \right\rangle - \left\langle \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}, \frac{\mathbf{w}^{(t_1+1)}}{\|\mathbf{w}^{(t_1+1)}\|_2} \right\rangle \cdot \left\langle \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}, \mathbf{z}_i \right\rangle \right| \leq \max_i \|\mathbf{z}_i\|_2 \cdot \mathcal{E}^{(t_1+1)}.$$

Recall again that $\mathbf{w}^*$ is chosen such that $\langle \mathbf{w}^*, \mathbf{z}_i \rangle = 1$ for all $i \in [n]$. Therefore, rearranging terms and applying (C.40) then gives

$$\begin{aligned}
\left\langle \frac{\mathbf{w}^{(t_1+1)}}{\|\mathbf{w}^{(t_1+1)}\|_2}, \mathbf{z}_i \right\rangle &\geq \left\langle \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}, \frac{\mathbf{w}^{(t_1+1)}}{\|\mathbf{w}^{(t_1+1)}\|_2} \right\rangle \cdot \left\langle \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}, \mathbf{z}_i \right\rangle - \max_i \|\mathbf{z}_i\|_2 \cdot \mathcal{E}^{(t_1+1)} \\
&\geq \left\langle \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}, \mathbf{z}_i \right\rangle - 2 \max_i \|\mathbf{z}_i\|_2 \cdot \mathcal{E}^{(t_1+1)} \\
&= \|\mathbf{w}^*\|_2^{-1} - 2 \max_i \|\mathbf{z}_i\|_2 \cdot \mathcal{E}^{(t_1+1)}, \\
\left\langle \frac{\mathbf{w}^{(t_1+1)}}{\|\mathbf{w}^{(t_1+1)}\|_2}, \mathbf{z}_i \right\rangle &\leq \left\langle \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}, \frac{\mathbf{w}^{(t_1+1)}}{\|\mathbf{w}^{(t_1+1)}\|_2} \right\rangle \cdot \left\langle \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}, \mathbf{z}_i \right\rangle + \max_i \|\mathbf{z}_i\|_2 \cdot \mathcal{E}^{(t_1+1)} \\
&\leq \left\langle \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}, \mathbf{z}_i \right\rangle + \max_i \|\mathbf{z}_i\|_2 \cdot \mathcal{E}^{(t_1+1)} \\
&= \|\mathbf{w}^*\|_2^{-1} + \max_i \|\mathbf{z}_i\|_2 \cdot \mathcal{E}^{(t_1+1)}.
\end{aligned}$$

Therefore, we have

$$\left| \left\langle \frac{\mathbf{w}^{(t_1+1)}}{\|\mathbf{w}^{(t_1+1)}\|_2}, \mathbf{z}_i \right\rangle - \alpha \right| = \left| \left\langle \frac{\mathbf{w}^{(t_1+1)}}{\|\mathbf{w}^{(t_1+1)}\|_2}, \mathbf{z}_i \right\rangle - \|\mathbf{w}^*\|_2^{-1} \right| \leq 2 \max_i \|\mathbf{z}_i\|_2 \cdot \mathcal{E}^{(t_1+1)}. \tag{C.41}$$

Moreover, by the induction hypothesis (iii) at iteration $t_1 + 1$ (which has been proved), we have

$$\begin{aligned}
\gamma^{(t_1+1)} &\leq \log[8\eta \cdot (t_1 + 1 - t_0) + 2\exp(\gamma^{(t_0)})] \\
&= \log(9) + \log[(8/9)\eta \cdot (t_1 + 1 - t_0) + (2/9) \cdot \exp(\gamma^{(t_0)})] \\
&\leq \log(9) + \log[(8/9)\eta \cdot (t_1 + 1 - t_0) + (2/9) \cdot \exp(1.5)] \\
&\leq \log(9) + \log[(8/9)\eta \cdot (t_1 + 1 - t_0) + 1].
\end{aligned} \tag{C.42}$$

Therefore by (C.39), (C.41) and (C.42), we have

$$\begin{aligned}
&\left| \left\langle \frac{\mathbf{w}^{(t_1+1)}}{\|\mathbf{w}^{(t_1+1)}\|_2}, \mathbf{z}_i \right\rangle - \alpha \right| \cdot \gamma^{(t_1+1)} \\
&\quad \leq \frac{2 \max_i \|\mathbf{z}_i\|_2 \cdot \epsilon \cdot \{\log(9) + \log[(8/9)\eta \cdot (t_1 + 1 - t_0) + 1]\}}{\exp\left[\frac{\lambda_{\min}}{1024 \lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \log^2((8/9)\eta \cdot (t_1 + 1 - t_0) + 1)\right]} \\
&\quad \leq 2\log(9) \cdot \max_i \|\mathbf{z}_i\|_2 \cdot \epsilon + \frac{32\sqrt{2} \cdot \exp(1/2) \cdot \lambda_{\max}^{3/4}}{\lambda_{\min}^{1/2}} \cdot \max_i \|\mathbf{z}_i\|_2 \cdot \epsilon \cdot \|\mathbf{w}^{(0)}\|_2 \\
&\quad \leq 6 \max_i \|\mathbf{z}_i\|_2 \cdot \epsilon + \frac{80 \lambda_{\max}^{3/4}}{\lambda_{\min}^{1/2}} \cdot \max_i \|\mathbf{z}_i\|_2 \cdot \epsilon \cdot \|\mathbf{w}^{(0)}\|_2 \\
&\quad \leq \|\mathbf{w}^*\|_2^{-1}/4.
\end{aligned}$$

where the second inequality follows by the fact that $\exp(-Az^2) \cdot z \leq \exp(-1/2) \cdot (2A)^{-1/2}$ for all $A, z > 0$, and the last inequality follows by the definition of ϵ which ensures that $\epsilon \leq (16 \max_i \|\mathbf{z}_i\|_2)^{-1} \cdot \|\mathbf{w}^*\|_2^{-1} \cdot \min\{1/3, \|\mathbf{w}^{(0)}\|_2^{-1} \cdot \lambda_{\min}^{1/2}/(40\lambda_{\max}^{3/4})\}$. This finishes the proof of

induction hypothesis (v) at iteration $t_1 + 1$, and thus the first five results in Lemma 4.6 hold for all $t \geq t_0$.

Proof of the last result in Lemma 4.6. As we have shown by induction, the first five results in Lemma 4.6 hold for all $t \geq t_0$. Therefore (C.27) and (C.27) also hold for all $t \geq t_0$. Therefore by the monotonicity of $\ell(\cdot)$ and the third result in Lemma 4.6, we have

$$\begin{aligned}\ell(y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)) &\leq \log(1 + \exp(-\gamma^{(t)} + 1/4)) \\ &\leq \exp\{-\log[(\eta/8) \cdot (t - t_0) + \exp(\gamma^{(t_0)})] + 1/4\} \\ &= \frac{\exp(1/4)}{(\eta/8) \cdot (t - t_0) + \exp(\gamma^{(t_0)})} \\ &\leq \frac{12}{\eta \cdot (t - t_0) + 1}\end{aligned}$$

for all t $\geq t_0$, where the second inequality follows by $\log(1 + z) \geq z$ for all $z \in \mathbb{R}$. Similarly, we also have

$$\begin{aligned}\ell(y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)) &\geq \log(1 + \exp(-\gamma^{(t)} - 3/4)) \\ &\geq \exp\{-\log[8\eta \cdot (t - t_0) + 2\exp(\gamma^{(t_0)})] - 3/4\}/2 \\ &= \frac{\exp(-3/4)/2}{8\eta \cdot (t - t_0) + 2\exp(\gamma^{(t_0)})} \\ &\geq \frac{1}{40} \cdot \frac{1}{\eta \cdot (t - t_0) + 1}\end{aligned}$$

for all t $\geq t_0$, where the second inequality follows by the fact that $\log(1+z) \geq z/2$ for all $z \in [0, 1]$. This proves the first part of the result.

As for the second part of the result, by Lemma 4.3, for all $t \geq t_0$ we have

$$\begin{aligned}\frac{1}{n^2} \sum_{i,i'=1}^{n} (\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - \langle \mathbf{w}^{(t_1+1)}, \mathbf{z}_i \rangle)^2 &= 2\|\mathbf{w}^{(t)} - \langle \mathbf{w}^*, \mathbf{w}^{(t)} \rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_{\mathbf{\Sigma}}^2 \\ &\leq 2\lambda_{\max} \cdot \left\| \mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)} \rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^* \right\|_2^2. \qquad \text{(C.43)}\end{aligned}$$

Now that the fourth result in Lemma 4.6, which has been proved to hold for all $t \geq t_0$, we have

$$\left\| \mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)} \rangle}{\|\mathbf{w}^*\|_2^2} \cdot \mathbf{w}^* \right\|_2^2 \leq \epsilon \cdot \|\mathbf{w}^{(0)}\|_2 \cdot \exp\left[- \frac{\lambda_{\min} \cdot \log^2((8/9)\eta \cdot (t - t_0) + 1)}{1024\lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \right] \qquad \text{(C.44)}$$

for all $t \geq t_0$. Combining (C.44) and (C.43) then gives

$$\begin{aligned}&\frac{1}{n^2} \sum_{i,i'=1}^{n} (\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle)^2 \\ &\quad \leq \lambda_{\max} \cdot \epsilon^2 \cdot \|\mathbf{w}^{(0)}\|_2^2 \cdot \exp\left[- \frac{\lambda_{\min}}{512\lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \log^2((8/9)\eta \cdot (t - t_0) + 1) \right].\end{aligned}$$

