How Does the ReLU Activation Affect the Implicit Bias of Gradient Descent on High-dimensional Neural Network Regression?

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Overparameterized ML models, including neural networks, typically induce underdetermined training objectives with multiple global minima. The *implicit bias* refers to the limiting global minimum that is attained by a common optimization algorithm, such as gradient descent (GD). In this paper, we characterize the implicit bias of GD for training a shallow ReLU model with the squared loss on high-dimensional random features. Prior work [15] showed that the implicit bias does not exist in the worst-case, or corresponds exactly to the minimum-ℓ_2-norm interpolating solution under *exactly* orthogonal data [3]. Our work interpolates between these two extremes and shows that, for sufficiently high-dimensional random data, the implicit bias approximates the minimum-ℓ_2-norm solution with high probability with a gap on the order $\Theta\left(\sqrt{\frac{n}{d}}\right)$, where n is the number of data and d is the feature dimension. Our results are obtained through a novel primal-dual analysis that carefully tracks the evolution of predictions, data-span coefficients, as well as their interactions, and show that the ReLU activation pattern quickly stabilizes with high probability over random data.

1. Introduction

Modern machine learning objectives are often *underdetermined*, admitting infinitely many global minima. Nevertheless, empirical evidence suggests that gradient descent (GD) converges to solutions with strong generalization properties even without explicit regularization [13, 17]. This phenomenon, termed implicit bias [11, 14], can be treated as a critical factor of scaling laws; it ensures that as model capacity increases, the optimizer consistently selects high quality solutions [9]. Early research focused on linear models to characterize this bias. In linear classification, GD applied to exponentially-tailed losses converges in direction to the max-margin solution [11, 14]. For linear regression with squared loss, GD converges to the interpolation with the minimum ℓ_2-norm [6].

Understanding the implicit bias in nonlinear models such as neural networks remains a significant challenge, primarily due to the induced non-convexity of the optimization objective. In this work, we focus on regression with a one-hidden-layer ReLU neural network and the squared loss. Remarkably, Vardi and Shamir [15] showed that establishing an implicit bias is known to be hard in the worst case, even when the model consists only of a single neuron and assuming global convergence. At the other extreme, Boursier et al. [3] showed that the implicit bias of gradient flow for *exactly* orthogonal features is exactly the minimum-ℓ_2-norm solution. Notably, high-dimensional random features are *near*-orthogonal, raising the question of whether the implicit bias can be characterized in this more realistic but also more challenging case.

2. Problem Setup

We consider a regression problem with feature vector $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$ and label $y \in \mathcal{Y} \subset \mathbb{R}$. We consider random feature vectors drawn according to a distribution $\mathcal{D}$ with zero mean, i.e., $\mathbb{E}[\boldsymbol{x}] = \mathbf{0}$,

	Orthogonal data	High dimensional data	Worst-case data
ReLU models $h(\boldsymbol{x}) := \sum_{k=1}^{m} s_k \sigma(\boldsymbol{w}_k^\top \boldsymbol{x})$	Implicit bias characterization [3]	Global convergence only [5] This work: Implicit bias characterization	No implicit bias in general [15]
Linear models $h(\boldsymbol{x}) := \boldsymbol{w}^\top \boldsymbol{x}$	Implicit bias coincides with maximum margin SVM [10]		$\boldsymbol{w}^{(\infty)} = \underset{\boldsymbol{w}:\boldsymbol{X}\boldsymbol{w}=\boldsymbol{y}}{\arg\min} \left\Vert \boldsymbol{w} - \boldsymbol{w}^{(0)} \right\Vert_2$ [6]

Table 1: Our results contextualized with related literature.

and covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$. Let $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top \in \mathbb{R}^{d\times d}$ be the eigendecomposition of the covariance matrix, where $\boldsymbol{V} \in \mathbb{R}^{d\times d}$ is the matrix of eigenvectors and $\boldsymbol{\Lambda} \in \mathbb{R}^{d\times d}$ is a diagonal matrix whose entries are the eigenvalues of $\boldsymbol{\Sigma}$ in descending order. We make the mild assumption that $\boldsymbol{x} = \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{z}$ where $\boldsymbol{z} \in \mathbb{R}^d$ has independent, mean-zero, 1-subgaussian components. We observe a dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ where the features $\{\boldsymbol{x}_i\}_{i=1}^n$ are drawn i.i.d. from the distribution $\mathcal{D}$. We denote the data matrix by $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and the label vector by $\boldsymbol{y} \in \mathbb{R}^n$. Since we operate in a high-dimensional regime ($d > n$), we make the mild assumption that $\operatorname{rank}(\boldsymbol{X}) = n$.

Next, we assume that the magnitudes of the labels are bounded away from zero and infinity.

Assumption 1 (Bounded Labels) *For all $i \in [n]$, $y_{\min} \le |y_i| \le y_{\max}$ for some $y_{\min}, y_{\max} \in \mathbb{R}_+$.*

We next impose a high-dimensional assumption. We define effective dimensions based on the spectrum $\boldsymbol{\lambda} := [\lambda_1, \cdots, \lambda_d]^\top$ of the covariance matrix $\boldsymbol{\Sigma}$, by $d_2 := \|\boldsymbol{\lambda}\|_1^2 \,/\, \|\boldsymbol{\lambda}\|_2^2$, $d_\infty := \|\boldsymbol{\lambda}\|_1 \,/\, \|\boldsymbol{\lambda}\|_\infty$.

Assumption 2 (High-dim Features) $d_2 \ge C_0^2 \frac{n^2 y_{\max}^2}{y_{\min}^2}$ *and* $d_\infty \ge C_0 \frac{n^{1.5} y_{\max}}{y_{\min}}$ *for a large* $C_0 > 1$.

General ReLU Models and Empirical Risk Minimization. We denote by $h_{\boldsymbol{\Theta}} : \mathcal{X} \to \mathcal{Y}$ the general ReLU model used for the regression task, defined as $h_{\boldsymbol{\Theta}}(\boldsymbol{x}) := \sum_{k=1}^{m} s_k \sigma(\boldsymbol{w}_k^\top \boldsymbol{x})$, where $\boldsymbol{\Theta}$ denotes the collection of model parameters $\{\boldsymbol{w}_k\}_{k=1}^m$ and $\{s_k\}_{k=1}^m$. Here, $s_k \in \{-1, +1\}$ denotes the sign of the k-th neuron, $\sigma(z) := \max\{z, 0\}$ is the ReLU activation, $\boldsymbol{w}_k \in \mathbb{R}^d$ is its weight, and $m \ge 1$ is the number of neurons. We minimize the empirical risk under the squared loss

$$\mathcal{R}(\boldsymbol{\Theta}) = \frac{1}{2}\sum_{i=1}^{n}\big(h_{\boldsymbol{\Theta}}(\boldsymbol{x}_i) - y_i\big)^2 = \frac{1}{2}\left\Vert h_{\boldsymbol{\Theta}}(\boldsymbol{X}) - \boldsymbol{y}\right\Vert_2^2, \tag{1}$$

where we define $h_{\boldsymbol{\Theta}}$ as $h_{\boldsymbol{\Theta}}(\boldsymbol{X}) := [h_{\boldsymbol{\Theta}}(\boldsymbol{x}_1), \cdots, h_{\boldsymbol{\Theta}}(\boldsymbol{x}_n)]^\top \in \mathbb{R}^n$. We employ the GD to minimize (1), and we only update the neuron weights $\{\boldsymbol{w}_k\}_{k=1}^m$ and fix the signs of the neurons $\{s_k\}_{k=1}^m$.

Gradient Descent and Primal-dual Representation. The gradient of the empirical risk in (1) toward $\boldsymbol{w}_k$ is $\nabla_{\boldsymbol{w}_k}\mathcal{R}(\boldsymbol{\Theta}) = s_k \boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}_k)\big(h_{\boldsymbol{\Theta}}(\boldsymbol{X}) - \boldsymbol{y}\big)$, where $\boldsymbol{D}(\boldsymbol{z}) : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ denotes the diagonal matrix with entries $D_{ii} := \mathbb{1}_{z_i > 0}$. Accordingly, the GD update for $\boldsymbol{w}_k$ takes the form

$$\boldsymbol{w}_k^{(t+1)} = \boldsymbol{w}_k^{(t)} - \eta\nabla_{\boldsymbol{w}_k}\mathcal{R}(\boldsymbol{\Theta}^{(t)}) = \boldsymbol{w}_k^{(t)} - \eta s_k \boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}_k^{(t)})\big(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}\big). \tag{2}$$

To analyze these updates, we introduce a primal–dual representation used in mirror descent [12].

$$\boldsymbol{\beta}_k := \boldsymbol{X}\boldsymbol{w}_k, \quad \boldsymbol{\alpha}_k := \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{X}\boldsymbol{w}_k, \text{ with } \boldsymbol{\beta}_k = \boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{\alpha}_k \quad \forall k \in [m]. \tag{3}$$

This representation restricts attention to the components of $\boldsymbol{w}_k$ that lie in the span of the data matrix $\boldsymbol{X}$.[1] Under this parameterization, the gradient descent update (2) in primal–dual form becomes

(Primal) $$\boldsymbol{\beta}_k^{(t+1)} = \boldsymbol{\beta}_k^{(t)} - \eta s_k \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\beta}_k^{(t)})(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}), \tag{4a}$$

(Dual) $$\boldsymbol{\alpha}_k^{(t+1)} = \boldsymbol{\alpha}_k^{(t)} - \eta s_k \boldsymbol{D}(\boldsymbol{\beta}_k^{(t)})(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}). \tag{4b}$$

1. In general, $\boldsymbol{w}_k$ may contain components orthogonal to $\operatorname{span}(\{\boldsymbol{x}_i\}_{i=1}^n)$. However, since the gradient updates act only within $\operatorname{span}(\{\boldsymbol{x}_i\}_{i=1}^n)$, the components not in the span remain unchanged throughout training.

The sign of $\beta_{k,i}^{(t)}$ determines whether the corresponding $\alpha_{k,i}^{(t+1)}$ is updated through $\boldsymbol{D}(\boldsymbol{\beta}_k^{(t)})$. The sign of $\boldsymbol{\beta}_k^{(t)}$ and the dynamics of $\boldsymbol{\alpha}_k^{(t)}$ is key to characterizing the behavior and implicit bias of GD.

Minimum-ℓ_2-norm Solution. For linear regression with zero initialization, GD converges to the minimum-ℓ_2-norm interpolation: $\boldsymbol{w}_{\text{linear-MNI}} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y} = \arg\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|_2^2$, s.t. $\boldsymbol{w}^\top\boldsymbol{x}_i = y_i,\ \forall i \in [n]$. Motivated by this, we consider the minimum-ℓ_2-norm for ReLU regression problem

$$\{\boldsymbol{w}_k^\star\}_{k=1}^m = \arg\min_{\{\boldsymbol{w}_k\}_{k=1}^m} \frac{1}{2}\sum_{k=1}^m \|\boldsymbol{w}_k\|_2^2 \qquad \text{s.t.} \sum_{k=1}^m s_k\sigma(\boldsymbol{w}_k^\top\boldsymbol{x}_i) = y_i,\ \text{for all } i \in [n]. \tag{5}$$

3. Implicit Bias of Single ReLU Models ($m = 1$) Under Gradient Descent

We begin by analyzing the single positive ReLU neuron model $h_{\boldsymbol{\Theta}}(\boldsymbol{x}) := \sigma(\boldsymbol{w}^\top\boldsymbol{x})$.

3.1. Gradient Descent Updates and Convergence

For the single ReLU model, the GD update in (2) simplifies to $\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta\nabla_{\boldsymbol{w}}\mathcal{R}(\boldsymbol{w}^{(t)}) = \boldsymbol{w}^{(t)} - \eta\boldsymbol{X}^\top\boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}^{(t)})(\boldsymbol{X}\boldsymbol{w}^{(t)} - \boldsymbol{y})$. Compared to linear regression, the difference is the diagonal matrix $\boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}^{(t)})$. This matrix selects a subset of examples to contribute to each GD update.

3.1.1. SUFFICIENT CONDITIONS FOR GRADIENT DESCENT CONVERGENCE

In general, the optimization trajectory and loss landscape induced by gradient descent even on a single ReLU model are difficult to characterize. Suppose there exists $t_0 \geq 0$ that the activation pattern remains unchanged after t_0. We formalize this observation in the following lemma.