Then by the definitions of $D(\mathbf{w})$ and ϵ, we have

$$
\begin{aligned}
&D(\mathbf{w}^{(t)}) \\
&\leq \frac{1}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \cdot \frac{1}{n^2} \sum_{i,i'=1}^{n} (\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle)^2 \\
&\leq \frac{1}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \cdot \frac{\lambda_{\max} \cdot \|\mathbf{w}^{(0)}\|_2^2}{2304 \max_i \|\mathbf{x}_i\|_2^2 \cdot \|\mathbf{w}^*\|_2^2} \cdot \exp\bigg[- \frac{\lambda_{\min}}{512 \lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \log^2((8/9)\eta \cdot (t - t_0) + 1) \bigg].
\end{aligned}
$$

By the definition of $\mathbf{w}^*$, clearly we have $\max_i \|\mathbf{x}_i\|_2^2 \cdot \|\mathbf{w}^*\|_2^2 \geq \min_i \langle y_i \cdot \mathbf{x}_i, \mathbf{w}^* \rangle^2 = 1$. Therefore

$$
\begin{aligned}
D(\mathbf{w}^{(t)}) &\leq \frac{\lambda_{\max} \cdot \|\mathbf{w}^{(0)}\|_2^2}{2304 \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2} \cdot \exp\bigg[- \frac{\lambda_{\min}}{512 \lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \log^2((8/9)\eta \cdot (t - t_0) + 1) \bigg] \\
&\leq \frac{\lambda_{\max} \cdot \|\mathbf{w}^{(0)}\|_2^2}{2304 \lambda_{\min} \|\mathbf{w}^{(t)}\|_2^2} \cdot \exp\bigg[- \frac{\lambda_{\min}}{512 \lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \log^2((8/9)\eta \cdot (t - t_0) + 1) \bigg] \\
&\leq \frac{\lambda_{\max}}{2304 \lambda_{\min}} \cdot \exp\bigg[- \frac{\lambda_{\min}}{512 \lambda_{\max}^{3/2} \cdot \|\mathbf{w}^{(0)}\|_2^2} \cdot \log^2((8/9)\eta \cdot (t - t_0) + 1) \bigg]
\end{aligned}
$$

for all $t \geq t_0$, where the last inequality follows by Lemma 4.4. Therefore the last result in Lemma 4.6 holds, and the proof of Lemma 4.6 is thus complete. ■

Appendix D. Proofs for Batch Normalization in Two-Layer Linear CNNs

D.1. Proof of Theorem 3.2

As most part of the proof of Theorem 3.2 are the same as the proof of Theorem 2.2, here we only highlight the differences between these two proofs. Specifically, we give the proofs of the counterparts of Lemmas 4.1, 4.2, 4.4. The rest of the proofs are essentially the same based on the multi-patch versions of these three results.

By Assumption 3.1, we can define $\mathbf{w}^*$ as the minimum norm solution of the system:

$$
\mathbf{w}^* := \operatorname*{argmin}_{\mathbf{w}} \|\mathbf{w}\|_2^2, \quad \text{subject to } \langle \mathbf{w}, y_i \cdot \mathbf{x}_i^{(p)} \rangle = 1, \ i \in [n], \ p \in [P]. \tag{D.1}
$$

The following lemma is the counterpart of Lemma 4.1.

Lemma D.1 *Under Assumption 2.1, for any $\mathbf{w} \in \mathbb{R}^d$, it holds that*

$$
\begin{aligned}
\langle -\nabla_{\mathbf{w}} L(\mathbf{w}, \gamma), \mathbf{w}^* \rangle = &\frac{\gamma}{2n^2 P \|\mathbf{w}\|_{\mathbf{\Sigma}}^3} \sum_{i,i'=1}^{n} |\ell_i'| \cdot |\ell_{i'}'| \cdot (\langle \mathbf{w}, \mathbf{z}_{i'} \rangle - \langle \mathbf{w}, \mathbf{z}_i \rangle) \cdot (|\ell_{i'}'|^{-1} \cdot \langle \mathbf{w}, \mathbf{z}_{i'} \rangle - |\ell_i'|^{-1} \cdot \langle \mathbf{w}, \mathbf{z}_i \rangle) \\
&+ \frac{\gamma}{2n^2 P \|\mathbf{w}\|_{\mathbf{\Sigma}}^3} \cdot \bigg(\sum_{i=1}^{n} |\ell_i'| \bigg) \cdot \sum_{i'=1}^{n} \sum_{p,p'=1}^{P} \big(\langle \mathbf{w}, \mathbf{z}_{i'}^{(p)} \rangle - \langle \mathbf{w}, \mathbf{z}_{i'}^{(p')} \rangle \big)^2,
\end{aligned}
$$

where $\ell_i' = \ell'[y_i \cdot f(\mathbf{w}, \gamma, \mathbf{x}_i)]$, $\mathbf{z}_i^{(p)} = y_i \cdot \mathbf{x}_i^{(p)}$, and $\mathbf{z}_i = \sum_{p=1}^{P} \mathbf{z}_i^{(p)}$ for $i \in [n]$, $p \in [P]$.

Proof [Proof of Lemma D.1] By definition, we have

$$\nabla_{\mathbf{w}} L(\mathbf{w}, \gamma) = \frac{1}{n} \sum_{i=1}^{n} \ell'[y_i \cdot f(\mathbf{w}, \gamma, \mathbf{x}_i)] \cdot y_i \cdot \nabla_{\mathbf{w}} f(\mathbf{w}, \gamma, \mathbf{x}_i).$$

Then by the definition of $f(\mathbf{w}, \gamma, \mathbf{x})$, we have the following calculation using chain rule:

$$\begin{aligned}
\nabla_{\mathbf{w}} L(\mathbf{w}, \gamma) &= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-1} \cdot \frac{1}{n} \sum_{i=1}^{n} \ell'[y_i \cdot f(\mathbf{w}, \gamma, \mathbf{x}_i)] \cdot y_i \cdot \sum_{p=1}^{P} \gamma \cdot \Big(\mathbf{I} - \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-2} \cdot \mathbf{\Sigma} \mathbf{w} \mathbf{w}^\top\Big) \mathbf{x}_i^{(p)} \\
&= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n} \sum_{i=1}^{n} \sum_{p=1}^{P} \ell_i' \cdot y_i \cdot \Big(\|\mathbf{w}\|_{\mathbf{\Sigma}}^{2} - \mathbf{\Sigma} \mathbf{w} \mathbf{w}^\top\Big) \mathbf{x}_i^{(p)} \\
&= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n} \sum_{i=1}^{n} \sum_{p=1}^{P} \ell_i' \cdot \left(\frac{1}{nP} \sum_{i'=1}^{n} \sum_{p'=1}^{P} \langle \mathbf{w}, \mathbf{z}_{i'}^{(p')} \rangle^2 - \frac{1}{nP} \sum_{i'=1}^{n} \sum_{p'=1}^{P} \langle \mathbf{w}, \mathbf{z}_{i'}^{(p')} \rangle \cdot \mathbf{z}_{i'}^{(p')} \mathbf{w}^\top \right) \mathbf{z}_i^{(p)} \\
&= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2 P} \sum_{i,i'=1}^{n} \sum_{p,p'=1}^{P} \Big[\ell_i' \cdot \langle \mathbf{w}, \mathbf{z}_{i'}^{(p')} \rangle^2 \cdot \mathbf{z}_i^{(p)} - \ell_i' \cdot \langle \mathbf{w}, \mathbf{z}_{i'}^{(p')} \rangle \cdot \langle \mathbf{w}, \mathbf{z}_i^{(p)} \rangle \cdot \mathbf{z}_{i'}^{(p')} \Big],
\end{aligned}$$

where we denote $\mathbf{z}_i^{(p)} = y_i \cdot \mathbf{x}_i^{(p)}$, $i \in [n]$, $p \in [P]$. By Assumption 3.1, taking inner product with $-\mathbf{w}^*$ on both sides above then gives

$$\begin{aligned}
-\langle \nabla_{\mathbf{w}} L(\mathbf{w}, \gamma), \mathbf{w}^* \rangle &= -\|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2 P} \sum_{i,i'=1}^{n} \sum_{p,p'=1}^{P} \Big[\ell_i' \cdot \langle \mathbf{w}, \mathbf{z}_{i'}^{(p')} \rangle^2 - \ell_i' \cdot \langle \mathbf{w}, \mathbf{z}_{i'}^{(p')} \rangle \cdot \langle \mathbf{w}, \mathbf{z}_i^{(p)} \rangle \Big] \\
&= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2 P} \sum_{i,i'=1}^{n} \sum_{p,p'=1}^{P} \Big[|\ell_i'| \cdot \langle \mathbf{w}, \mathbf{z}_{i'}^{(p')} \rangle^2 - |\ell_i'| \cdot \langle \mathbf{w}, \mathbf{z}_{i'}^{(p')} \rangle \cdot \langle \mathbf{w}, \mathbf{z}_i^{(p)} \rangle \Big],
\end{aligned}$$

where the second equality follows by the fact that $\ell_i' < 0$, $i \in [n]$. Further denote $u_i^{(p)} = \langle \mathbf{w}, \mathbf{z}_i \rangle$ for $i \in [n]$, $p \in [P]$. Then we have

$$\begin{aligned}
-\langle \nabla_{\mathbf{w}} L(\mathbf{w}, \gamma), \mathbf{w}^* \rangle &= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2 P} \sum_{i,i'=1}^{n} \sum_{p,p'=1}^{P} \big[|\ell_i'| \cdot u_{i'}^{(p')2} - |\ell_i'| \cdot u_{i'}^{(p')} u_i^{(p)} \big] \\
&= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2 P} \sum_{i,i'=1}^{n} \left[|\ell_i'| \cdot \left(\sum_{p,p'=1}^{P} u_{i'}^{(p')2} - \sum_{p,p'=1}^{P} u_{i'}^{(p')} u_i^{(p)} \right) \right] \\
&= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2 P} \sum_{i,i'=1}^{n} \left\{ |\ell_i'| \cdot \left[P \cdot \sum_{p'=1}^{P} (u_{i'}^{(p')})^2 - \left(\sum_{p=1}^{P} u_i^{(p)} \right) \cdot \left(\sum_{p'=1}^{P} u_{i'}^{(p')} \right) \right] \right\}.
\end{aligned}$$

Adding and subtracting a term

$$\|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2 P} \sum_{i,i'=1}^{n} \left[|\ell_i'| \cdot \left(\sum_{p'=1}^{P} u_{i'}^{(p')} \right)^2 \right]$$

then gives

$$-\langle \nabla_{\mathbf{w}} L(\mathbf{w},\gamma), \mathbf{w}^*\rangle = \underbrace{\|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2 P} \sum_{i,i'=1}^{n} \left\{ |\ell_i'| \cdot \left[\left(\sum_{p'=1}^{P} u_{i'}^{(p')} \right)^2 - \left(\sum_{p=1}^{P} u_i^{(p)} \right) \cdot \left(\sum_{p'=1}^{P} u_{i'}^{(p')} \right) \right] \right\}}_{I_1}$$
$$+ \underbrace{\|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2 P} \sum_{i,i'=1}^{n} \left\{ |\ell_i'| \cdot \left[P \cdot \sum_{p'=1}^{P} (u_{i'}^{(p')})^2 - \left(\sum_{p'=1}^{P} u_{i'}^{(p')} \right)^2 \right] \right\}}_{I_2}. \tag{D.2}$$

We then calculate the two terms I_1 and I_2 in (D.2) separately. The calculation for I_1 is the same as the derivation in the proof of Lemma 4.1. To see this, let $u_i = \sum_{p=1}^{P} u_i^{(p)}$ for $i \in [n], p \in [P]$. Then we have

$$I_1 = \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2 P} \sum_{i,i'=1}^{n} \big(|\ell_i'| \cdot u_{i'}^2 - |\ell_i'| \cdot u_{i'} u_i \big). \tag{D.3}$$

Switching the index notations i, i' in the above equation also gives

$$I_1 = \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2 P} \sum_{i,i'=1}^{n} \big(|\ell_{i'}'| \cdot u_i^2 - |\ell_{i'}'| \cdot u_{i'} u_i \big). \tag{D.4}$$

We can add (D.3) and (D.4) together to obtain

$$\begin{aligned}
2I_1 &= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2 P} \sum_{i,i'=1}^{n} \big(|\ell_i'| \cdot u_{i'}^2 - |\ell_i'| \cdot u_{i'} u_i + |\ell_{i'}'| \cdot u_i^2 - |\ell_{i'}'| \cdot u_{i'} u_i \big) \\
&= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2 P} \sum_{i,i'=1}^{n} (u_{i'} - u_i) \cdot \big(|\ell_i'| \cdot u_{i'} - |\ell_{i'}'| \cdot u_i \big) \\
&= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2 P} \sum_{i,i'=1}^{n} |\ell_i'| \cdot |\ell_{i'}'| \cdot (u_{i'} - u_i) \cdot \big(|\ell_{i'}'|^{-1} \cdot u_{i'} - |\ell_i'|^{-1} \cdot u_i \big).
\end{aligned}$$

Therefore we have

$$I_1 = \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{2n^2 P} \sum_{i,i'=1}^{n} |\ell_i'| \cdot |\ell_{i'}'| \cdot (u_{i'} - u_i) \cdot \big(|\ell_{i'}'|^{-1} \cdot u_{i'} - |\ell_i'|^{-1} \cdot u_i \big). \tag{D.5}$$

This completes the calculation for I_1. We then proceed to calculate I_2. for any $i' \in [n]$, we can directly check that the following identity holds:

$$\frac{1}{2} \cdot \sum_{p,p'=1}^{P} \big(u_{i'}^{(p)} - u_{i'}^{(p')} \big)^2 = \frac{1}{2} \cdot \sum_{p,p'=1}^{P} \big[(u_{i'}^{(p)})^2 + (u_{i'}^{(p')})^2 - 2 u_{i'}^{(p)} u_{i'}^{(p')} \big] = P \cdot \sum_{p'=1}^{P} (u_{i'}^{(p')})^2 - \left(\sum_{p'=1}^{P} u_{i'}^{(p')} \right)^2.$$

Note that the right hand side above appears in I_2. Plugging the above calculation into the definition of I_2 gives

$$\begin{aligned} I_2 &= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{n^2 P} \sum_{i,i'=1}^{n} \left\{ |\ell_i'| \cdot \left[\frac{1}{2} \cdot \sum_{p,p'=1}^{P} \big(u_{i'}^{(p)} - u_{i'}^{(p')}\big)^2 \right] \right\} \\ &= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{2n^2 P} \sum_{i,i'=1}^{n} \sum_{p,p'=1}^{P} |\ell_i'| \cdot \big(u_{i'}^{(p)} - u_{i'}^{(p')}\big)^2 \\ &= \|\mathbf{w}\|_{\mathbf{\Sigma}}^{-3} \cdot \frac{\gamma}{2n^2 P} \cdot \left(\sum_{i=1}^{n} |\ell_i'| \right) \cdot \sum_{i'=1}^{n} \sum_{p,p'=1}^{P} \big(u_{i'}^{(p)} - u_{i'}^{(p')}\big)^2. \end{aligned} \tag{D.6}$$

Finally, plugging (D.5) and (D.6) into (D.2) completes the proof. ■

The following lemma follows by exactly the same proof as Lemma 4.3.

Lemma D.2 *For any $\mathbf{w} \in \mathbb{R}^d$, it holds that*

$$\frac{1}{n^2 P^2} \sum_{i,i'=1}^{n} \sum_{p,p'=1}^{P} \big(\langle \mathbf{w}, \mathbf{z}_{i'} \rangle - \langle \mathbf{w}, \mathbf{z}_i \rangle\big)^2 = \|\mathbf{w} - \langle \mathbf{w}^*, \mathbf{w} \rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_{\mathbf{\Sigma}}^2.$$

The following lemma is the counterpart of Lemma 4.2.

Lemma D.3 *For all $t \geq 0$, it holds that*

$$\langle \mathbf{w}^{(t+1)}, \mathbf{w}^* \rangle \geq \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle + \frac{\eta P \gamma^{(t)} \cdot \exp(-\gamma^{(t)})}{16 \cdot \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \cdot \|\mathbf{w} - \langle \mathbf{w}^*, \mathbf{w} \rangle_{\mathbf{\Sigma}} \cdot \mathbf{w}^*\|_{\mathbf{\Sigma}}^2.$$

Proof [Proof of Lemma D.3]

By Lemma D.1 and the gradient descent update rule, we have

$$\begin{aligned} &\langle \mathbf{w}^{(t+1)}, \mathbf{w}^* \rangle - \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle = \eta \cdot \langle -\nabla_{\mathbf{w}} L(\mathbf{w}^{(t)}, \gamma^{(t)}), \mathbf{w}^* \rangle \\ &= \frac{\eta \gamma^{(t)}}{2n^2 P \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \sum_{i,i'=1}^{n} |\ell_i'^{(t)}| \cdot |\ell_{i'}'^{(t)}| \cdot \big(\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle\big) \cdot \big(|\ell_{i'}'^{(t)}|^{-1} \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - |\ell_i'^{(t)}|^{-1} \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle\big) \\ &\quad + \frac{\eta \gamma^{(t)}}{2n^2 P \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \cdot \left(\sum_{i=1}^{n} |\ell_i'^{(t)}| \right) \cdot \sum_{i'=1}^{n} \sum_{p,p'=1}^{P} \big(\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}^{(p)} \rangle - \langle \mathbf{w}, \mathbf{z}_{i'}^{(p')} \rangle\big)^2 \end{aligned} \tag{D.7}$$

for all $t \geq 0$, $\mathbf{z}_i^{(p)} = y_i \cdot \mathbf{x}_i^{(p)}$, and $\mathbf{z}_i = \sum_{p=1}^{P} \mathbf{z}_i^{(p)}$ for $i \in [n]$, $p \in [P]$. Note that

$$y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i) = y_i \cdot \sum_{p=1}^{P} \gamma^{(t)} \cdot \frac{\langle \mathbf{w}^{(t)}, \mathbf{x}_i^{(p)} \rangle}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}} = \gamma^{(t)} \cdot \frac{\langle \mathbf{w}^{(t)}, \mathbf{z}_i^{(p)} \rangle}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}},$$

and

$$|\ell_i'^{(t)}|^{-1} = -\{\ell'[y_i \cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)]\}^{-1} = 1 + \exp(\gamma^{(t)} \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle / \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}).$$

With the exact same derivation as (C.8) in the proof of Lemma 4.1, we have

$$\sum_{i,i'=1}^{n} |\ell_i'^{(t)}| \cdot |\ell_{i'}'^{(t)}| \cdot (\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle) \cdot (|\ell_{i'}'^{(t)}|^{-1} \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - |\ell_i'^{(t)}|^{-1} \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle)$$
$$\geq \frac{1}{8} \exp(-\gamma^{(t)}) \sum_{i,i'=1}^{n} (\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle)^2. \tag{D.8}$$