Lemma 1 *Suppose there exists $t_0 \geq 0$ such that $\boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}^{(t_0)}) = \boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}^{(t)})$ for all $t \geq t_0$. Define the subset of examples $S := \{i \in [n] : \boldsymbol{x}_i^\top\boldsymbol{w}^{(t_0)} > 0\}$. Then, for all $t \geq t_0$, the gradient descent dynamics of the single ReLU model are equivalent to gradient descent applied to a linear model initialized at $\boldsymbol{w}^{(t_0)}$ and trained only on the subset S. (Lemma 1 is proved in Appendix D.1)*

Convergence of Lemma 1 follows from classical convergence guarantees for linear regression.

Lemma 2 *Suppose the effective dimension $d_\infty \geq cn$, for some constant $c \geq 1$, the step size satisfies $\eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$, and there exists $t_0 \geq 0$ such that $\boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}^{(t_0)}) = \boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}^{(t)})$ for all $t \geq t_0$. Then, gradient descent applied to the single ReLU model converges to $\boldsymbol{w}^{(\infty)} = \arg\min_{\boldsymbol{w}:\boldsymbol{X}_S\boldsymbol{w}=\boldsymbol{y}_S} \left\|\boldsymbol{w} - \boldsymbol{w}^{(t_0)}\right\|_2$ with probability at least $1 - 2e^{-n/C_g}$, where $S := \{i \in [n] : \boldsymbol{x}_i^\top\boldsymbol{w}^{(t_0)} > 0\}$. (Proof in Appendix D.1)*

3.2. Minimum-ℓ_2-norm Solution of Single ReLU models

Compared to $\boldsymbol{w}_{\text{linear}-\text{MNI}}$ in Section 2, single ReLU models are restricted to non-negative outputs. We write a convex optimization problem where the constraints with nonpositive labels are relaxed

$$\boldsymbol{w}^\star = \arg\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 \text{ s.t. } \boldsymbol{w}^\top\boldsymbol{x}_i = y_i,\ \text{for all } y_i > 0,\ \boldsymbol{w}^\top\boldsymbol{x}_j \leq 0,\ \text{for all } y_j \leq 0. \tag{6}$$

We show that the solution of (6) coincides with the minimum-ℓ_2-norm solution associated with *linearly* fitting a suitable subset of training examples, after setting all negative labels to zero. We define a subset $S \subseteq [n]$ and $\boldsymbol{w}_{\text{linear}-\text{MNI},S} = \boldsymbol{X}_S^\top(\boldsymbol{X}_S\boldsymbol{X}_S^\top)^{-1}\tilde{\boldsymbol{y}}_S$, where $\tilde{\boldsymbol{y}}_S \in \mathbb{R}^{|S|}$ is the corresponding label subvector with all negative entries replaced by zero. Lemma 3 is proved in Appendix D.1.

Lemma 3 *Consider a single ReLU model $h_{\mathbf{\Theta}}(\boldsymbol{x}) = \sigma(\boldsymbol{w}^\top \boldsymbol{x})$. The minimum-$\ell_2$-norm solution $\boldsymbol{w}^\star$ of $h_{\mathbf{\Theta}}(\boldsymbol{x})$ satisfies $\boldsymbol{w}^\star = \boldsymbol{w}_{\text{linear}-\text{MNI},S}$ for some index subset $S \subseteq [n]$ that necessarily contains all indices i such that $y_i > 0$, where the corresponding labels are given by $\tilde{y}_{S,i} = \max\{y_i, 0\}$.*

Note that $\boldsymbol{w}^\star$ is a fundamentally different inductive bias from $\boldsymbol{w}_{\text{linear}-\text{MNI}}$ as the subset S does not have an explicit formula, and is training data-dependent.

3.3. High-dimensional Implicit Bias of Single ReLU Models

Our first main result characterizes the GD dynamics of single ReLU models on high-dim data.

Theorem 4 *Consider Assumptions 1 and 2, suppose the initialization is $\boldsymbol{w}^{(0)} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{\epsilon}$, where $0 < \epsilon_i \leq \frac{1}{C_\alpha} y_{\min}$ for all $i \in [n]$, and the step size to satisfy $\frac{1}{CC_g\|\boldsymbol{\lambda}\|_1} \leq \eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$. Then, the gradient descent limit $\boldsymbol{w}^{(\infty)}$ for the single ReLU model coincides with the solution obtained by linear regression trained only on the positively labeled examples with initialization $\boldsymbol{w}^{(1)} = \eta\boldsymbol{X}^\top\Big(\boldsymbol{y} - \boldsymbol{\epsilon} + \frac{1}{\eta}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{\epsilon}\Big)$ with probability at least $1 - 2\exp(-cn)$. Formally, we have $\boldsymbol{w}^{(\infty)} = \underset{\boldsymbol{w}:\boldsymbol{X}_+\boldsymbol{w}=\boldsymbol{y}_+}{\arg\min} \left\|\boldsymbol{w} - \boldsymbol{w}^{(1)}\right\|_2$ and $\boldsymbol{X}_-\boldsymbol{w}^{(\infty)} \preceq \mathbf{0}$. (Theorem 4 is proved in Appendix D.2)*

Theorem 4 characterizes a tractable regime of GD. Due to the ReLU activation, the main challenge lies in monitoring which examples are active and which are inactive during gradient descent.

3.4. Approximation to Minimum-ℓ_2-norm Solution in High Dimensions

We show the limiting solution from Theorem 4 is different from, but close to the minimum-ℓ_2-norm.

Theorem 5 *Consider Assumptions 1 and 2, suppose the initialization is $\boldsymbol{w}^{(0)} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{\epsilon}$, where $0 < \epsilon_i \leq \frac{1}{C_\alpha} y_{\min}$ for all $i \in [n]$, and the step size to satisfy $\frac{1}{CC_g\|\boldsymbol{\lambda}\|_1} \leq \eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$. Then, we have $\sqrt{\frac{n_- y_{\min}^2}{CC_g\|\boldsymbol{\lambda}\|_1}} \leq \left\|\boldsymbol{w}^{(\infty)} - \boldsymbol{w}^\star\right\|_2 \leq \sqrt{\frac{16 n_- y_{\max}^2}{C_g\|\boldsymbol{\lambda}\|_1}}$ with probability at least $1 - 2\exp(-cn)$.*

Theorem 5 is proved in Appendix D.3 and heavily leverages our characterization result in Lemma 3.

4. Implicit Bias of Two ReLU Models ($m = 2$) Under Gradient Descent

We extend our analysis to a 2-ReLU model, which combines one positive ReLU neuron and one negative ReLU neuron: $h_{\mathbf{\Theta}}(\boldsymbol{x}) = \sigma(\boldsymbol{w}_\oplus^\top\boldsymbol{x}) - \sigma(\boldsymbol{w}_\ominus^\top\boldsymbol{x})$, The GD update in (2) simplifies to $\boldsymbol{w}_\oplus^{(t+1)} = \boldsymbol{w}_\oplus^{(t)} - \eta\nabla_{\boldsymbol{w}_\oplus}\mathcal{R}(\mathbf{\Theta}^{(t)}) = \boldsymbol{w}_\oplus^{(t)} - \eta\boldsymbol{X}^\top\boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}_\oplus^{(t)})\big(h_{\mathbf{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}\big)$. $\boldsymbol{w}_\ominus$ follows the same format.

4.1. Minimum-ℓ_2-norm Solution of 2-ReLU Models

First, we characterize the minimum-ℓ_2-norm solution for the 2-ReLU model, defined below:

$$\boldsymbol{w}_\oplus^\star, \boldsymbol{w}_\ominus^\star = \underset{\boldsymbol{w}_\oplus, \boldsymbol{w}_\ominus}{\arg\min} \frac{1}{2}\|\boldsymbol{w}_\oplus\|_2^2 + \frac{1}{2}\|\boldsymbol{w}_\ominus\|_2^2 \quad \text{s.t. } \sigma(\boldsymbol{w}_\oplus^\top\boldsymbol{x}_i) - \sigma(\boldsymbol{w}_\ominus^\top\boldsymbol{x}_i) = y_i, \text{ for all } i \in [n]. \tag{7}$$

To analyze (7), we show that the optimal solution is also the optimal solution to a *restricted convex program* obtained by fixing the activation pattern across the training examples. Let $S_+ = \{i : y_i > 0, \forall i \in [n]\}$, $S_- = \{j : y_j < 0, \forall j \in [n]\}$, so that $S_+ \cup S_- = [n]$ and $S_+ \cap S_- = \varnothing$.

Lemma 6 *The feasible set of (7) is nonempty, and there exist partitions $S_1 \cup S_2 = S_+,\ S_1 \cap S_2 = \varnothing$, and $S_3 \cup S_4 = S_-,\ S_3 \cap S_4 = \varnothing$ such that the optimal solution $\{\boldsymbol{w}_\oplus^\star, \boldsymbol{w}_\ominus^\star\}$ of (7) is also an optimal solution of the following convex program: (Lemma 6 is proved in Appendix E.1.)*

$$\boldsymbol{w}_\oplus^\star, \boldsymbol{w}_\ominus^\star = \arg\min_{\boldsymbol{w}_\oplus, \boldsymbol{w}_\ominus} \frac{1}{2}\|\boldsymbol{w}_\oplus\|_2^2 + \frac{1}{2}\|\boldsymbol{w}_\ominus\|_2^2 \tag{8}$$

$$\text{s.t. } \boldsymbol{w}_\oplus^\top \boldsymbol{x}_i = y_i,\ \boldsymbol{w}_\ominus^\top \boldsymbol{x}_i \le 0, i \in S_1, \quad (\boldsymbol{w}_\oplus - \boldsymbol{w}_\ominus)^\top \boldsymbol{x}_i = y_i,\ \boldsymbol{w}_\ominus^\top \boldsymbol{x}_i \ge 0, i \in S_2,$$
$$-\boldsymbol{w}_\ominus^\top \boldsymbol{x}_i = y_i,\ \boldsymbol{w}_\oplus^\top \boldsymbol{x}_i \le 0, i \in S_3, \quad (\boldsymbol{w}_\oplus - \boldsymbol{w}_\ominus)^\top \boldsymbol{x}_i = y_i,\ \boldsymbol{w}_\oplus^\top \boldsymbol{x}_i \ge 0, i \in S_4.$$

4.2. High-dimensional Implicit Bias of 2-ReLU Models

Next, we characterize the gradient descent dynamics of two ReLU models in the high dimensions.

Theorem 7 *Consider Assumptions 1 and 2, suppose the initialization is $\boldsymbol{w}_\oplus^{(0)} = \boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{\epsilon}_\oplus$ and $\boldsymbol{w}_\ominus^{(0)} = \boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{\epsilon}_\ominus$, where $0 < \epsilon_{\oplus,i}, \epsilon_{\ominus,i} \le \frac{1}{2C_\alpha} y_{\min}$ for all $i \in [n]$, and the step size to satisfy $\frac{1}{CC_g\|\boldsymbol{\lambda}\|_1} \le \eta \le \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$. Then, with probability at least $1 - 2\exp(-cn)$, the gradient descent limit $\boldsymbol{w}_\oplus^{(\infty)}$ coincides with the solution obtained by linear regression trained only on the positively labeled examples, with the initialization $\boldsymbol{w}_\oplus^{(1)} = \eta \boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{\epsilon}_\oplus + \boldsymbol{\epsilon}_\ominus + \frac{1}{\eta}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{\epsilon}_\oplus\right)$, and $\boldsymbol{w}_\oplus^{(\infty)} = \arg\min_{\boldsymbol{w}:\boldsymbol{X}_+\boldsymbol{w}=\boldsymbol{y}_+} \left\|\boldsymbol{w} - \boldsymbol{w}_\oplus^{(1)}\right\|_2$; similarly, for $\boldsymbol{w}_\ominus^{(\infty)}$, we have $\boldsymbol{w}_\ominus^{(1)} = \eta \boldsymbol{X}^\top\left(-\boldsymbol{y} + \boldsymbol{\epsilon}_\oplus - \boldsymbol{\epsilon}_\ominus + \frac{1}{\eta}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{\epsilon}_\ominus\right)$ and $\boldsymbol{w}_\ominus^{(\infty)} = \arg\min_{\boldsymbol{w}:\boldsymbol{X}_-\boldsymbol{w}=-\boldsymbol{y}_-} \left\|\boldsymbol{w} - \boldsymbol{w}_\ominus^{(1)}\right\|_2$, with $\boldsymbol{X}_-\boldsymbol{w}_\oplus^{(\infty)} \preceq \boldsymbol{0}$ and $\boldsymbol{X}_+\boldsymbol{w}_\ominus^{(\infty)} \preceq \boldsymbol{0}$.*

Theorem 7 is proved in Appendix E.2. The optimization dynamics naturally decouple: $\boldsymbol{w}_\oplus$ learns to fit all positively labeled examples, while $\boldsymbol{w}_\ominus$ learns to fit all negatively labeled examples.

4.3. Approximation to Minimum-ℓ_2-norm Solution in High Dimensions

Finally, we show that the limiting solution of Theorem 7 is close to the minimum-ℓ_2-norm $\{\boldsymbol{w}_\oplus^\star, \boldsymbol{w}_\ominus^\star\}$.

Theorem 8 *Consider Assumptions 1 and 2, suppose the initialization is $\boldsymbol{w}_\oplus^{(0)} = \boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{\epsilon}_\oplus$, $\boldsymbol{w}_\ominus^{(0)} = \boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{\epsilon}_\ominus$, where $0 < \epsilon_{\oplus,i}, \epsilon_{\ominus,i} \le \frac{1}{2C_\alpha} y_{\min}$ for all $i \in [n]$, and the step size satisfies $\frac{1}{CC_g\|\boldsymbol{\lambda}\|_1} \le \eta \le \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$. Then, we have $\sqrt{\frac{n_- y_{\min}^2}{CC_g\|\boldsymbol{\lambda}\|_1}} \le \left\|\boldsymbol{w}_\oplus^{(\infty)} - \boldsymbol{w}_\oplus^\star\right\|_2 \le \sqrt{\frac{16 n_- y_{\max}^2}{C_g\|\boldsymbol{\lambda}\|_1}}$ and $\sqrt{\frac{n_+ y_{\min}^2}{CC_g\|\boldsymbol{\lambda}\|_1}} \le \left\|\boldsymbol{w}_\ominus^{(\infty)} - \boldsymbol{w}_\ominus^\star\right\|_2 \le \sqrt{\frac{16 n_+ y_{\max}^2}{C_g\|\boldsymbol{\lambda}\|_1}}$ with probability at least $1 - 2\exp(-cn)$.*

Proof is in Appendix E.3 and leverages the restricted convex program that we derived in Lemma 6.

5. Discussion

We showed that the implicit bias of single and 2-ReLU models, under appropriate initialization, is remarkably close to the minimum-norm solution if the features are sufficiently high-dimensional. Natural open questions include: 1) characterizing the dynamics for $m > 2$ neurons, and 2) studying the effect of moderate dimension where $d > n$ but not $d \gg n$. We provide partial results of $m > 2$ neurons in Appendices F and G and simulate the effect of moderate-dimensional data on the dynamics in Appendix H and observe that the primal dual variables intricately influence each other.

References

[1] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences (PNAS)*, 117(48):30063–30070, 2020.

[2] Rajendra Bhatia and Fuad Kittaneh. On the singular values of a product of operators. *SIAM Journal on Matrix Analysis and Applications*, 11(2):272–277, 1990.

[3] Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. *Advances in Neural Information Processing Systems*, 35:20105–20118, 2022.

[4] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[5] Léo Dana, Francis Bach, and Loucas Pillaud-Vivien. Convergence of shallow relu networks on weakly interacting data. *arXiv preprint arXiv:2502.16977*, 2025.

[6] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[7] Spencer Frei, Gal Vardi, Peter Bartlett, and Nathan Srebro. Benign overfitting in linear classifiers and leaky relu networks from kkt conditions for margin maximization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3173–3228. PMLR, 2023.

[8] Spencer Frei, Gal Vardi, Peter L Bartlett, and Wei Hu. Implicit bias in leaky relu networks trained on high-dimensional data. *ICLR*, 2023.

[9] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

[10] Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 91–99. PMLR, 2021.

[11] Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

[12] Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

[13] 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.

[14] 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.

[15] Gal Vardi and Ohad Shamir. Implicit regularization in relu networks with the square loss. In *Conference on Learning Theory*, pages 4224–4258. PMLR, 2021.

[16] Ke Wang and Christos Thrampoulidis. Binary classification of Gaussian mixtures: Abundance of support vectors, benign overfitting, and regularization. *SIAM Journal on Mathematics of Data Science (SIMODS)*, 4:260–284, 2022.

[17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Appendix

Table of Contents

Appendix A. Related work, our contribution and notation

We now briefly discuss the most closely related work and highlight key differences of our approach. We contextualize our results within the most closely related prior studies on implicit bias of regression models in Table 1. Boursier et al. [3] study the dynamics of gradient flow on two-layer ReLU networks under an *exact* orthogonality assumption on the data. Exact orthogonality removes interactions between examples and significantly simplifies the activation patterns induced by the ReLU nonlinearity. As a result, their analysis primarily focuses on how the second-layer weights evolve to fit all examples, leading to a multi-phase gradient flow dynamic. Under these assumptions, they show that gradient flow converges to the minimum-ℓ_2-norm solution (their Theorem 1). In contrast, our work focuses on understanding how interactions between examples—captured through the Gram matrix—shape the active and inactive patterns in ReLU models under more realistic, controllable high-dimensional settings. Interestingly, we are able to show that the implicit bias is no longer exactly the minimum-ℓ_2-norm solution, but is close to it (Theorems 5 and 8). Dana et al. [5] also analyze the high-dimensional regime and establish global convergence by showing that each example can be fitted by at least one neuron with high probability and all active examples stay active (their Theorem 1). However, their analysis does not address the behavior of inactive examples suppressed by the ReLU nonlinearity and does not shed light on the implicit bias. As a result, their work provides only a partial view of the gradient dynamics. In contrast, we introduce a novel primal–dual framework that allows us to simultaneously track both active and inactive examples (Lemmas 9 and 10). This framework enables a full characterization of the gradient dynamics and, consequently, the implicit bias in high dimensions. We use some of the observations of Dana et al. [5] as a starting point for our primal-dual characterizations. More generally, most existing analyses [3, 5, 15] rely on gradient flow and continuous-time ODE techniques, which assume infinitesimal step sizes. In contrast, our analysis directly studies gradient descent with finite (though still small) step sizes. This distinction is both theoretically and practically important, as gradient descent is the algorithm used in practice. Our primal–dual approach provides a new framework for analyzing discrete-time optimization dynamics in ReLU networks and opens a complementary direction to existing studies based on gradient flow.

Our contributions: In this paper, we provide a rich characterization of the implicit bias induced by gradient descent for ReLU networks trained with the squared loss on sufficiently high-dimensional data. Our main contributions are summarized as follows. First, we completely characterize the implicit bias of gradient descent dynamics on ReLU models with 1 or 2 neurons for high-dimensional data under sufficient conditions (Theorems 4 and 7). Second, we quantify the relationship between the implicit bias of gradient descent and the global minimum that achieves the minimum-ℓ_2-norm. More specifically, we establish both upper and lower bounds on the distance between the gradient descent limit and the minimum-ℓ_2-norm solution, showing that it scales as $\Theta(\sqrt{n/\|\boldsymbol{\lambda}\|_1})$ where n is the number of training examples and $\boldsymbol{\lambda}$ denotes the spectrum of the data covariance matrix (Theorems 5 and 8). Consequently, the solutions are very close, but not identical, for high-dimensional features. Interestingly, a similar phenomenon was also shown to occur with exponentially-tailed losses [7, 8] for classification.

Our techniques in a nutshell: Our main results are obtained through a novel primal-dual formulation of the gradient descent dynamics under the squared loss with ReLU networks, which is inspired by mirror descent (first studied by Ji and Telgarsky [12] for linear models). Instead of directly tracking the weight vector in the original parameter space like previous work, we introduce primal variables representing the predictions on training examples, and dual variables capturing the coefficients in the data span. This representation is particularly well-suited for analyzing ReLU networks because the sign of each primal variable directly determines whether the corresponding example is active, and hence whether its dual variable receives a gradient update. Our analysis reveals that understanding the gradient dynamics hinges on tracking (i) the positivity of the primal variables and (ii) the interactions between training examples. We introduce new tools to carefully control the evolution of positive primal variables and sufficiently negative dual variables (Lemmas 9 and 10, which may be of independent interest). Underlying the proofs of our approximation results to the minimum-ℓ_2-norm solutions are novel characterizations of the latter as minimum-ℓ_2-norm *linear* interpolations of a (possibly data-dependent) subset of training examples. This data-dependent subset selection is a fundamental difference between the implicit bias of linear models and ReLU models.

Notation: We use lowercase boldface letters (e.g. $\boldsymbol{x}$) to denote vectors, lowercase letters (e.g. y) to denote scalars, and uppercase boldface letters (e.g. $\boldsymbol{X}$) to denote matrices. We use $\|\cdot\|_p$ to denote the ℓ_p-norm of a vector for $p \in [1, \infty)$ and $\|\cdot\|_2$ to additionally denote the operator norm of a matrix. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, we use x_i to denote its i-th component. We use $[n]$ to denote the set $\{1, \ldots, n\}$. For ease of subsequent notation, we consider without loss of generality the samples with positive labels to appear in the upper block of the data matrix $\boldsymbol{X}$, while those with negative labels appear in the lower block. Let n_+ denote the number of positive labels and $n_- = n - n_+$ denote the number of negative labels. Accordingly, we write $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_+^\top & \boldsymbol{X}_-^\top \end{bmatrix}^\top$ where $\boldsymbol{X}_+ \in \mathbb{R}^{n_+ \times d}$ contains the features corresponding to positive labels and $\boldsymbol{X}_- \in \mathbb{R}^{n_- \times d}$ contains the features corresponding to negative labels. We similarly partition the label vector as $\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_+^\top & \boldsymbol{y}_-^\top \end{bmatrix}^\top$. For a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, a vector $\boldsymbol{y} \in \mathbb{R}^n$, and any index set $S \subseteq [n]$, we use $\boldsymbol{X}_S \in \mathbb{R}^{|S| \times d}$ to denote the submatrix of $\boldsymbol{X}$ consisting of the rows indexed by S, and $\boldsymbol{y}_S \in \mathbb{R}^{|S|}$ denotes the corresponding subvector. We use C, c to denote universal constants that appear in upper and lower bounds respectively that may change from line to line. We also use the notation $C_{(\cdot)}$ to denote universal constants with a specific meaning that *do not* change from line to line. We specifically choose $C_0 \gtrsim C_\alpha^2$ and $C_\alpha \gtrsim \max\{C_g^2, C_y C_g\}$ in our analysis.

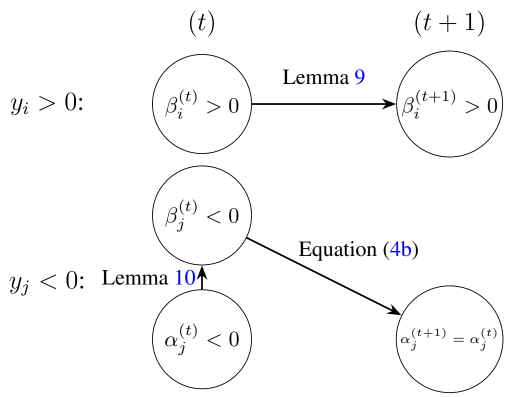


Figure 1: Gradient descent transition diagram for the k-th neuron.

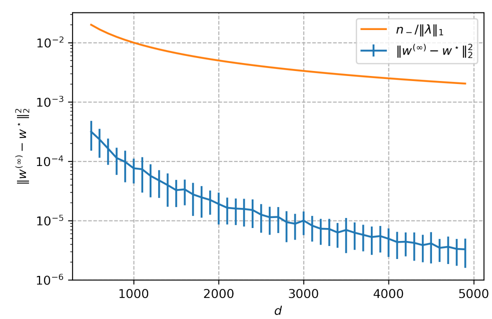


Figure 2: Approximation error between the implicit bias of the single ReLU model $\boldsymbol{w}^{(\infty)}$ and the minimum-ℓ_2-norm solution $\boldsymbol{w}^\star$.