Moreover, we also have

$$\frac{1}{n} \sum_{i=1}^{n} |\ell_i'^{(t)}| = \frac{1}{n} \sum_{i=1}^{n} \left[1 + \exp\left(\frac{\gamma^{(t)} \cdot \langle \mathbf{w}^{(t)}, \mathbf{z}_j \rangle}{\sqrt{\frac{1}{n} \sum_{j=1}^{n} \langle \mathbf{w}^{(t)}, \mathbf{z}_j \rangle^2}} \right) \right]^{-1}$$
$$\geq \frac{1}{n} \sum_{i=1}^{n} \left[1 + \exp\left(\frac{\gamma^{(t)} \cdot |\langle \mathbf{w}^{(t)}, \mathbf{z}_j \rangle|}{\sqrt{\frac{1}{n} \sum_{j=1}^{n} \langle \mathbf{w}^{(t)}, \mathbf{z}_j \rangle^2}} \right) \right]^{-1},$$

where the inequality follows by the fact that $[1 + \exp(z)]^{-1}$ is a decreasing function. Further note that $[1 + \exp(z)]^{-1}$ is convex over $z \in [0, +\infty)$. Therefore by Jensen's inequality, we have

$$\frac{1}{n} \sum_{i=1}^{n} |\ell_i'^{(t)}| \geq \frac{1}{n} \sum_{i=1}^{n} \left[1 + \exp\left(\frac{\gamma^{(t)} \cdot |\langle \mathbf{w}^{(t)}, \mathbf{z}_j \rangle|}{\sqrt{\frac{1}{n} \sum_{j=1}^{n} \langle \mathbf{w}^{(t)}, \mathbf{z}_j \rangle^2}} \right) \right]^{-1}$$
$$\geq \left[1 + \exp\left(\frac{1}{n} \sum_{i=1}^{n} \frac{\gamma^{(t)} \cdot |\langle \mathbf{w}^{(t)}, \mathbf{z}_j \rangle|}{\sqrt{\frac{1}{n} \sum_{j=1}^{n} \langle \mathbf{w}^{(t)}, \mathbf{z}_j \rangle^2}} \right) \right]^{-1}$$
$$\geq [1 + \exp(\gamma^{(t)})]^{-1}$$
$$\geq \exp(-\gamma^{(t)})/2. \tag{D.9}$$

Plugging (D.8) and (D.9) into (D.7), we obtain

$$\langle \mathbf{w}^{(t+1)}, \mathbf{w}^* \rangle - \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle$$
$$\geq \frac{\eta \gamma^{(t)}}{16 n^2 P \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \cdot \exp(-\gamma^{(t)}) \sum_{i,i'=1}^{n} (\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle)^2$$
$$+ \frac{\eta \gamma^{(t)}}{4 n^2 P \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \cdot \exp(-\gamma^{(t)}) \cdot \sum_{i=1}^{n} \sum_{p,p'=1}^{P} \big(\langle \mathbf{w}^{(t)}, \mathbf{z}_i^{(p)} \rangle - \langle \mathbf{w}, \mathbf{z}_i^{(p')} \rangle \big)^2$$
$$\geq \frac{\eta \gamma^{(t)} \cdot \exp(-\gamma^{(t)})}{16 P \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \cdot \left[\frac{1}{n^2} \sum_{i,i'=1}^{n} (\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'} \rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_i \rangle)^2 + \frac{1}{n} \sum_{i=1}^{n} \sum_{p,p'=1}^{P} \big(\langle \mathbf{w}^{(t)}, \mathbf{z}_i^{(p)} \rangle - \langle \mathbf{w}, \mathbf{z}_i^{(p')} \rangle \big)^2 \right].$$

Recall that $\mathbf{z}_i = \sum_{p=1}^{P} \mathbf{z}_i^{(p)}$. Denoting

$$u_{i,p} = \langle \mathbf{w}^{(t)}, \mathbf{z}_i^{(p)} \rangle - \frac{1}{nP} \sum_{i'=1}^{n} \sum_{p'=1}^{P} \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}^{(p')} \rangle$$

for $i \in [n]$ and $p \in [P]$, we have

$$\langle \mathbf{w}^{(t+1)}, \mathbf{w}^* \rangle - \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle \geq \frac{\eta \gamma^{(t)} \cdot \exp(-\gamma^{(t)})}{16P \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \cdot (I_1 + I_2), \tag{D.10}$$

where

$$I_1 = \frac{1}{n^2} \sum_{i,i'=1}^{n} \left(\sum_{p=1}^{p} u_{i,p} - \sum_{p=1}^{p} u_{i',p} \right)^2, \qquad I_2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{p,p'=1}^{P} (u_{i,p} - u_{i,p'})^2.$$

By direct calculation, we have

$$\begin{aligned} I_1 &= \frac{1}{n^2} \sum_{i,i'=1}^{n} \left[\left(\sum_{p=1}^{p} u_{i,p} \right)^2 - 2 \cdot \left(\sum_{p=1}^{p} u_{i,p} \right) \cdot \left(\sum_{p=1}^{p} u_{i',p} \right) + \left(\sum_{p=1}^{p} u_{i',p} \right)^2 \right] \\ &= \frac{2}{n} \sum_{i=1}^{n} \left(\sum_{p=1}^{p} u_{i,p} \right)^2 - \frac{2}{n^2} \cdot \left(\sum_{i=1}^{n} \sum_{p=1}^{p} u_{i,p} \right)^2 \\ &= \frac{2}{n} \sum_{i=1}^{n} \left(\sum_{p=1}^{p} u_{i,p} \right)^2, \end{aligned} \tag{D.11}$$

where the last equality follows by the definition of $u_{i,p}$. Moreover, we have

$$I_2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{p,p'=1}^{P} (u_{i,p}^2 - 2u_{i,p}u_{i,p'} + u_{i,p'}^2) = \frac{2P}{n} \sum_{i=1}^{n} \sum_{p=1}^{P} u_{i,p}^2 - \frac{2}{n} \sum_{i=1}^{n} \left(\sum_{p=1}^{P} u_{i,p} \right)^2. \tag{D.12}$$

Plugging (D.11), (D.12) and the definition of $u_{i,p}$ into (D.10) gives

$$\langle \mathbf{w}^{(t+1)}, \mathbf{w}^* \rangle - \langle \mathbf{w}^{(t)}, \mathbf{w}^* \rangle \geq \frac{\eta \gamma^{(t)} \cdot \exp(-\gamma^{(t)})}{8n \cdot \|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3} \cdot \sum_{i=1}^{n} \sum_{p=1}^{P} \left(\langle \mathbf{w}^{(t)}, \mathbf{z}_i^{(p)} \rangle - \frac{1}{nP} \sum_{i'=1}^{n} \sum_{p'=1}^{P} \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}^{(p')} \rangle \right)^2.$$

We continue the calculation as follows:

$$
\begin{aligned}
&\frac{1}{nP}\sum_{i=1}^{n}\sum_{p=1}^{P}\left(\langle \mathbf{w}^{(t)}, \mathbf{z}_i^{(p)}\rangle - \frac{1}{nP}\sum_{i'=1}^{n}\sum_{p'=1}^{P}\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}^{(p')}\rangle\right)^2 \\
&\quad = \frac{1}{nP}\sum_{i=1}^{n}\sum_{p=1}^{P}\langle \mathbf{w}^{(t)}, \mathbf{z}_i^{(p)}\rangle^2 - \left(\frac{1}{nP}\sum_{i=1}^{n}\sum_{p=1}^{P}\langle \mathbf{w}^{(t)}, \mathbf{z}_i^{(p)}\rangle\right)^2 \\
&\quad = \frac{1}{2}\cdot\left[\frac{2}{nP}\sum_{i=1}^{n}\sum_{p=1}^{P}\langle \mathbf{w}^{(t)}, \mathbf{z}_i^{(p)}\rangle^2 - 2\cdot\left(\frac{1}{nP}\sum_{i=1}^{n}\sum_{p=1}^{P}\langle \mathbf{w}^{(t)}, \mathbf{z}_i^{(p)}\rangle\right)^2\right] \\
&\quad = \frac{1}{2}\cdot\left[\frac{1}{nP}\sum_{i=1}^{n}\sum_{p=1}^{P}\langle \mathbf{w}^{(t)}, \mathbf{z}_i^{(p)}\rangle^2 - 2\cdot\left(\frac{1}{nP}\sum_{i=1}^{n}\sum_{p=1}^{P}\langle \mathbf{w}^{(t)}, \mathbf{z}_i^{(p)}\rangle\right)^2 + \frac{1}{nP}\sum_{i'=1}^{n}\sum_{p'=1}^{P}\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}^{(p')}\rangle^2\right] \\
&\quad = \frac{1}{2n^2P^2}\cdot\sum_{i,i'=1}^{n}\sum_{p,p'=1}^{P}\left(\langle \mathbf{w}^{(t)}, \mathbf{z}_i^{(p)}\rangle^2 - 2\cdot\langle \mathbf{w}^{(t)}, \mathbf{z}_i^{(p)}\rangle\langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}^{(p')}\rangle + \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}^{(p')}\rangle^2\right) \\
&\quad = \frac{1}{2n^2P^2}\cdot\sum_{i,i'=1}^{n}\sum_{p,p'=1}^{P}\left(\langle \mathbf{w}^{(t)}, \mathbf{z}_i^{(p)}\rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}^{(p')}\rangle\right)^2.
\end{aligned}
$$

Therefore we have

$$
\langle \mathbf{w}^{(t+1)}, \mathbf{w}^*\rangle - \langle \mathbf{w}^{(t)}, \mathbf{w}^*\rangle \geq \frac{\eta\gamma^{(t)}\cdot\exp(-\gamma^{(t)})}{16n^2P\cdot\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^3}\cdot\sum_{i,i'=1}^{n}\sum_{p,p'=1}^{P}\left(\langle \mathbf{w}^{(t)}, \mathbf{z}_i^{(p)}\rangle - \langle \mathbf{w}^{(t)}, \mathbf{z}_{i'}^{(p')}\rangle\right)^2.
$$

Applying Lemma D.2 finishes the proof. ■

The following lemma is the counterpart of Lemma 4.2.