Appendix B. Main Proof Ideas

Our analysis hinges on **precisely tracking the activation patterns of ReLU neurons across all training examples.** By controlling which examples remain active or inactive throughout training, we are able to understand the resulting gradient dynamics and, consequently, the implicit bias of the converged solution. To establish these results, we introduce two key lemmas. Lemma 9, inspired by ideas in Dana et al. [5], shows that once the primal variable $\beta_{k,i}$ corresponding to the k-th neuron and the i-th example is active—and the sign of the neuron s_k agrees with the label y_i—it remains active in the next iteration. This ensures that such an example is not suppressed by the ReLU nonlinearity and continues to contribute to the gradient updates.

Lemma 9 *Under Assumptions 1 and 2, suppose the gradient descent step size satisfies* $\eta \leq \frac{1}{C_g \|\boldsymbol{\lambda}\|_1}$. *Consider the k-th ReLU neuron in* $h_{\boldsymbol{\Theta}}$. *For any* $t \geq 0$ *and any index* $i \in [n]$ *such that* $s_k \cdot y_i > 0$, *if* $\beta_{k,i}^{(t)} > 0$, $\beta_{k,i}^{(t)} \geq s_k \cdot h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{x}_i)$, *and* $\left\|h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}\right\|_2 \leq C_y \|\boldsymbol{y}\|_2$, *then* $\beta_{k,i}^{(t+1)} > 0$ *with probability at least* $1 - 2\exp(-cn)$.

This lemma is proved in Appendix C.2. The main idea behind Lemma 9 is that as long as the primal variable $\beta_{k,i}^{(t)}$ is positive and the empirical risk remains uniformly bounded, the gradient update of $\beta_{k,i}^{(t)}$ is dominated by its self-interaction term for high-dimensional data — the reason, at a high level, is that cross-sample interactions can be bounded in high dimensions (due to the concentration of the random Gram matrix $\boldsymbol{X}\boldsymbol{X}^\top$ around $\|\boldsymbol{\lambda}\|_1 \boldsymbol{I}$). As a result, the magnitude of the update is strictly smaller than $\beta_{k,i}^{(t)}$, ensuring that $\beta_{k,i}^{(t+1)}$ remains positive.

Lemma 10 concerns the behavior of inactive examples. It shows that once a dual variable $\alpha_{k,j}$ associated with the k-th neuron and the j-th example becomes sufficiently negative, the corresponding primal variable $\beta_{k,j}$ remains inactive. Consequently, the dual variable is no longer updated and stays frozen throughout training. This mechanism effectively removes certain examples from the optimization dynamics.

Lemma 10 *Under Assumptions 1 and 2, suppose the step size satisfies* $\eta \leq \frac{1}{C_g \|\boldsymbol{\lambda}\|_1}$. *Consider the* k*-th ReLU neuron in* $h_{\boldsymbol{\Theta}}$. *For any* $t \geq 0$ *and any index* $j \in [n]$, *if* $\alpha_{k,j}^{(t)} \leq -\frac{y_{\min}}{C_\alpha \|\boldsymbol{\lambda}\|_1}$ *and* $\left\|\boldsymbol{\alpha}_k^{(t)}\right\|_2 \leq \frac{C_\alpha \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1}$, *then* $\beta_{k,j}^{(t)} \leq 0$ *and* $\alpha_{k,j}^{(t+1)} = \alpha_{k,j}^{(t)}$ *with probability at least* $1 - 2\exp(-cn)$.

The proof of Lemma 10 (see Appendix C.3) relies on the primal–dual relationship $\boldsymbol{\beta}_k = \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\alpha}_k$ from Equation (3), together with concentration results for the Gram matrix. Specifically, if a dual variable is sufficiently negative, then the corresponding primal variable $\beta_{k,j}^{(t)}$ is strictly negative. According to the dual update rule in Equation (4b), a negative $\beta_{k,j}^{(t)}$ implies that the ReLU is inactive and the dual coordinate receives no further updates. As a result, $\alpha_{k,j}^{(t+1)} = \alpha_{k,j}^{(t)}$, and sufficiently negative dual variables remain frozen throughout training. Figure 1 depicts the transition of primal-dual updates in Lemma 9 and Lemma 10.

In the following paragraphs, we outline the proof sketch for single ReLU models. The proof ideas for the 2-ReLU case follow analogously.

Proof Sketch of Theorem 4: The proof combines the insights from Lemma 9 and Lemma 10 to obtain a complete picture of how activation patterns evolve during training. Together, these lemmas allow us to track which examples remain active or inactive throughout gradient descent. Our goal is to reach—and maintain—a configuration in which positive-labeled examples remain active while negative-labeled examples remain inactive, as formalized by the sufficient conditions in Lemma 14 in Appendix D.2. To achieve this, we leverage two key properties of the initialization. First, the positive initialization guarantees that every example initially has at least one active neuron capable of fitting it. Second, the small initialization ensures that, after the first gradient step, positive-labeled examples remain in the active regime while negative-labeled examples acquire sufficiently negative dual variables and become inactive. Together, these properties place positive and negative examples into their respective regimes after a single update. We then apply Lemma 14 to show that this configuration is stable under subsequent iterations. As a result, the activation pattern becomes fixed after the first step, and the dynamics enter the final phase described in Lemma 1.

Proof Sketch of Theorem 5: To compare the gradient descent limit $\boldsymbol{w}^{(\infty)}$ with the minimum-ℓ_2-norm solution $\boldsymbol{w}^\star$, we relate their distance in parameter space to their distance in prediction space. Since both solutions interpolate all positive-labeled examples exactly, any discrepancy between them must arise from their predictions on negative-labeled examples. We bound this discrepancy using the KKT conditions characterizing $\boldsymbol{w}^\star$, as established in Lemma 3. These conditions precisely describe how $\boldsymbol{w}^\star$ treats negative-labeled examples and allow us to control the prediction distance in terms of the distance between the primal and dual variables. In particular, the KKT conditions imply that this gap is nonzero, showing that $\boldsymbol{w}^{(\infty)} \neq \boldsymbol{w}^\star$. Translating our bounds back to parameter space yields matching upper and lower bounds on $\|\boldsymbol{w}^{(\infty)} - \boldsymbol{w}^\star\|_2$.

Appendix C. Proofs of Key Lemmas Tracking Primal–Dual Gradient Dynamics

In this section, we present the proofs of the key lemmas used to track the gradient dynamics of the primal and dual variables. The central factor governing these dynamics is the sign pattern of the primal variables, which determines whether individual examples are active or inactive under the ReLU nonlinearity and, consequently, whether the corresponding dual variables are updated.

Before presenting the proofs, we first recall two key technical lemmas: 1) a concentration result on the eigenvalues of random Gram matrices in high dimensions from Bartlett et al. [1]; 2) a concentration bound on the operator norm of random Gram matrices from Hsu et al. [10]. Both these lemmas play a crucial role throughout the analysis.

C.1. Concentration of Random Gram Matrices in High Dimensions

Our analysis relies heavily on properties of the Gram matrix on high-dimensional data. These concentration results allow us to control cross-sample interactions and isolate the dominant self-interaction terms that drive the gradient updates. As a result, we can rigorously characterize how positivity and negativity patterns in the primal and dual variables evolve over time.

In Lemma 11, we characterize the typical behavior of the eigenvalues of a weighted sum of outer products of independent subgaussian vectors. Recall from Section 2 that the feature vector $\boldsymbol{x} \in \mathbb{R}^d$ admits the representation $\boldsymbol{x} = \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{z}$, where $\boldsymbol{z} \in \mathbb{R}^d$ has independent, mean-zero, σ_z^2-subgaussian components, and we take $\sigma_z = 1$. Under this model, the empirical Gram matrix can be written as $\boldsymbol{X}\boldsymbol{X}^\top = \sum_{j=1}^d \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^\top$ where each $\boldsymbol{v}_j \in \mathbb{R}^n$ is an independent random vector with independent, mean-zero, subgaussian entries. Concretely, Lemma 11 provides high-probability bounds on the extreme eigenvalues of $\boldsymbol{X}\boldsymbol{X}^\top$.

Lemma 11 (1, Lemma 9, 16, Lemma 12) *There exists a constant c such that with probability at least $1 - 2e^{-n/c}$, we have*

$$\frac{1}{c}\sum_{j=1}^d \lambda_j - c\lambda_1 n \le \mu_n(\boldsymbol{X}\boldsymbol{X}^\top) \le \mu_1(\boldsymbol{X}\boldsymbol{X}^\top) \le c\left(\sum_{j=1}^d \lambda_j + \lambda_1 n\right).$$

Moreover, if the effective dimension satisfies $d_\infty = \frac{\sum_{j=1}^d \lambda_j}{\lambda_1} \ge bn$ for some constant $b \ge 1$, then there exists a constant $C_g \ge 1$ such that

$$\frac{1}{C_g}\sum_{j=1}^d \lambda_j \le \mu_n(\boldsymbol{X}\boldsymbol{X}^\top) \le \mu_1(\boldsymbol{X}\boldsymbol{X}^\top) \le C_g \sum_{j=1}^d \lambda_j.$$

with probability at least $1 - 2e^{-n/C_g}$.

Next, Lemma 12 provides a high-probability bound on the operator norm deviation between the Gram matrix $\boldsymbol{X}\boldsymbol{X}^\top$ and $\|\boldsymbol{\lambda}\|_1 \boldsymbol{I}$, which is fruitful for high-dimensional data, and Corollary 13 shows that the typical value of this deviation can be expressed in terms of n and effective dimensions d_2, d_∞.

Lemma 12 (10, Lemma 8) *There exists a universal constant $c > 0$, for any $\tau > 0$,*

$$\Pr\left(\left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right\|_2 \ge \tau\right) \le 2 \cdot 9^n \cdot \exp\left(-c \cdot \min\left\{\frac{\tau^2}{\|\boldsymbol{\lambda}\|_2^2}, \frac{\tau}{\|\boldsymbol{\lambda}\|_\infty}\right\}\right),$$

where $\|\boldsymbol{\lambda}\|_1 \coloneqq \sum_{j=1}^d \lambda_j$, $\|\boldsymbol{\lambda}\|_2^2 \coloneqq \sum_{j=1}^d \lambda_j^2$, *and* $\|\boldsymbol{\lambda}\|_\infty \coloneqq \max_{j\in[d]} \lambda_j$.

Corollary 13 *With the choice of* $\tau = C \cdot \max(\|\boldsymbol{\lambda}\|_2 \sqrt{n}, \|\boldsymbol{\lambda}\|_\infty n)$ *and the constant* $C \cdot c > \ln 9$*, we have*

$$\left\| \frac{1}{\|\boldsymbol{\lambda}\|_1} \boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{I} \right\|_2 \le C \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty} \right),$$

with probability at least $1 - 2\exp(-n(Cc - \ln 9))$*, where we have defined* $d_2 \coloneqq \frac{\|\boldsymbol{\lambda}\|_1^2}{\|\boldsymbol{\lambda}\|_2^2}, d_\infty \coloneqq \frac{\|\boldsymbol{\lambda}\|_1}{\|\boldsymbol{\lambda}\|_\infty}$*. Similarly, we have*

$$\left\| \|\boldsymbol{\lambda}\|_1 \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \boldsymbol{I} \right\|_2 \le C_g C \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty} \right),$$

with probability at least $1 - 2\exp(-n(Cc - \ln 9))$.

C.2. Proof of Lemma 9 (Primal Variable Gradient Dynamics in High Dimensions)

In this proof, we show that under the assumptions of the lemma, if the sign of any ReLU neuron agrees with the label of an example, then the corresponding primal variable remains positive after one gradient descent step.

Proof (Lemma 9) According to the primal gradient descent update in Equation (4a) for the k-th neuron, we have

$$\boldsymbol{\beta}_k^{(t+1)} = \boldsymbol{\beta}_k^{(t)} - \eta s_k \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\beta}_k^{(t)})(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}).$$

We aim to separate the gradient contribution arising from the diagonal and off-diagonal components of the Gram matrix and to show that the updated primal coordinate remains positive, i.e., $\beta_{k,i}^{(t+1)} > 0$. Fix any $t \ge 0$ and any index i such that $s_k \cdot y_i > 0$ and $\beta_{k,i}^{(t)} > 0$. Then, the update of the i-th coordinate can be written as

$$\begin{aligned}
\beta_{k,i}^{(t+1)} &= \beta_{k,i}^{(t)} - \eta s_k \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\beta}_k^{(t)})(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}) \\
&= \beta_{k,i}^{(t)} - \eta s_k \boldsymbol{e}_i^\top \left[\|\boldsymbol{\lambda}\|_1 \boldsymbol{I} + \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I} \right) \right] \boldsymbol{D}(\boldsymbol{\beta}_k^{(t)})(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}) \\
&= \left[\beta_{k,i}^{(t)} - \eta \|\boldsymbol{\lambda}\|_1 \left(s_k h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{x}_i) - s_k y_i \right) \right] - \eta s_k \boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I} \right) \boldsymbol{D}(\boldsymbol{\beta}_k^{(t)})(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}),
\end{aligned} \tag{9}$$

where the last equality uses the assumption $\beta_{k,i}^{(t)} > 0$, which implies $D_{ii} = \mathbb{1}_{\beta_{k,i}^{(t)} > 0} = 1$. We now lower bound $\beta_{k,i}^{(t+1)}$. By the step size condition $\eta \le \frac{1}{C_g \|\boldsymbol{\lambda}\|_1}$ and the assumption $\beta_{k,i}^{(t)} \ge s_k \cdot h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{x}_i)$, the first term in Equation (9) satisfies

$$\beta_{k,i}^{(t)} - \eta \|\boldsymbol{\lambda}\|_1 \left(s_k h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{x}_i) - s_k y_i \right) \ge \eta \|\boldsymbol{\lambda}\|_1 |y_i|.$$

Substituting this into Equation (9) yields

$$\begin{aligned}
(9) &\ge \eta \|\boldsymbol{\lambda}\|_1 |y_i| - \eta s_k \boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I} \right) \boldsymbol{D}(\boldsymbol{\beta}_k^{(t)})(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}) \\
&\ge \eta \|\boldsymbol{\lambda}\|_1 |y_i| - \eta \left\| \boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I} \right\|_2 \left\| h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y} \right\|_2,
\end{aligned} \tag{10}$$

where the last inequality follows from the Cauchy–Schwarz inequality and the sub-multiplicativity of the operator norm. Next, we upper bound the second term in Equation (10) using Corollary 13. With probability at least $1-2\exp(-n(Cc-\ln 9))$, we obtain

$$\begin{aligned}
(10) &\geq \eta \left\|\boldsymbol{\lambda}\right\|_1 \left[|y_i| - C\cdot\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\left\|h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X})-\boldsymbol{y}\right\|_2\right] \\
&\overset{\text{(i)}}{\geq} \eta \left\|\boldsymbol{\lambda}\right\|_1 \left[y_{\min} - C\cdot\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\cdot C_y\sqrt{n}y_{\max}\right] \\
&\overset{\text{(ii)}}{\geq} \eta \left\|\boldsymbol{\lambda}\right\|_1 \left[y_{\min} - C\cdot C_y\cdot\frac{y_{\min}}{C_0 y_{\max}}\cdot y_{\max}\right] \\
&> 0.
\end{aligned}$$

Inequality (i) applies the lemma assumption that $\left\|h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X})-\boldsymbol{y}\right\|_2 \leq C_y\left\|\boldsymbol{y}\right\|_2 \leq C_y\sqrt{n}y_{\max}$. Inequality (ii) follows from Assumption 2, which guarantees that $d_2 \geq C_0^2\frac{n^2y_{\max}^2}{y_{\min}^2}$ and $d_\infty \geq C_0\frac{n^{1.5}y_{\max}}{y_{\min}}$ with large enough $C_0 > C\cdot C_y$. This completes the proof of the lemma. ■

C.3. Proof of Lemma 10 (Dual Variable Gradient Dynamics in High Dimensions)

In this proof, we show that under the assumptions of the lemma, if the dual variable $\alpha_{k,j}^{(t)}$ for the k-th neuron and j-th example is sufficiently negative, then it remains unchanged in the next iteration, i.e., $\alpha_{k,j}^{(t+1)} = \alpha_{k,j}^{(t)}$.

Proof (Lemma 10) By the definition of primal and dual variables in Equation (3), we have

$$\boldsymbol{\beta}_k^{(t)} = \boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{\alpha}_k^{(t)}.$$

According to the dual gradient update in Equation (4b), we have

$$\boldsymbol{\alpha}_k^{(t+1)} = \boldsymbol{\alpha}_k^{(t)} - \eta\boldsymbol{D}(\boldsymbol{\beta}_k^{(t)})(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X})-\boldsymbol{y}).$$

This update reveals that each coordinate $\alpha_{k,j}^{(t)}$ evolves independently and is governed by the sign of the corresponding primal variable $\beta_{k,j}^{(t)}$. In particular, if $\beta_{k,j}^{(t)} \leq 0$, then the j-th diagonal entry of $\boldsymbol{D}(\boldsymbol{\beta}_k^{(t)})$ vanishes, and consequently $\alpha_{k,j}^{(t+1)} = \alpha_{k,j}^{(t)}$.

We therefore establish a sufficient condition under which $\beta_{k,j}^{(t)} \leq 0$ in terms of the dual variable $\alpha_{k,j}^{(t)}$. Specifically, we separate the diagonal and off-diagonal components of the Gram matrix as

$$\begin{aligned}
\beta_{k,j}^{(t)} &= \boldsymbol{e}_j^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{\alpha}_k^{(t)} \\
&= \boldsymbol{e}_j^\top\left[\left\|\boldsymbol{\lambda}\right\|_1\boldsymbol{I} + \left(\boldsymbol{X}\boldsymbol{X}^\top - \left\|\boldsymbol{\lambda}\right\|_1\boldsymbol{I}\right)\right]\boldsymbol{\alpha}_k^{(t)} \\
&= \left\|\boldsymbol{\lambda}\right\|_1\alpha_{k,j}^{(t)} + \boldsymbol{e}_j^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \left\|\boldsymbol{\lambda}\right\|_1\boldsymbol{I}\right)\boldsymbol{\alpha}_k^{(t)} \\
&\leq \left\|\boldsymbol{\lambda}\right\|_1\alpha_{k,j}^{(t)} + \left\|\boldsymbol{X}\boldsymbol{X}^\top - \left\|\boldsymbol{\lambda}\right\|_1\boldsymbol{I}\right\|_2\left\|\boldsymbol{\alpha}_k^{(t)}\right\|_2, \qquad (11)
\end{aligned}$$

where the last inequality follows from the sub-multiplicativity of the operator norm. Next, we upper bound the two terms $\left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right\|_2$ and $\left\|\boldsymbol{\alpha}_k^{(t)}\right\|_2$ appearing in Equation (11). Following the same argument as in the proof of Lemma 9, we apply Corollary 13. Consequently, with probability at least $1 - 2\exp(-n(Cc - \ln 9))$, we obtain

$$(11) \leq \|\boldsymbol{\lambda}\|_1 \left[\alpha_{k,j}^{(t)} + C \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right) \left\|\boldsymbol{\alpha}_k^{(t)}\right\|_2\right]. \tag{12}$$

Finally, substituting the upper bound of $\alpha_{k,j}^{(t)}$ and $\left\|\boldsymbol{\alpha}_k^{(t)}\right\|_2$ in lemma assumptions into Equation (12), we obtain

$$\begin{aligned}
(12) &\leq \|\boldsymbol{\lambda}\|_1 \left[-\frac{y_{\min}}{C_\alpha \|\boldsymbol{\lambda}\|_1} + C \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right) \frac{C_\alpha \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right] \\
&= \|\boldsymbol{\lambda}\|_1 \left[-\frac{y_{\min}}{C_\alpha \|\boldsymbol{\lambda}\|_1} + C \cdot \frac{y_{\min}}{C_0 y_{\max}} \frac{C_\alpha y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right] \\
&\leq 0,
\end{aligned}$$

where the last inequality follows from Assumption 2, which ensures $d_2 \geq C_0^2 \frac{n^2 y_{\max}^2}{y_{\min}^2}$ and $d_\infty \geq C_0 \frac{n^{1.5} y_{\max}}{y_{\min}}$ with large enough $C_0 > C \cdot C_\alpha^2$. We have thus shown that if $\alpha_{k,j}^{(t)}$ is sufficiently negative, then $\beta_{k,j}^{(t)} \leq 0$, and consequently $\alpha_{k,j}^{(t+1)} = \alpha_{k,j}^{(t)}$. This completes the proof of the lemma. ■

Appendix D. Proofs for the Single ReLU model ($m = 1$) Trained with Gradient Descent

In this section, we present the proofs concerning the behavior of the single ReLU model trained with gradient descent.

D.1. Proofs of Lemmas 1, 2 and 3 (Gradient Descent Convergence and $w^\star$)

We present complete proofs of the gradient descent convergence for single ReLU models in Lemmas 1 and 2, as well as a characterization of the minimum-ℓ_2-norm solution in Lemma 3.

Proof (Lemma 1) We prove this lemma by showing that after iteration $t_0 \geq 0$, since the activation pattern is fixed, the gradient of the single ReLU model is equivalent to the gradient of a linear model using only a subset of examples. Consider a linear model

$$h(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x},$$

where $\boldsymbol{w} \in \mathbb{R}^d$ is the linear model parameter (also called weight). Let $S \subseteq [n]$ denote the active set for the single ReLU model at iteration t_0, defined by $S \coloneqq \{i \in [n] : \boldsymbol{x}_i^\top \boldsymbol{w}^{(t_0)} > 0\}$. We write the empirical risk with the linear model using only the examples in S as

$$\mathcal{R}_S(\boldsymbol{w}) = \frac{1}{2} \sum_{i \in S} (\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2.$$

The gradient descent update for this linear model is

$$\begin{aligned} \boldsymbol{w}^{(t+1)} &= \boldsymbol{w}^{(t)} - \eta \nabla \mathcal{R}_S(\boldsymbol{w}^{(t)}) \\ &= \boldsymbol{w}^{(t)} - \eta \sum_{i \in S} (\boldsymbol{w}^{(t)\top} \boldsymbol{x}_i - y_i) \boldsymbol{x}_i. \end{aligned} \tag{13}$$

On the other hand, the original gradient descent dynamic for the single ReLU model tells us that

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta \boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}^{(t)})(\boldsymbol{X}\boldsymbol{w}^{(t)} - \boldsymbol{y}).$$

Under the lemma assumption, $\boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}^{(t_0)}) = \boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}^{(t)})$ for all $t \geq t_0$. Thus, we know that $D_{ii} = \mathbb{1}_{i \in S}$ for all $t \geq t_0$. Therefore, for $t \geq t_0$, we can write the gradient update of the original single ReLU model as

$$\begin{aligned} \boldsymbol{w}^{(t+1)} &= \boldsymbol{w}^{(t)} - \eta \boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}^{(t)})(\boldsymbol{X}\boldsymbol{w}^{(t)} - \boldsymbol{y}) \\ &= \boldsymbol{w}^{(t)} - \eta \sum_{i \in S} (\boldsymbol{w}^{(t)\top} \boldsymbol{x}_i - y_i) \boldsymbol{x}_i. \end{aligned}$$

This gradient update is equivalent to the gradient update of the linear model in Equation (13) for all $t \geq t_0$. As a result, for $t \geq t_0$, the gradient update of the single ReLU model is equivalent to a linear model using only data in S. This completes the proof of the lemma. ■

Proof (Lemma 2) By Lemma 1, the activation pattern is fixed for all $t \geq t_0$, so the gradient descent update reduces to linear regression restricted to the active subset S, given by

$$\begin{aligned} \boldsymbol{w}^{(t+1)} &= \boldsymbol{w}^{(t)} - \eta \boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}^{(t)})(\boldsymbol{X}\boldsymbol{w}^{(t)} - \boldsymbol{y}) \\ &= \boldsymbol{w}^{(t)} - \eta \sum_{i \in S} (\boldsymbol{w}^{(t)\top} \boldsymbol{x}_i - y_i) \boldsymbol{x}_i \\ &= \boldsymbol{w}^{(t)} - \eta \boldsymbol{X}_S^\top (\boldsymbol{X}_S \boldsymbol{w}^{(t)} - \boldsymbol{y}_S). \end{aligned}$$

The final phase empirical risk is given by

$$\mathcal{R}(\boldsymbol{w}) = \frac{1}{2}\left\|\boldsymbol{X}_S\boldsymbol{w} - \boldsymbol{y}_S\right\|_2^2 + \frac{1}{2}\left\|\boldsymbol{y}_{S^{\mathsf{c}}}\right\|_2^2,$$

where the second term comes from the examples in S^{c} with negative pre-activations, and it does not depend on $\boldsymbol{w}$ because the activation pattern does not change after t_0. Note that $\mathcal{R}(\boldsymbol{w})$ is a convex quadratic with

$$\nabla\mathcal{R}(\boldsymbol{w}) = \boldsymbol{X}_S^\top(\boldsymbol{X}_S\boldsymbol{w} - \boldsymbol{y}_S), \qquad \nabla^2\mathcal{R}(\boldsymbol{w}) = \boldsymbol{X}_S^\top\boldsymbol{X}_S.$$

Therefore, $\mathcal{R}$ is L-smooth with

$$L = \left\|\nabla^2\mathcal{R}(\boldsymbol{w})\right\|_2 = \left\|\boldsymbol{X}_S^\top\boldsymbol{X}_S\right\|_2 = \mu_1(\boldsymbol{X}_S\boldsymbol{X}_S^\top).$$

A standard smoothness/descent result (e.g., Boyd and Vandenberghe 4, Equation 9.17) implies that for any $\eta \leq \frac{1}{L}$,

$$\mathcal{R}(\boldsymbol{w}^{(t+1)}) \leq \mathcal{R}(\boldsymbol{w}^{(t)}) - \frac{\eta}{2}\left\|\nabla\mathcal{R}(\boldsymbol{w}^{(t)})\right\|_2^2,$$

and in particular, $\mathcal{R}(\boldsymbol{w}^{(t)})$ is non-increasing for all $t \geq t_0$.