Lemma D.4 *For all $t \geq 0$, it holds that*

$$
\|\mathbf{w}^{(t)}\|_2^2 \leq \|\mathbf{w}^{(t+1)}\|_2^2 \leq \|\mathbf{w}^{(t)}\|_2^2 + 4\eta^2\cdot\frac{P^2\gamma^{(t)2}\cdot\max_i\|\mathbf{x}_i\|_2^3}{\lambda_{\min}^2\cdot\|\mathbf{w}^{(t)}\|_2^2}.
$$

Moreover, if $\|\mathbf{w}^{(t)} - \langle \mathbf{w}^{(t)}, \mathbf{w}^*\rangle\cdot\|\mathbf{w}^*\|_2^{-1}\cdot\mathbf{w}^*\|_2 \leq \|\mathbf{w}^{(0)}\|_2/2$, then

$$
\begin{aligned}
\|\mathbf{w}^{(t+1)}\|_2^2 \leq \|\mathbf{w}^{(t)}\|_2^2 &+ \eta^2 G\cdot\max\{\gamma^{(t)2}, \gamma^{(t)4}\}\cdot\max\{|\ell_1'^{(t)}|^2, \ldots, |\ell_n'^{(t)}|^2, \exp(-2\gamma^{(t)})\} \\
&\cdot\left\|\mathbf{w}^{(t)} - \frac{\langle \mathbf{w}^*, \mathbf{w}^{(t)}\rangle}{\|\mathbf{w}^*\|_2^2}\cdot\mathbf{w}^*\right\|_2^2,
\end{aligned}
$$

where $G = 64P^3\lambda_{\min}^{-3}\cdot\max_i\|\mathbf{x}_i\|_2^6\cdot\|\mathbf{w}^{(0)}\|_2^{-4}$.

Proof [Proof of Lemma D.4] Note that

$$
\begin{aligned}
\nabla_{\mathbf{w}}L(\mathbf{w}^{(t)}, \gamma^{(t)}) &= \frac{1}{n\cdot\|\mathbf{w}\|_{\mathbf{\Sigma}}}\sum_{i=1}^{n}\ell'[y_i\cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)]\cdot y_i\cdot\sum_{p=1}^{P}\gamma^{(t)}\cdot\left(\mathbf{I} - \frac{\mathbf{\Sigma}\mathbf{w}^{(t)}\mathbf{w}^{(t)\top}}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2}\right)\mathbf{x}_i^{(p)} \\
&= \frac{1}{n\cdot\|\mathbf{w}\|_{\mathbf{\Sigma}}}\sum_{i=1}^{n}\sum_{p=1}^{P}\ell'[y_i\cdot f(\mathbf{w}^{(t)}, \gamma^{(t)}, \mathbf{x}_i)]\cdot y_i\cdot\gamma^{(t)}\cdot\left(\mathbf{I} - \frac{\mathbf{\Sigma}\mathbf{w}^{(t)}\mathbf{w}^{(t)\top}}{\|\mathbf{w}^{(t)}\|_{\mathbf{\Sigma}}^2}\right)\mathbf{x}_i^{(p)}
\end{aligned}
$$

One can essentially treat $\mathbf{x}_i^{(p)}$, $i \in [n]$ and $p \in [P]$ as different data points and use the same proof as Lemma 4.4 to prove the first inequality. As for the second inequality, for any $\mathbf{w} \in \mathcal{B}^{(t)} := \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w} - \mathbf{w}^{(t)}\|_2 \le \|\mathbf{w}^{(t)}\|_2/2\}$, we have

$$\begin{aligned}
\|\nabla_{\mathbf{w}} f(\mathbf{w}, \gamma^{(t)}, \mathbf{x}_i)\|_2 &= \bigg\| \sum_{p=1}^{P} \frac{\gamma^{(t)}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}} \cdot \bigg(\mathbf{I} - \frac{\mathbf{\Sigma}\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|_{\mathbf{\Sigma}}^2} \bigg) \mathbf{x}_i^{(p)} \bigg\|_2 \\
&\le \frac{P\gamma^{(t)}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}} \cdot \max_{i,p} \|\mathbf{x}_i^{(p)}\|_2 \cdot \bigg(1 + \bigg\| \frac{\mathbf{\Sigma}\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|_{\mathbf{\Sigma}}^2} \bigg\|_2 \bigg) \\
&= \frac{P\gamma^{(t)}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}} \cdot \max_{i,p} \|\mathbf{x}_i^{(p)}\|_2 \cdot \bigg(1 + \frac{\|\mathbf{\Sigma}\mathbf{w}\|_2 \cdot \|\mathbf{w}\|_2}{\|\mathbf{w}\|_{\mathbf{\Sigma}}^2} \bigg) \\
&\le \frac{2P\gamma^{(t)}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}} \cdot \max_{i,p} \|\mathbf{x}_i^{(p)}\|_2 \cdot \frac{\|\mathbf{\Sigma}\mathbf{w}\|_2 \cdot \|\mathbf{w}\|_2}{\|\mathbf{w}\|_{\mathbf{\Sigma}}^2},
\end{aligned}$$

where the last inequality follows by the fact that $\|\mathbf{w}\|_{\mathbf{\Sigma}}^2 = \langle \mathbf{w}, \mathbf{\Sigma}\mathbf{w} \rangle \le \|\mathbf{\Sigma}\mathbf{w}\|_2 \cdot \|\mathbf{w}\|_2$. Further plugging in the definition of $\mathbf{\Sigma}$ gives

$$\begin{aligned}
\|\nabla_{\mathbf{w}} f(\mathbf{w}, \gamma^{(t)}, \mathbf{x}_i)\|_2 &\le \frac{2P\gamma^{(t)}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}^3} \cdot \max_{i,p} \|\mathbf{x}_i^{(p)}\|_2 \cdot \|\mathbf{w}\|_2 \cdot \bigg\| \frac{1}{nP} \sum_{i'=1}^{n} \sum_{p'=1}^{P} \mathbf{x}_{i'}^{(p)} \cdot \langle \mathbf{w}, \mathbf{x}_{i'}^{(p)} \rangle \bigg\|_2 \\
&\le \frac{2P\gamma^{(t)}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}^3} \cdot \max_{i,p} \|\mathbf{x}_i^{(p)}\|_2^2 \cdot \|\mathbf{w}\|_2 \cdot \frac{1}{nP} \sum_{i'=1}^{n} \sum_{p'=1}^{P} |\langle \mathbf{w}, \mathbf{x}_{i'}^{(p)} \rangle| \\
&\le \frac{2P\gamma^{(t)}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}^3} \cdot \max_{i,p} \|\mathbf{x}_i^{(p)}\|_2^2 \cdot \sqrt{\frac{1}{nP} \sum_{i'=1}^{n} \sum_{p'=1}^{P} |\langle \mathbf{w}, \mathbf{x}_{i'}^{(p)} \rangle|^2} \\
&= \frac{2P\gamma^{(t)}}{\|\mathbf{w}\|_{\mathbf{\Sigma}}^2} \cdot \max_{i,p} \|\mathbf{x}_i^{(p)}\|_2^2 \cdot \|\mathbf{w}\|_2 \\
&\le \frac{2P\gamma^{(t)}}{\lambda_{\min} \cdot \|\mathbf{w}\|_2^2} \cdot \max_{i,p} \|\mathbf{x}_i^{(p)}\|_2^2 \cdot \|\mathbf{w}\|_2 \\
&\le \frac{4P\gamma^{(t)}}{\lambda_{\min} \cdot \|\mathbf{w}^{(t)}\|_2} \cdot \max_{i,p} \|\mathbf{x}_i^{(p)}\|_2^2,
\end{aligned}$$

where the third inequality follows by Jensen's inequality, and the last inequality follows by $\|\mathbf{w}\|_2 \ge \|\mathbf{w}^{(t)}\|_2/2$ for all $\mathbf{w} \in \mathcal{B}^{(t)}$. Therefore $f(\mathbf{w}, \gamma^{(t)}, \mathbf{x}_i)$ is $(4\gamma^{(t)}\lambda_{\min}^{-1} \cdot \|\mathbf{w}^{(t)}\|_2^{-1} \cdot \max_{i,p} \|\mathbf{x}_i^{(p)}\|_2^2)$-Lipschitz. The rest of the proof is the same as the proof of Lemma 4.4. ■

D.2. Proof of Theorem 3.3

Proof [Proof of Theorem 3.3] We first show the existence and uniqueness results for the maximum margin and patch-wise uniform margin classifiers. By Example 3.1, the training data patches are given as $\mathbf{x}_i^{(p)} = y_i \cdot \mathbf{u} + \boldsymbol{\xi}_i^{(p)}$ for $i \in [n]$, $p \in [P]$, where $\boldsymbol{\xi}_i^{(p)} \sim N(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{u}\mathbf{u}^\top/\|\mathbf{u}\|_2^2)))$ are orthogonal to $\mathbf{u}$. Therefore it is clear that the linear classifier defined by $\mathbf{u}$ can linearly separate all the data points $(\overline{\mathbf{x}}_i, y_i)$, $i \in [n]$. Therefore the maximum margin solution exists. Its uniqueness

then follows by the definition of the maximum margin problem (3.1), which has a strongly convex objective function and linear constraints.