It remains to upper bound L. Since S is a subset of the training indices, $|S| \leq n$. Under the effective dimension condition $d_\infty \geq bn$, Lemma 11 applies to the Gram matrix $\boldsymbol{X}\boldsymbol{X}^\top$, and yields that with probability at least $1 - 2e^{-n/C_g}$, $\mu_1(\boldsymbol{X}_S\boldsymbol{X}_S^\top) \leq \mu_1(\boldsymbol{X}\boldsymbol{X}^\top) \leq C_g\left\|\boldsymbol{\lambda}\right\|_1$. On this event, we have $L \leq C_g\left\|\boldsymbol{\lambda}\right\|_1$, Hence, choosing $\eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$ guarantees that $\mathcal{R}(\boldsymbol{w}^{(t)})$ is non-increasing for all $t \geq t_0$. This establishes the desired step size condition in the final phase (and thus convergence in function value for the single ReLU dynamics after t_0).

Finally, according to Gunasekar et al. [9, Section 2.1], the set of minimizers of $\mathcal{R}(\boldsymbol{w})$ is the affine subspace,

$$\mathcal{W}_S = \{\boldsymbol{w} : \boldsymbol{X}_S\boldsymbol{w} = \boldsymbol{y}_S\},$$

and gradient descent with constant step size converges to the Euclidean projection of the initialization $\boldsymbol{w}^{(t_0)}$ onto this subspace $\boldsymbol{w}^{(\infty)} = \underset{\boldsymbol{w}\in\mathcal{W}_S}{\arg\min}\left\|\boldsymbol{w} - \boldsymbol{w}^{(t_0)}\right\|_2$. This completes the proof of the lemma. ■

Proof (Lemma 3) We prove the lemma by showing that the optimal solution $\boldsymbol{w}^\star$ of the original convex program for single ReLU models also solves a reduced convex program whose solution is the minimum-ℓ_2-norm interpolation (MNI) over an index subset $S \subseteq [n]$ with modified labels. First, we restate the convex program (6) and its KKT conditions below:

$$\begin{aligned} \boldsymbol{w}^\star \in \arg\min_{\boldsymbol{w}} \frac{1}{2}\left\|\boldsymbol{w}\right\|_2^2 \\ \text{s.t. } \boldsymbol{w}^\top\boldsymbol{x}_i = y_i, \text{ for all } i \in S_1, \\ \boldsymbol{w}^\top\boldsymbol{x}_j \leq 0, \text{ for all } j \in S_2, \end{aligned}$$

where we denote $S_1 = \{i : y_i > 0, \text{ for all } i \in [n]\}$, $S_2 = \{j : y_j \leq 0, \text{ for all } j \in [n]\}$ and $S_1 \cup S_2 = [n]$. Since $n \leq d$ and we have assumed $\operatorname{rank}(\boldsymbol{X}) = n$, we can always find a feasible solution satisfying all n equality constraints. This implies that the solution set is nonempty, and $\boldsymbol{w}^\star$ always exists. Hence, the following KKT conditions are necessary (and also sufficient) to $\boldsymbol{w}^\star$ for some $\boldsymbol{\lambda}^\star \in \mathbb{R}^{|S_1|}$ and $\boldsymbol{\mu}^\star \in \mathbb{R}^{|S_2|}$:

Stationarity:

$$\boldsymbol{w}^\star + \sum_{i \in S_1} \lambda_i^\star \boldsymbol{x}_i + \sum_{j \in S_2} \mu_j^\star \boldsymbol{x}_j = 0 \Leftrightarrow \boldsymbol{w}^\star = -\sum_{i \in S_1} \lambda_i^\star \boldsymbol{x}_i - \sum_{j \in S_2} \mu_j^\star \boldsymbol{x}_j.$$

Primal feasibility:

$$\begin{aligned} \boldsymbol{w}^{\star\top} \boldsymbol{x}_i &= y_i, \text{ for all } i \in S_1, \\ \boldsymbol{w}^{\star\top} \boldsymbol{x}_j &\leq 0, \text{ for all } j \in S_2. \end{aligned}$$

Dual feasibility:

$$\begin{aligned} \lambda_i^\star &\in \mathbb{R}, \text{ for all } i \in S_1, \\ \mu_j^\star &\geq 0, \text{ for all } j \in S_2. \end{aligned}$$

Complementary slackness:

$$\sum_{j \in S_2} \mu_j^\star \left(\boldsymbol{w}^{\star\top} \boldsymbol{x}_j \right) = 0.$$

Next, we further denote a subset $\tilde{S}_2 \subseteq S_2$ such that $\tilde{S}_2 = \{j : \mu_j^\star > 0 \text{ for all } j \in S_2\}$ (note that $\tilde{S}_2$ can be empty). By the KKT conditions, it is necessary for $\boldsymbol{w}^\star$ to satisfy the following:

$$\boldsymbol{w}^\star = -\sum_{i \in S_1} \lambda_i^\star \boldsymbol{x}_i - \sum_{j \in \tilde{S}_2} \mu_j^\star \boldsymbol{x}_j, \text{ with } \lambda_i^\star \in \mathbb{R} \text{ and } \mu_j^\star > 0, \tag{14a}$$

$$\boldsymbol{w}^{\star\top} \boldsymbol{x}_i = y_i, \text{ for all } i \in S_1, \tag{14b}$$

$$\boldsymbol{w}^{\star\top} \boldsymbol{x}_j = 0, \text{ for all } j \in \tilde{S}_2. \tag{14c}$$

Now, we consider a reduced convex program:

$$\begin{aligned} \tilde{\boldsymbol{w}} &\in \arg\min_{\boldsymbol{w}} \frac{1}{2} \|\boldsymbol{w}\|_2^2 \\ \text{s.t. } \boldsymbol{w}^\top \boldsymbol{x}_i &= y_i, \text{ for all } i \in S_1, \\ \boldsymbol{w}^\top \boldsymbol{x}_j &= 0, \text{ for all } j \in \tilde{S}_2. \end{aligned} \tag{15}$$

Its KKT conditions are give below.

Stationarity:

$$\tilde{\boldsymbol{w}} + \sum_{i \in S_1} \tilde{\lambda}_i \boldsymbol{x}_i + \sum_{j \in \tilde{S}_2} \tilde{\lambda}_j \boldsymbol{x}_j = 0 \Leftrightarrow \tilde{\boldsymbol{w}} = -\sum_{i \in S_1} \tilde{\lambda}_i \boldsymbol{x}_i - \sum_{j \in \tilde{S}_2} \tilde{\lambda}_j \boldsymbol{x}_j.$$

Primal feasibility:

$$\tilde{\boldsymbol{w}}^\top \boldsymbol{x}_i = y_i, \text{ for all } i \in S_1,$$
$$\tilde{\boldsymbol{w}}^\top \boldsymbol{x}_j = 0, \text{ for all } j \in \tilde{S}_2.$$

Dual feasibility:

$$\tilde{\lambda}_i \in \mathbb{R}, \text{ for all } i \in S_1,$$
$$\tilde{\lambda}_j \in \mathbb{R}, \text{ for all } j \in \tilde{S}_2.$$

Since $\boldsymbol{w}^\star$ satisfies all the conditions in Equation (14), it also satisfies the KKT conditions for the reduced convex program (15). Thus, $\boldsymbol{w}^\star$ is also the optimal solution of the reduced convex program. Finally, we have a closed-form solution for the reduced convex program such that $\boldsymbol{w}^\star = \tilde{\boldsymbol{w}} = \boldsymbol{w}_{\text{linear}-\text{MNI},S} = \boldsymbol{X}_S^\top(\boldsymbol{X}_S\boldsymbol{X}_S^\top)^{-1}\tilde{\boldsymbol{y}}_S$ where $S = S_1 \cup \tilde{S}_2$ and $\tilde{\boldsymbol{y}}_S$ denotes the corresponding label subvector with all negative entries replaced by zero. This completes the proof of the lemma. ■

D.2. Proof of Theorem 4 (High-dimensional Implicit Bias)

In this section, we present the proof of Theorem 4. For the single ReLU model ($m = 1$), the primal–dual gradient update in (4) simplifies to

$$\text{(Primal)} \qquad \boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\beta}^{(t)})(\boldsymbol{\beta}^{(t)} - \boldsymbol{y}), \tag{16a}$$

$$\text{(Dual)} \qquad \boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} - \eta \boldsymbol{D}(\boldsymbol{\beta}^{(t)})(\boldsymbol{\beta}^{(t)} - \boldsymbol{y}). \tag{16b}$$

Before proceeding to the proof, we introduce a set of sufficient conditions under which the signs of the primal variables agree with the signs of the labels at iteration t. Moreover, these conditions are preserved at iteration $t+1$.

Lemma 14 *Under Assumptions 1 and 2, suppose the gradient descent step size satisfies $\eta \le \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$. For any single ReLU model, if the following six conditions hold at some iteration $t \ge 0$, then they also hold at iteration $t+1$.*

a. $\beta_i^{(t)} > 0$, *for all $i \in [n]$ with $y_i > 0$.*

b. $-\frac{3y_{\max}}{C_g\|\boldsymbol{\lambda}\|_1} \le \alpha_j^{(t)} \le -\frac{y_{\min}}{C_\alpha\|\boldsymbol{\lambda}\|_1}$, *for all $j \in [n]$ with $y_j < 0$.*

c. $\left\|\boldsymbol{\beta}_+^{(t)} - \boldsymbol{y}_+\right\|_2 \le C_y \left\|\boldsymbol{y}_+\right\|_2$.

d. $\left\|\boldsymbol{\alpha}^{(t)}\right\|_2 \le \frac{C_\alpha\sqrt{n}y_{\max}}{\|\boldsymbol{\lambda}\|_1}$.

e. $\beta_j^{(t)} \le 0$, *for all $j \in [n]$ with $y_j < 0$.*

f. $\sigma(\boldsymbol{\beta}^{(t)}) = \begin{bmatrix} \boldsymbol{\beta}_+^{(t)} \\ \boldsymbol{0} \end{bmatrix}$.

Consequently, the set of active examples consists exactly of the positively labeled examples, and the activation pattern remains unchanged, i.e., $\boldsymbol{D}(\boldsymbol{\beta}^{(t)}) = \boldsymbol{D}(\boldsymbol{\beta}^{(t+1)})$.

Proof (Lemma 14) In the following, we show that if the six sufficient conditions hold at some iteration $t \geq 0$, then they also hold at iteration $t+1$.