As for the patch-wise uniform margin classifier, the existence also follows by the observation that $\mathbf{u}$ gives such a classifier with patch-wise uniform margin. Moreover, for a patch-wise uniform margin classifier $\mathbf{w}$, by definition we have

$$y_{i'} \cdot \langle \mathbf{w}, \mathbf{x}_{i'}^{(p')} \rangle = y_i \cdot \langle \mathbf{w}, \mathbf{x}_i^{(p)} \rangle$$

for all $i, i' \in [n]$ and $p, p' \in [P]$. Plugging in the data model $\mathbf{x}_i^{(p)} = y_i \cdot \mathbf{u} + \boldsymbol{\xi}_i^{(p)}$ then gives

$$\langle \mathbf{w}, y_{i'} \cdot \boldsymbol{\xi}_{i'}^{(p')} - y_i \cdot \boldsymbol{\xi}_i^{(p)} \rangle = 0 \tag{D.13}$$

for all $i, i' \in [n]$ and $p, p' \in [P]$. Note that $nP \geq 4n = 2d$. Therefore it is easy to see that with probability 1,

$$\mathrm{span}\big\{ y_{i'} \cdot \boldsymbol{\xi}_{i'}^{(p')} - y_i \cdot \boldsymbol{\xi}_i^{(p)} : i, i' \in [n],\ p, p' \in [P] \big\} = \mathrm{span}\{\mathbf{u}\}^{\perp}.$$

Therefore by (D.13), we conclude that $\mathbf{w}$ is parallel to $\mathbf{u}$, and thus the patch-wise uniform margin classifier is unique up to a scaling factor. Moreover, this immediately implies that

$$\begin{aligned}\mathbb{P}_{(\mathbf{x}_{\text{test}}, y_{\text{test}}) \sim \mathcal{D}}(y_{\text{test}} \cdot \langle \mathbf{w}_{\text{uniform}}, \overline{\mathbf{x}}_{\text{test}} \rangle < 0) &= \mathbb{P}_{(\mathbf{x}_{\text{test}}, y_{\text{test}}) \sim \mathcal{D}}\left[y_{\text{test}} \cdot \left\langle \mathbf{u}, \sum_{p=1}^{P} (y_{\text{test}} \cdot \mathbf{u} + \boldsymbol{\xi}_{\text{test}}^{(p)}) \right\rangle < 0 \right] \\ &= \mathbb{P}_{(\mathbf{x}_{\text{test}}, y_{\text{test}}) \sim \mathcal{D}}(\|\mathbf{u}\|_2^2 < 0) = 0.\end{aligned}$$

This proves the first result.

As for the maximum margin classifier $\mathbf{w}_{\max}$, we first denote $\overline{\boldsymbol{\xi}}_i = y_i \cdot \sum_{p=1}^{P} \boldsymbol{\xi}_i^{(p)}$. Then $\overline{\boldsymbol{\xi}}_i$, $i \in [n]$ are independent Gaussian random vectors from $N(\mathbf{0}, \sigma^2 P(\mathbf{I} - \mathbf{u}\mathbf{u}^\top / \|\mathbf{u}\|_2^2))$. Define $\boldsymbol{\Gamma} = [\overline{\boldsymbol{\xi}}_1, \overline{\boldsymbol{\xi}}_2, \ldots, \overline{\boldsymbol{\xi}}_n] \in \mathbb{R}^{d \times n}$. Then focusing on the subspace $\mathrm{span}\{\mathbf{u}\}^{\perp}$, by Corollary 5.35 in Vershynin (2010), with probability at least $1 - \exp(-\Omega(\sigma^2 P d))$ we have

$$\sigma_{\min}(\boldsymbol{\Gamma}) \geq (\sqrt{d-1} - \sqrt{n} - \sqrt{d}/10) \cdot \sigma\sqrt{P} \geq \sigma\sqrt{Pd}/10, \tag{D.14}$$

$$\sigma_{\max}(\boldsymbol{\Gamma}) \leq (\sqrt{d-1} + \sqrt{n} + \sqrt{d}/10) \cdot \sigma\sqrt{P} \leq 2\sigma\sqrt{Pd}, \tag{D.15}$$

where we use $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ to denote the smallest and largest non-zero singular values of a matrix. Note that by $d = 2n > n + 1$, $\boldsymbol{\Gamma}$ has full column rank. Let $\hat{\mathbf{w}} = \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^\top \boldsymbol{\Gamma})^{-1}\mathbf{1}$, then we have $\hat{\mathbf{w}} \in \mathrm{span}\{\mathbf{u}\}^{\perp}$ and $\boldsymbol{\Gamma}^\top \hat{\mathbf{w}} = \mathbf{1}$. Therefore

$$\langle \hat{\mathbf{w}}, \overline{\boldsymbol{\xi}}_i \rangle = \langle \hat{\mathbf{w}}, P \cdot y_i \cdot \mathbf{u} + \overline{\boldsymbol{\xi}}_i \rangle = 0 + \langle \hat{\mathbf{w}}, \overline{\boldsymbol{\xi}}_i \rangle = \mathbf{e}_i^\top \boldsymbol{\Gamma}^\top \hat{\mathbf{w}} = 1$$

for all $i \in [n]$. This implies that $\hat{\mathbf{w}}$ is a feasible solution to the maximum margin problem (3.1). Therefore by the optimality of $\mathbf{w}_{\max}$, we have

$$\langle \mathbf{w}_{\max}, \mathbf{u} \rangle \cdot \|\mathbf{u}\|_2^{-1} \leq \|\mathbf{w}_{\max}\|_2 \leq \|\hat{\mathbf{w}}\|_2 = \|\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^\top \boldsymbol{\Gamma})^{-1}\mathbf{1}\|_2 = \sqrt{\mathbf{1}^\top (\boldsymbol{\Gamma}^\top \boldsymbol{\Gamma})^{-1} \mathbf{1}} \leq n \cdot \sigma_{\min}^{-1}(\boldsymbol{\Gamma}).$$

Then by (D.14), we have

$$\langle \mathbf{w}_{\max}, \mathbf{u} \rangle \leq \|\mathbf{u}\|_2 \cdot n \cdot \sigma_{\min}^{-1}(\boldsymbol{\Gamma}) \leq 10\|\mathbf{u}\|_2 \cdot nP^{-1/2} d^{-1/2} \sigma^{-1}.$$

Now by the assumption that $\sigma \geq 20\|\mathbf{u}\|_2 \cdot nP^{1/2}d^{-1/2}$, we have

$$\langle \mathbf{w}_{\max}, \mathbf{u} \rangle \leq 1/(2P), \tag{D.16}$$

and therefore

$$\langle \mathbf{w}_{\max}, \overline{\boldsymbol{\xi}}_i \rangle \geq 1 - \langle \mathbf{w}_{\max}, P \cdot \mathbf{u} \rangle \geq 1/2. \tag{D.17}$$

Further note that (3.1) indicates that $\mathbf{w}_{\max} \in \mathrm{span}\{\mathbf{u}, \overline{\boldsymbol{\xi}}_1, \overline{\boldsymbol{\xi}}_2, \ldots, \overline{\boldsymbol{\xi}}_n\}$. Therefore we have

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{u}\mathbf{u}^\top/\|\mathbf{u}\|_2^2)\mathbf{w}_{\max}\|_2 &= \|\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^\top\mathbf{w}_{\max}\|_2 \\
&= \sqrt{(\boldsymbol{\Gamma}^\top\mathbf{w}_{\max})^\top(\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma})^{-1}(\boldsymbol{\Gamma}^\top\mathbf{w}_{\max})} \\
&\geq \|\boldsymbol{\Gamma}^\top\mathbf{w}_{\max}\|_2 \cdot \sigma_{\max}^{-1}(\boldsymbol{\Gamma}) \\
&\geq \sqrt{n} \cdot (1/2) \cdot (2\sigma\sqrt{Pd})^{-1} \\
&\geq \sigma^{-1}P^{-1/2}/8,
\end{aligned} \tag{D.18}$$

where the second inequality follows by (D.17) and (D.15). Now for a new test data point $(\mathbf{x}_{\text{test}}, y_{\text{test}})$, we have $\overline{\mathbf{x}}_{\text{test}} = P \cdot y_{\text{test}} \cdot \mathbf{u} + \overline{\boldsymbol{\xi}}_{\text{test}}$, where $\overline{\boldsymbol{\xi}}_{\text{test}} = y_{\text{test}} \cdot \sum_{p=1}^{P} \boldsymbol{\xi}_{\text{test}}^{(p)} \sim N(\mathbf{0}, \sigma^2 P(\mathbf{I} - \mathbf{u}\mathbf{u}^\top/\|\mathbf{u}\|_2^2))$. Then by (D.16), we have

$$\langle \mathbf{w}_{\max}, P\mathbf{u} \rangle \leq 1/2.$$

Moreover, by (D.18) we also have

$$\langle \mathbf{w}_{\max}, \overline{\boldsymbol{\xi}}_{\text{test}} \rangle \sim N(0, \overline{\sigma}_{\text{test}}^2), \ \overline{\sigma}_{\text{test}} \geq \|(\mathbf{I} - \mathbf{u}\mathbf{u}^\top/\|\mathbf{u}\|_2^2)\mathbf{w}_{\max}\|_2 \cdot \sigma\sqrt{P} \geq 1/8.$$

Therefore

$$\mathbb{P}_{(\mathbf{x}_{\text{test}}, y_{\text{test}}) \sim \mathcal{D}}(y_{\text{test}} \cdot \langle \mathbf{w}_{\max}, \overline{\mathbf{x}}_{\text{test}} \rangle < 0) = \mathbb{P}_{(\mathbf{x}_{\text{test}}, y_{\text{test}}) \sim \mathcal{D}}(\langle \mathbf{w}_{\max}, P\mathbf{u} \rangle + \langle \mathbf{w}_{\max}, \overline{\boldsymbol{\xi}}_{\text{test}} \rangle < 0) = \Theta(1).$$

This finishes the proof. ■

D.3. Proof of Theorem 3.4

Proof [Proof of Theorem 3.4] **Proof for uniform margin solution $\mathbf{w}_{\text{uniform}}$.** Without loss of generality, we assume $\|\mathbf{w}_{\text{uniform}}\|_2 = \|\mathbf{w}_{\max}\|_2 = 1$. Then according to the definition of uniform margin, we can get that for all $i \in [n]$, it holds that for all strong signal data

$$|\langle \mathbf{w}_{\text{uniform}}, \mathbf{u} \rangle| = |\langle \mathbf{w}_{\text{uniform}}, \mathbf{v} \rangle| = |\langle \mathbf{w}_{\text{uniform}}, \boldsymbol{\xi}_i \rangle|; \tag{D.19}$$

and for all weak signal data:

$$|\langle \mathbf{w}_{\text{uniform}}, \mathbf{u} \rangle| = |\langle \mathbf{w}_{\text{uniform}}, \mathbf{v} \rangle| = |\langle \mathbf{w}_{\text{uniform}}, \boldsymbol{\xi}_i + \alpha\zeta_i\mathbf{u} \rangle|. \tag{D.20}$$

Then note that $\mathbf{w}_{\text{uniform}}$ lies in the span of $\{\mathbf{u}, \mathbf{v}\} \cup \{\boldsymbol{\xi}_i\}_{i=1,\ldots,n}$, we can get,

$$\left|\left\langle \mathbf{w}_{\text{uniform}}, \frac{\mathbf{u}}{\|\mathbf{u}\|_2} \right\rangle\right|^2 + \left|\left\langle \mathbf{w}_{\text{uniform}}, \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right\rangle\right|^2 + \sum_{i=1}^{n} \left|\left\langle \mathbf{w}_{\text{uniform}}, \frac{\boldsymbol{\xi}_i + \alpha\zeta_i\mathbf{u}}{\|\boldsymbol{\xi}_i + \alpha\zeta_i\mathbf{u}\|_2} \right\rangle\right|^2 \geq \|\mathbf{w}_{\text{uniform}}\|_2^2 = 1,$$

where we slightly abuse the notation by setting $\zeta_i = 0$ if $(\mathbf{x}_i, y_i)$ is a strong signal data. Then use the fact that $\|\mathbf{u}\|_2 = 1$, $\|\mathbf{v}\|_2 = \alpha^2$, and $\sigma = d^{-1/2}$, we can get that with probability at least $1 - \exp(-\Omega(d))$ with respect to the randomness of training data,

$$|\langle \mathbf{w}_{\text{uniform}}, \mathbf{u}\rangle|^2 + \alpha^{-4}|\langle \mathbf{w}_{\text{uniform}}, \mathbf{v}\rangle|^2 + \sum_{i=1}^{n} |\langle \mathbf{w}_{\text{uniform}}, \boldsymbol{\xi}_i + \alpha\zeta_i \mathbf{u}\rangle|^2 \geq c$$

for some absolute positive constant c. Then by (D.19) and (D.20) and using the fact that $\alpha = n^{-1/4}$, we can immediately get that

$$\langle \mathbf{w}_{\text{uniform}}, \mathbf{u}\rangle = \langle \mathbf{w}_{\text{uniform}}, \mathbf{v}\rangle = \Omega(n^{-1/2}).$$

We will then move on to the test phase. Consider a new strong signal data $(\mathbf{x}, y)$ with $\mathbf{x} = [\mathbf{u}, \boldsymbol{\xi}]$ and $y = 1$, we have

$$\langle \mathbf{w}_{\text{uniform}}, \mathbf{x}^{(1)}\rangle + \langle \mathbf{w}_{\text{uniform}}, \mathbf{x}^{(2)}\rangle = \langle \mathbf{w}_{\text{uniform}}, \mathbf{u}\rangle + \langle \mathbf{w}_{\text{uniform}}, \boldsymbol{\xi}\rangle = \Omega(n^{-1/2}) + \xi,$$

where ξ is an independent Gaussian random variable with variance smaller than σ^2. Additionally, given a weak signal data $(\mathbf{x}, y)$ with $\mathbf{x} = [\mathbf{v}, \boldsymbol{\xi}]$ and $y = 1$, we have

$$\begin{aligned}\langle \mathbf{w}_{\text{uniform}}, \mathbf{x}^{(1)}\rangle + \langle \mathbf{w}_{\text{uniform}}, \mathbf{x}^{(2)}\rangle &= \langle \mathbf{w}_{\text{uniform}}, \mathbf{v}\rangle + \langle \mathbf{w}_{\text{uniform}}, \boldsymbol{\xi} + \alpha\zeta\mathbf{u}\rangle \\ &= \langle \mathbf{w}_{\text{uniform}}, \mathbf{v}\rangle \cdot (1 + \alpha\zeta) + \xi \\ &= \Omega(n^{-1/2}) + \xi,\end{aligned}$$

where the second equation is due to $\langle \mathbf{w}_{\text{uniform}}, \mathbf{u}\rangle = \langle \mathbf{w}_{\text{uniform}}, \mathbf{v}\rangle$ and ξ is an independent Gaussian random variable with variance smaller than σ^2. Further note that $d = \omega(n\log(n))$ and $\sigma = d^{-1/2}$, we can immediately get that with probability at most $1 - 1/\text{poly}(n) \geq 1 - 1/n^{10}$, the random variable ξ will exceed $n^{-1/2}$. This completes the proof for the uniform margin solution.

Proof for maximum margin solution $\mathbf{w}_{\max}$. For the maximum margin solution, we consider

$$\mathbf{w}_{\max} = \arg\min_{\mathbf{w}} \|\mathbf{w}\|_2 \quad \text{s.t. } y_i \mathbf{w}^\top [\mathbf{x}_i^{(0)} + \mathbf{x}_i^{(1)}] \geq 1. \tag{D.21}$$

We first prove an upper bound on the norm of $\mathbf{w}_{\max}$ as follows. Based on the above definition, the upper bound can be obtained by simply finding a $\mathbf{w}$ that satisfies the margin requirements. Therefore, let $\mathbf{z}_i = y_i \cdot [\mathbf{x}_i^{(0)} + \mathbf{x}_i^{(1)}]$, we consider a candidate solution $\hat{\mathbf{w}}$ that satisfies

$$\begin{aligned}&\text{for all strong signal data:} \quad \hat{\mathbf{w}}^\top \mathbf{u} = 1, y_i \hat{\mathbf{w}}^\top \boldsymbol{\xi}_i = 0; \\ &\text{for all weak signal data:} \quad \hat{\mathbf{w}}^\top \mathbf{v} = 0, \mathbf{w}^\top y_i(\boldsymbol{\xi}_i + \alpha\zeta_i \mathbf{u}) = 1\end{aligned}$$

Then, let $\mathbf{P}_{\mathcal{E}^c}$ be the projection on the subspace that is orthogonal to all noise vectors of the strong signal data. Then the above condition for weak signal data requires

$$y_i \langle \hat{\mathbf{w}}, \mathbf{P}_{\mathcal{E}^c} \boldsymbol{\xi}_i\rangle = 1 - \alpha y_i \zeta_i,$$

where we use the fact that $\langle \hat{\mathbf{w}}, \mathbf{u}\rangle = 1$. Let n_1 and n_2 be the numbers of strong signal data and weak signal data respectively, which clearly satisfy $n_1 = \Theta(n)$ and $n_2 = \Theta(\rho n) = \Theta(n^{1/4})$ with probability at least $1 - \exp(-\Omega(n^{1/4}))$. Define $\mathbf{r} \in \mathbb{R}^{n_2 \times 1}$ be the collection of $y_i(1 - \alpha y_i \zeta_i)$ for

all weak signal data, let $\boldsymbol{\omega}_i = \mathbf{P}_{\mathcal{E}^c}\boldsymbol{\xi}_i$ and set $\boldsymbol{\Omega} \in \mathbb{R}^{n_2 \times d}$ as the collection of $\boldsymbol{\omega}_i$'s. Then the above equation can be rewritten as $\boldsymbol{\Omega}\hat{\mathbf{w}} = \mathbf{r}$, which leads to a feasible solution

$$\hat{\mathbf{w}} = \mathbf{u} + \boldsymbol{\Omega}^\top(\boldsymbol{\Omega}\boldsymbol{\Omega}^\top)^{-1}\mathbf{r}.$$

Then note that the noise vectors for all data points are independent, conditioning on $\mathbf{P}_{\mathcal{E}^c}$, the random vector $\omega_i = \mathbf{P}_{\mathcal{E}^c}\boldsymbol{\xi}_i$ can be still regarded as a Gaussian random vector in $d - n_1 - 2$ dimensional space. Then by standard random matrix theory, we can obtain that we can get that with probability at least $1 - \exp(-\Omega(d))$ with respect to the randomness of training data, $\lambda_{\min}(\boldsymbol{\Omega}\boldsymbol{\Omega}^\top), \lambda_{\min}(\boldsymbol{\Omega}\boldsymbol{\Omega}^\top) = \Theta\big(\sigma^2 \cdot (d - n_1 - 2)\big) = \Theta(1)$. Further note that $\|\mathbf{r}\|_2 = \Theta(n_2^{1/2})$, we can finally obtain that

$$\|\mathbf{w}_{\max}\|_2^2 \leq \|\hat{\mathbf{w}}\|_2^2 = 1 + \big\|\boldsymbol{\Omega}^\top(\boldsymbol{\Omega}\boldsymbol{\Omega}^\top)^{-1}\mathbf{r}\big\|_2^2 = \Theta(n_2) = \Theta(n^{1/4}). \tag{D.22}$$

Therefore, we can further get

$$\langle \mathbf{w}_{\max}, \mathbf{v}\rangle \leq \|\mathbf{w}_{\max}\|_2 \cdot \|\mathbf{v}\|_2 = O\big(\alpha^2 n^{1/8}\big) = O(n^{-7/8}).$$

Next, we will show that $\langle \mathbf{w}_{\max}, \mathbf{u}\rangle > 1/2$. In particular, under the margin condition in (D.21), we have for all strong signal data

$$\langle \mathbf{w}_{\max}, \mathbf{u}\rangle + \langle \mathbf{w}_{\max}, y_i\boldsymbol{\xi}_i\rangle \geq 1.$$

Then if $\langle \mathbf{w}_{\max}, \mathbf{u}\rangle \leq 1/2$, we will then get $\langle \mathbf{w}_{\max}, y_i\boldsymbol{\xi}_i\rangle \geq 1/2$ for all strong signal data. Let $\mathcal{S}_{\mathbf{u}}$ be the collection of indices of strong signal data, we further have

$$\Big\langle \mathbf{w}_{\max}, \sum_{i\in\mathcal{S}_{\mathbf{u}}} y_i\boldsymbol{\xi}_i \Big\rangle \geq n_1.$$

Let $\boldsymbol{\xi}' = \sum_{i\in\mathcal{S}_{\mathbf{u}}} y_i\boldsymbol{\xi}_i$, it can be seen that $\boldsymbol{\xi}'$ is also a Gaussian random vector with covariance matrix $n_1\sigma^2\big(\mathbf{I} - \mathbf{u}\mathbf{u}^\top/\|\mathbf{u}\|_2^2 - \mathbf{v}\mathbf{v}^\top/\|\mathbf{v}\|_2^2\big)$. This implies that conditioning on $\mathcal{S}_{\mathbf{u}}$, with probability at least $1 - \exp(-\Omega(d))$, it holds that $\|\boldsymbol{\xi}'\|_2 = \Theta(n_1^{1/2})$. Then using the fact that $n_1 = \Theta(n)$, it further suggests that

$$\|\mathbf{w}_{\max}\|_2 \geq \frac{n_1}{\|\boldsymbol{\xi}'\|_2} = \Theta(n^{1/2}),$$

which contradicts the upper bound on the norm of maximum margin solution we have proved in (D.22). Therefore we must have $\langle \mathbf{w}_{\max}, \mathbf{u}\rangle > 1/2$.