Part (a): By conditions (c) and (f) at iteration t, we have $\left\|h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}\right\|_2^2 = \left\|\sigma(\boldsymbol{\beta}^{(t)}) - \boldsymbol{y}\right\|_2^2 = \left\|\boldsymbol{\beta}_+^{(t)} - \boldsymbol{y}_+\right\|_2^2 + \left\|\boldsymbol{y}_-\right\|_2^2 \leq C_y^2 \left\|\boldsymbol{y}\right\|_2^2$. Together with $h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{x}_i) = \beta_i^{(t)}$ and condition (a), all the assumptions of Lemma 9 are satisfied for all i with $y_i > 0$. Consequently, we obtain $\beta_i^{(t+1)} > 0$ for all $i \in [n]$ with $y_i > 0$, and thus condition (a) holds at iteration $t+1$.

Part (b): According to the dual gradient update in Equation (16b), and using condition (e) at iteration t, we conclude that the dual variables corresponding to negatively labeled examples remain unchanged, i.e., $\alpha_j^{(t+1)} = \alpha_j^{(t)}$ for all $j \in [n]$ with $y_j < 0$. Therefore, condition (b) continues to hold at iteration $t+1$.

Part (c): By conditions (a) and (e), the gradient update at iteration t depends only on the positively labeled examples. Consequently, the update is equivalent to a linear regression gradient descent step using only the positive-labeled subset. As similarly argued in the proof of Lemma 2, since the step size satisfies $\eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$, the squared loss is monotonically non-increasing, and we obtain $\left\|\boldsymbol{\beta}_+^{(t+1)} - \boldsymbol{y}_+\right\|_2 \leq \left\|\boldsymbol{\beta}_+^{(t)} - \boldsymbol{y}_+\right\|_2 \leq C_y \left\|\boldsymbol{y}_+\right\|_2$ by condition (c) at iteration t. Therefore, condition (c) holds at iteration $t+1$.

Part (d): For this part, we use conditions (b) and (c) at iteration $t+1$. By the triangle inequality, we have

$$\left\|\boldsymbol{\alpha}^{(t+1)}\right\|_2 \leq \left\|\boldsymbol{\alpha}_+^{(t+1)}\right\|_2 + \left\|\boldsymbol{\alpha}_-^{(t+1)}\right\|_2.$$

By condition (b) at iteration $t+1$, it follows that $\left\|\boldsymbol{\alpha}_-^{(t+1)}\right\|_2 \leq \frac{3\sqrt{n}y_{\max}}{C_g\|\boldsymbol{\lambda}\|_1}$. It therefore remains to upper bound $\left\|\boldsymbol{\alpha}_+^{(t+1)}\right\|_2$. By condition (c) at iteration $t+1$, we have $\left\|\boldsymbol{\beta}_+^{(t+1)}\right\|_2 \leq C_y \left\|\boldsymbol{y}_+\right\|_2 + \left\|\boldsymbol{y}_+\right\|_2 \leq (C_y+1)\left\|\boldsymbol{y}\right\|_2$. Moreover, we have

$$\begin{aligned}
\left\|\boldsymbol{\beta}_+^{(t+1)}\right\|_2 &= \left\|\boldsymbol{X}_+\boldsymbol{X}^\top\boldsymbol{\alpha}^{(t+1)}\right\|_2 \\
&= \left\|\boldsymbol{X}_+\left[\boldsymbol{X}_+^\top \boldsymbol{X}_-^\top\right]\begin{bmatrix}\boldsymbol{\alpha}_+^{(t+1)} \\ \boldsymbol{\alpha}_-^{(t+1)}\end{bmatrix}\right\|_2 \\
&= \left\|\boldsymbol{X}_+\boldsymbol{X}_+^\top\boldsymbol{\alpha}_+^{(t+1)} + \boldsymbol{X}_+\boldsymbol{X}_-^\top\boldsymbol{\alpha}_-^{(t+1)}\right\|_2.
\end{aligned}$$

Applying the triangle inequality yields

$$\left\|\boldsymbol{X}_+\boldsymbol{X}_+^\top\boldsymbol{\alpha}_+^{(t+1)}\right\|_2 \leq \left\|\boldsymbol{\beta}_+^{(t+1)}\right\|_2 + \left\|\boldsymbol{X}_+\boldsymbol{X}_-^\top\boldsymbol{\alpha}_-^{(t+1)}\right\|_2 \leq (C_y+1)\left\|\boldsymbol{y}\right\|_2 + \left\|\boldsymbol{X}_+\boldsymbol{X}_-^\top\boldsymbol{\alpha}_-^{(t+1)}\right\|_2.$$

Since $\boldsymbol{X}_+\boldsymbol{X}_+^\top \in \mathbb{R}^{n_+\times n_+}$ is full rank, we obtain

$$\left\|\boldsymbol{\alpha}_+^{(t+1)}\right\|_2 \leq \frac{(C_y+1)\left\|\boldsymbol{y}\right\|_2 + \left\|\boldsymbol{X}_+\boldsymbol{X}_-^\top\boldsymbol{\alpha}_-^{(t+1)}\right\|_2}{\mu_{n_+}(\boldsymbol{X}_+\boldsymbol{X}_+^\top)}.$$

For the denominator, the variational formulation for eigenvalues of a submatrix and Lemma 11 imply that

$$\mu_{n_+}(\boldsymbol{X}_+\boldsymbol{X}_+^\top) \geq \mu_n(\boldsymbol{X}\boldsymbol{X}^\top) \geq \frac{1}{C_g}\sum_{j=1}^d \lambda_j = \frac{\|\boldsymbol{\lambda}\|_1}{C_g},$$

with probability at least $1-2e^{-n/C_g}$. For the numerator, we have $(C_y+1)\left\|\boldsymbol{y}\right\|_2 \leq (C_y+1)\sqrt{n}y_{\max}$. Moreover, by Bhatia and Kittaneh [2, Theorem 1], we have

$$\begin{aligned}\left\|\boldsymbol{X}_+\boldsymbol{X}_-^\top\right\|_2 &\leq \frac{1}{2}\left\|\boldsymbol{X}_+\boldsymbol{X}_+^\top + \boldsymbol{X}_-\boldsymbol{X}_-^\top\right\|_2 \\ &\leq \frac{1}{2}\left(\left\|\boldsymbol{X}_+\boldsymbol{X}_+^\top\right\|_2 + \left\|\boldsymbol{X}_-\boldsymbol{X}_-^\top\right\|_2\right) \\ &\leq C_g\sum_{j=1}^d \lambda_j \\ &= C_g\left\|\boldsymbol{\lambda}\right\|_1,\end{aligned}$$

where the last inequality follows from Lemma 11. Combining these bounds yields

$$\left\|\boldsymbol{\alpha}_+^{(t+1)}\right\|_2 \leq \frac{(C_y+1)\sqrt{n}y_{\max} + C_g\left\|\boldsymbol{\lambda}\right\|_1 \cdot \frac{3\sqrt{n}y_{\max}}{C_g\|\boldsymbol{\lambda}\|_1}}{\left\|\boldsymbol{\lambda}\right\|_1/C_g} = ((C_y+1)C_g + 3C_g)\frac{\sqrt{n}y_{\max}}{\left\|\boldsymbol{\lambda}\right\|_1}.$$

Consequently, we have

$$\left\|\boldsymbol{\alpha}^{(t+1)}\right\|_2 \leq ((C_y+1)C_g+3C_g)\frac{\sqrt{n}y_{\max}}{\left\|\boldsymbol{\lambda}\right\|_1} + \frac{3\sqrt{n}y_{\max}}{C_g\left\|\boldsymbol{\lambda}\right\|_1} \leq \frac{C_\alpha\sqrt{n}y_{\max}}{\left\|\boldsymbol{\lambda}\right\|_1},$$

with $C_\alpha \gtrsim \max\{C_g^2, C_yC_g\}$, and thus condition (d) holds at iteration $t+1$.

Part (e): By Lemma 10, and since conditions (b) and (d) hold at iteration $t+1$, we conclude that $\beta_j^{(t+1)} \leq 0$ for all $j \in [n]$ with $y_j < 0$. Thus, condition (e) holds at iteration $t+1$.

Part (f): By conditions (a) and (e) at iteration $t+1$, the signs of the primal variables continue to agree with the signs of the labels. Consequently, $\sigma(\boldsymbol{\beta}^{(t+1)}) = \begin{bmatrix}\boldsymbol{\beta}_+^{(t+1)} \\ \boldsymbol{0}\end{bmatrix}$, and thus condition (f) holds at iteration $t+1$.

We have shown that the six sufficient conditions hold at iteration $t+1$. Consequently, the signs of the primal variables continue to agree with the signs of the labels, and hence $\boldsymbol{D}(\boldsymbol{\beta}^{(t)}) = \boldsymbol{D}(\boldsymbol{\beta}^{(t+1)})$. This completes the proof. ■

Equipped with Lemma 14, we are now ready to prove Theorem 4.

Proof (Theorem 4) In the proof, we first show that after the first gradient step, the iterate at $t = 1$ satisfies the conditions in Lemma 14. Next, since the conditions hold at $t = 1$ and are preserved from $t = \tilde{t}$ to $t = \tilde{t} + 1$ by Lemma 14, we fully characterize the gradient descent dynamics by induction.

We begin by verifying that the iterate at $t = 1$ satisfies the sufficient conditions in Lemma 14. With the initialization $\boldsymbol{w}^{(0)} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{\epsilon}$, we have $\boldsymbol{\beta}^{(0)} = \boldsymbol{X}\boldsymbol{w}^{(0)} = \boldsymbol{\epsilon}$. Therefore, using the primal gradient update in Equation (16a), we obtain

$$\begin{aligned}\boldsymbol{\beta}^{(1)} &= \boldsymbol{\beta}^{(0)} - \eta\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{D}(\boldsymbol{\beta}^{(0)})(\boldsymbol{\beta}^{(0)} - \boldsymbol{y}) \\ &= \boldsymbol{\epsilon} - \eta\boldsymbol{X}\boldsymbol{X}^\top(\boldsymbol{\epsilon} - \boldsymbol{y}) \\ &= \boldsymbol{X}\boldsymbol{X}^\top\left[\underbrace{\eta\left(\boldsymbol{y} - \boldsymbol{\epsilon} + \frac{1}{\eta}\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{\epsilon}\right)}_{=:\boldsymbol{\alpha}^{(1)}}\right]. \end{aligned} \tag{17}$$

We denote $\boldsymbol{\alpha}^{(1)} := \eta\left(\boldsymbol{y} - \boldsymbol{\epsilon} + \frac{1}{\eta}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{\epsilon}\right)$ according to the primal-dual formulation $\boldsymbol{\beta}^{(1)} = \boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{\alpha}^{(1)}$ in Equation (3). In the below, we show that at iteration $t = 1$, the variables $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\alpha}^{(1)}$ satisfy all the conditions in Lemma 14.

Part (a): For all $i \in [n]$ with $y_i > 0$, we apply Lemma 9. Since $\beta_i^{(0)} = \epsilon_i > 0$, $\beta_i^{(0)} = h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{x}_i)$ and $\left\|\sigma(\boldsymbol{\beta}^{(0)}) - \boldsymbol{y}\right\|_2 \leq \|\boldsymbol{\epsilon}\|_2 + \|\boldsymbol{y}\|_2 \leq \frac{\sqrt{n}}{C_\alpha} y_{\min} + \|\boldsymbol{y}\|_2 \leq C_y\|\boldsymbol{y}\|_2$ with $C_y > 1 + \frac{1}{C_\alpha}$, it follows that $\beta_i^{(1)} > 0$ for all $i \in [n]$ with $y_i > 0$.