Finally, we are ready to evaluate the test error of $\mathbf{w}_{\max}$. In particular, since we are proving a lower bound on the test error, we will only consider the weak signal data while assuming all strong feature data can be correctly classified. Consider a weak signal data $(\mathbf{x}, y)$ with with $\mathbf{x} = [\mathbf{v}, \boldsymbol{\xi}]$ and $y = 1$, we have

$$\begin{aligned}\langle \mathbf{w}_{\max}, \mathbf{x}^{(1)}\rangle + \langle \mathbf{w}_{\max}, \mathbf{x}^{(2)}\rangle &= \langle \mathbf{w}_{\max}, \mathbf{v}\rangle + \langle \mathbf{w}_{\max}, \boldsymbol{\xi} + \alpha\zeta\mathbf{u}\rangle \\ &= \langle \mathbf{w}_{\max}, \mathbf{v}\rangle + \zeta \cdot \alpha\langle \mathbf{w}_{\max}, \mathbf{u}\rangle + \|\mathbf{w}_{\max}\|_2 \cdot \xi,\end{aligned}$$

where ξ is a random variable with variance smaller than σ_0. Note that $\zeta = -1$ with half probability. In this case, using our previous results on $\langle \mathbf{w}_{\max}, \mathbf{v}\rangle$ and $\langle \mathbf{w}_{\max}, \mathbf{u}\rangle$, and the fact that ξ is independent of $\mathbf{v}$ and $\mathbf{u}$, we can get with probability at least $1/2$,

$$\langle \mathbf{w}_{\max}, \mathbf{x}^{(1)}\rangle + \langle \mathbf{w}_{\max}, \mathbf{x}^{(2)}\rangle \leq \Theta(n^{-7/8}) - \Theta(n^{-1/2}) < 0$$

which will lead to an incorrect prediction. This implies that for weak signal data, the population prediction error will be at least $1/4$. Noting that the weak signal data will appear with probability ρ, combining them will be able to complete the proof. ■

Appendix E. Proofs of Technical Lemmas

E.1. Proof of Lemma C.1

Proof [Proof of Lemma C.1] First, it is easy to see that

$$\sum_{i,i'=1}^{n} \max\{b_i, b_{i'}\} \cdot (a_i - a_{i'})^2 = 2\sum_{i=1}^{n}\sum_{i'>i} b_i \cdot (a_{i'} - a_i)^2. \tag{E.1}$$

Then, we can also observe that for any $i < j$, we have

$$\sum_{i'>i}(a_{i'} - a_i)^2 \geq \sum_{i'>j}(a_{i'} - a_i)^2 \geq \sum_{i'>j}(a_{i'} - a_j)^2, \tag{E.2}$$

where we use the fact that $a_i \leq a_j$. Then let $\bar{b} = \frac{\sum_{i=1}^{n} b_i}{n}$ and k^* be the index satisfying $b_{k^*-1} \geq \bar{b}/2 > b_{k^*}$, we can immediately get that

$$\sum_{i \geq k^*} b_i \leq \frac{\sum_{i=1}^{n} b_i}{2n} \cdot n = \frac{\sum_{i=1}^{n} b_i}{2}, \tag{E.3}$$

which further implies that $\sum_{i<k^*} b_i \geq n\bar{b}/2$. Then by (E.2), we can get

$$\frac{\bar{b}}{2} \cdot \sum_{i \geq k^*}\sum_{i'>i}(a_{i'} - a_i)^2 \leq \frac{\bar{b}}{2} \cdot \sum_{i \geq k^*}\sum_{i'>k^*}(a_{i'} - a_{k^*})^2 \leq \frac{n\bar{b}}{2} \cdot \sum_{i'>k^*}(a_{i'} - a_{k^*})^2.$$

Besides, we also have

$$\sum_{i<k^*}\sum_{i'>i} b_i \cdot (a_{i'} - a_i)^2 \geq \left(\sum_{i<k^*} b_i\right) \cdot \sum_{i'>k^*}(a_{i'} - a_{k^*})^2 \geq \frac{n\bar{b}}{2} \cdot \sum_{i'>k^*}(a_{i'} - a_{k^*})^2,$$

which immediately implies that

$$\sum_{i<k^*}\sum_{i'>i} b_i \cdot (a_{i'} - a_i)^2 \geq \frac{\bar{b}}{2} \cdot \sum_{i \geq k^*}\sum_{i'>i}(a_{i'} - a_i)^2.$$

Therefore, we can get

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{i'>i} b_i \cdot (a_{i'} - a_i)^2 &\geq \sum_{i<k^*}\sum_{i'>i} b_i \cdot (a_{i'} - a_i)^2 \\
&\geq \sum_{i<k^*}\sum_{i'>i} \frac{b_i}{2} \cdot (a_{i'} - a_i)^2 + \frac{\bar{b}}{4} \cdot \sum_{i \geq k^*}\sum_{i'>i}(a_{i'} - a_i)^2 \\
&\geq \frac{\bar{b}}{4}\sum_{i=1}^{n}\sum_{i'>i}(a_{i'} - a_i)^2,
\end{aligned}$$

where we use the fact that $b_i \geq \bar{b}/2$ for all $i < k^*$. Finally, putting the above inequality into (E.1) and applying the definition of $\bar{b}$, we are able to complete the proof. ■

E.2. Proof of Lemma C.2

Proof [Proof of Lemma C.2] We first show the lower bound of $a^{(t)}$. Consider a continuous-time sequence $\underline{a}^{(t)}$, $t \geq 0$ defined by the integral equation

$$\underline{a}^{(t)} = \underline{a}^{(0)} + c \cdot \int_0^t \exp(-\underline{a}^{(\tau)}) \mathrm{d}\tau, \quad \underline{a}^{(0)} = a^{(0)}. \tag{E.4}$$

Note that $\underline{a}^{(t)}$ is obviously an increasing function of t. Therefore we have

$$\begin{aligned}
\underline{a}^{(t+1)} &= \underline{a}^{(t)} + c \cdot \int_t^{t+1} \exp(-\underline{a}^{(\tau)}) \mathrm{d}\tau \\
&\leq \underline{a}^{(t)} + c \cdot \int_t^{t+1} \exp(-\underline{a}^{(t)}) \mathrm{d}\tau \\
&= \underline{a}^{(t)} + c \cdot \exp(-\underline{a}^{(t)})
\end{aligned}$$

for all $t \in \mathbb{N}$. Comparing the above inequality with the iterative formula $a^{(t+1)} = a^{(t)} + c \cdot \exp(-a^{(t)})$, we conclude by the comparison theorem that $a^{(t)} \geq \underline{a}^{(t)}$ for all $t \in \mathbb{N}$. Note that (E.4) has an exact solution

$$\underline{a}^{(t)} = \log(c \cdot t + \exp(a^{(0)})).$$

Therefore we have

$$a^{(t)} \geq \log(c \cdot t + \exp(a^{(0)}))$$

for all $t \in \mathbb{N}$, which completes the first part of the proof. Now for the upper bound of $a^{(t)}$, we have

$$\begin{aligned}
a^{(t)} &= a^{(0)} + c \cdot \sum_{\tau=0}^{t} \exp(-a^{(\tau)}) \\
&\leq a^{(0)} + c \cdot \sum_{\tau=0}^{t} \exp[-\log(c \cdot \tau + \exp(a^{(0)}))] \\
&= a^{(0)} + c \cdot \sum_{\tau=0}^{t} \frac{1}{c \cdot \tau + \exp(a^{(0)})} \\
&= a^{(0)} + \frac{c}{\exp(a^{(0)})} + c \cdot \sum_{\tau=1}^{t} \frac{1}{c \cdot \tau + \exp(a^{(0)})} \\
&\leq a^{(0)} + \frac{c}{\exp(a^{(0)})} + c \cdot \int_0^t \frac{1}{c \cdot \tau + \exp(a^{(0)})} \mathrm{d}\tau,
\end{aligned}$$

where the second inequality follows by the lower bound of $a^{(t)}$ as the first part of the result of this lemma. Therefore we have

$$\begin{aligned}
a^{(t)} &\leq a^{(0)} + \frac{c}{\exp(a^{(0)})} + \log(c \cdot t + \exp(a^{(0)})) - \log(\exp(a^{(0)})) \\
&= c \exp(-a^{(0)}) + \log(c \cdot t + \exp(a^{(0)})).
\end{aligned}$$

This finishes the proof. ■