Part (b): For all $j \in [n]$ with $y_j < 0$, we verify that $\alpha_j^{(1)}$ satisfies the required upper and lower bounds. For the upper bound, recall that

$$\begin{aligned}\alpha_j^{(1)} &= \eta\left(y_j - \epsilon_j + \frac{1}{\eta}\boldsymbol{e}_j^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{\epsilon}\right) \\ &= \eta\left(y_j - \epsilon_j + \frac{1}{\eta}\boldsymbol{e}_j^\top\left[\frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I} + \left(\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I}\right)\right]\boldsymbol{\epsilon}\right) \\ &= \eta\left(y_j - \epsilon_j + \frac{\epsilon_j}{\eta\|\boldsymbol{\lambda}\|_1} + \frac{1}{\eta}\boldsymbol{e}_j^\top\left(\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I}\right)\boldsymbol{\epsilon}\right) \\ &\overset{\text{(i)}}{\leq} \eta\left(y_j + \frac{\epsilon_j}{\eta\|\boldsymbol{\lambda}\|_1} + \frac{1}{\eta}\boldsymbol{e}_j^\top\left(\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I}\right)\boldsymbol{\epsilon}\right) \\ &\overset{\text{(ii)}}{\leq} \eta\left(y_j + \frac{\epsilon_j}{\eta\|\boldsymbol{\lambda}\|_1} + \frac{1}{\eta}\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I}\right\|_2\|\boldsymbol{\epsilon}\|_2\right),\end{aligned}$$

where inequality (i) drops the negative term $-\epsilon_j$, and inequality (ii) follows from the submultiplicativity of the operator norm. By Corollary 13, we have

$$\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I}\right\|_2 \leq \frac{C_g C}{\|\boldsymbol{\lambda}\|_1}\cdot\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right),$$

with probability at least $1 - 2\exp(-n(Cc - \ln 9))$. Moreover, by the theorem assumptions, $\|\boldsymbol{\epsilon}\|_2 \le \frac{\sqrt{n}}{C_\alpha} y_{\min}$ and $\frac{1}{\eta} \le CC_g \|\boldsymbol{\lambda}\|_1$. Combining these bounds yields

$$\begin{aligned}
\alpha_j^{(1)} &\le \frac{1}{CC_g \|\boldsymbol{\lambda}\|_1}\left(-y_{\min} + \frac{CC_g}{C_\alpha} y_{\min} + C^2 C_g^2 \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right) \cdot \frac{\sqrt{n}}{C_\alpha} y_{\min}\right) \\
&\le \frac{1}{CC_g \|\boldsymbol{\lambda}\|_1}\left(-y_{\min} + \frac{CC_g}{C_\alpha} y_{\min} + C^2 C_g^2 \cdot \frac{y_{\min}}{C_0 y_{\max}} \cdot \frac{1}{C_\alpha} y_{\min}\right) \\
&= -\frac{y_{\min}}{C_\alpha \|\boldsymbol{\lambda}\|_1}\left(\frac{C_\alpha}{CC_g} - 1 - \frac{CC_g y_{\min}}{C_0 y_{\max}}\right) \\
&\le -\frac{y_{\min}}{C_\alpha \|\boldsymbol{\lambda}\|_1}.
\end{aligned}$$

The second inequality follows from $d_2 \ge C_0^2 \frac{n^2 y_{\max}^2}{y_{\min}^2}$ and $d_\infty \ge C_0 \frac{n^{1.5} y_{\max}}{y_{\min}}$ in Assumption 2, and the last inequality uses the following relationships between constants: $C_0 > C \cdot C_\alpha^2$ and $C_\alpha > C \cdot \max\{C_g^2, C_y C_g\}$. For the lower bound, we have

$$\begin{aligned}
\alpha_j^{(1)} &= \eta\left(y_j - \epsilon_j + \frac{\epsilon_j}{\eta \|\boldsymbol{\lambda}\|_1} + \frac{1}{\eta} \boldsymbol{e}_j^\top \left(\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I}\right)\boldsymbol{\epsilon}\right) \\
&\ge \eta\left(-y_{\max} - \epsilon_j - \frac{1}{\eta}\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I}\right\|_2 \|\boldsymbol{\epsilon}\|_2\right) \\
&\ge \frac{1}{C_g \|\boldsymbol{\lambda}\|_1}\left(-y_{\max} - \frac{1}{C_\alpha} y_{\min} - C^2 C_g^2 \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right) \cdot \frac{\sqrt{n}}{C_\alpha} y_{\min}\right) \\
&\ge \frac{1}{C_g \|\boldsymbol{\lambda}\|_1}\left(-y_{\max} - \frac{1}{C_\alpha} y_{\min} - C^2 C_g^2 \cdot \frac{y_{\min}}{C_0 y_{\max}} \cdot \frac{1}{C_\alpha} y_{\min}\right) \\
&\ge \frac{-3 y_{\max}}{C_g \|\boldsymbol{\lambda}\|_1},
\end{aligned}$$

by the same arguments. Thus, $\alpha_j^{(1)}$ satisfies both the required upper and lower bounds for all j with $y_j < 0$.

Part (c): We now verify that the primal variables corresponding to positively labeled examples minus $\boldsymbol{y}_+$ satisfy the norm bound in Lemma 14. Specifically, we show that $\left\|\boldsymbol{\beta}_+^{(1)} - \boldsymbol{y}_+\right\|_2^2 \le C_y^2 \|\boldsymbol{y}_+\|_2^2$. According to Equation (17), we have

$$\begin{aligned}
\left\|\boldsymbol{\beta}_+^{(1)} - \boldsymbol{y}_+\right\|_2^2 &= \sum_{i: y_i > 0} \left(\beta_i^{(1)} - y_i\right)^2 \\
&= \sum_{i: y_i > 0} \Bigg(\underbrace{\epsilon_i - \eta \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top (\boldsymbol{\epsilon} - \boldsymbol{y}) - y_i}_{=: T_i}\Bigg)^2. \qquad (18)
\end{aligned}$$

Next, we bound the term $T_i := \epsilon_i - \eta \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top (\boldsymbol{\epsilon} - \boldsymbol{y}) - y_i$ for all $i \in [n]$ with $y_i > 0$. We have

$$T_i = \epsilon_i - \eta \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top (\boldsymbol{\epsilon} - \boldsymbol{y}) - y_i = (\epsilon_i - y_i) - \eta \boldsymbol{e}_i^\top \left[\|\boldsymbol{\lambda}\|_1 \boldsymbol{I} + \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I} \right) \right] (\boldsymbol{\epsilon} - \boldsymbol{y})$$
$$= (1 - \eta \|\boldsymbol{\lambda}\|_1)(\epsilon_i - y_i) - \eta \boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I} \right) (\boldsymbol{\epsilon} - \boldsymbol{y}).$$

Since the step size assumption guarantees that $\frac{1}{CC_g\|\boldsymbol{\lambda}\|_1} \leq \eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$, and $\epsilon_i \leq \frac{1}{C_\alpha} y_{\min}$, the term $(1 - \eta \|\boldsymbol{\lambda}\|_1)(\epsilon_i - y_i)$ is strictly negative. Hence, in order to upper bound T_i^2, it suffices to find the lower bound for T_i. We have

$$T_i = (1 - \eta \|\boldsymbol{\lambda}\|_1)(\epsilon_i - y_i) - \eta \boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I} \right) (\boldsymbol{\epsilon} - \boldsymbol{y})$$
$$\geq -y_i - \eta \left\| \boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I} \right\|_2 \|\boldsymbol{\epsilon} - \boldsymbol{y}\|_2,$$

where the inequality drops the positive terms $(1-\eta \|\boldsymbol{\lambda}\|_1)\epsilon_i$ and $\eta \|\boldsymbol{\lambda}\|_1 y_i$. We again upper bound $\left\| \boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I} \right\|_2$ by Corollary 13. With probability at least $1-2\exp(-n(Cc - \ln 9))$, we have

$$T_i \geq -y_i - \eta \cdot C \|\boldsymbol{\lambda}\|_1 \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty} \right) \|\boldsymbol{\epsilon} - \boldsymbol{y}\|_2$$
$$\geq -y_i - \frac{C}{C_g} \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty} \right) \|\boldsymbol{\epsilon} - \boldsymbol{y}\|_2,$$

by applying $\eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$. Finally, we apply the upper bounds for $\|\boldsymbol{\epsilon}\|_2$ and $\|\boldsymbol{y}\|_2$, and Assumption 2 ensures that $d_2 \geq C_0^2 \frac{n^2 y_{\max}^2}{y_{\min}^2}$ and $d_\infty \geq C_0 \frac{n^{1.5} y_{\max}}{y_{\min}}$. We have

$$T_i \geq -y_i - \frac{C}{C_g} \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty} \right) \left(\frac{\sqrt{n}}{C_\alpha} y_{\min} + \sqrt{n} y_{\max} \right)$$
$$\geq -y_i - \frac{C y_{\min}}{C_g C_0 y_{\max}} \left(\frac{1}{C_\alpha} y_{\min} + y_{\max} \right)$$
$$\geq -y_i \left(1 + \frac{2C}{C_g C_0} \right)$$
$$\geq -C_y y_i,$$

with the choice of $C_y \geq 2$. Substituting $T_i^2 \leq C_y^2 y_i^2$ into Equation (18), we have

$$\left\| \boldsymbol{\beta}_+^{(1)} - \boldsymbol{y}_+ \right\|_2^2 \leq \sum_{i: y_i > 0} C_y^2 y_i^2 = C_y^2 \|\boldsymbol{y}_+\|_2^2.$$

As a result, we conclude that $\left\| \boldsymbol{\beta}_+^{(1)} - \boldsymbol{y}_+ \right\|_2 \leq C_y \|\boldsymbol{y}_+\|_2$ as required.

Part (d): We next verify that $\boldsymbol{\alpha}^{(1)}$ satisfies the required norm bound. Recall that

$$\boldsymbol{\alpha}^{(1)} = \eta \left(\boldsymbol{y} - \boldsymbol{\epsilon} + \frac{1}{\eta} \left(\boldsymbol{X}\boldsymbol{X}^\top \right)^{-1} \boldsymbol{\epsilon} \right).$$

Taking the ℓ_2 norm and applying the triangle inequality yields

$$\left\|\boldsymbol{\alpha}^{(1)}\right\|_2 = \left\|\eta\left(\boldsymbol{y} - \boldsymbol{\epsilon} + \frac{1}{\eta}\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{\epsilon}\right)\right\|_2$$
$$\leq \eta\left[\|\boldsymbol{y}\|_2 + \|\boldsymbol{\epsilon}\|_2 + \frac{1}{\eta}\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\right\|_2\|\boldsymbol{\epsilon}\|_2\right].$$

We now bound each term on the right-hand side. We apply the label bound, $\|\boldsymbol{y}\|_2 \leq \sqrt{n}y_{\max}$ and the construction of the initialization, $\|\boldsymbol{\epsilon}\|_2 \leq \frac{\sqrt{n}}{C_\alpha}y_{\min}$. Moreover, Lemma 11 implies $\left\|(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\right\|_2 \leq \frac{C_g}{\|\boldsymbol{\lambda}\|_1}$ with probability at least $1 - 2e^{-n/C_g}$, and the step size condition ensures $\frac{1}{CC_g\|\boldsymbol{\lambda}\|_1} \leq \eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$. Substituting these bounds, we obtain

$$\left\|\boldsymbol{\alpha}^{(1)}\right\|_2 \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}\left[\sqrt{n}y_{\max} + \frac{\sqrt{n}}{C_\alpha}y_{\min} + CC_g\|\boldsymbol{\lambda}\|_1 \cdot \frac{C_g}{\|\boldsymbol{\lambda}\|_1} \cdot \frac{\sqrt{n}}{C_\alpha}y_{\min}\right]$$
$$\leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}\left(3\sqrt{n}y_{\max}\right)$$
$$\leq \frac{C_\alpha\sqrt{n}y_{\max}}{\|\boldsymbol{\lambda}\|_1},$$

with $C_\alpha \gtrsim \max\{C_g^2, C_yC_g\}$. Therefore, $\boldsymbol{\alpha}^{(1)}$ satisfies the required norm bound.

Part (e): Since we have shown that $\alpha_j^{(1)} \leq -\frac{y_{\min}}{C_\alpha\|\boldsymbol{\lambda}\|_1}$ and $\left\|\boldsymbol{\alpha}^{(1)}\right\|_2 \leq \frac{C_\alpha\sqrt{n}y_{\max}}{\|\boldsymbol{\lambda}\|_1}$ for all $j \in [n]$ with $y_j < 0$, it follows from Lemma 10 that $\beta_j^{(1)} \leq 0$ for all $j \in [n]$ with $y_j < 0$.

Part (f): Since we have shown that $\beta_i^{(1)} > 0$ for all $i \in [n]$ with $y_i > 0$ and $\beta_j^{(1)} \leq 0$ for all $j \in [n]$ with $y_j < 0$, the signs of the primal variables coincide with the signs of the labels. Consequently, $\sigma(\boldsymbol{\beta}^{(1)}) = \begin{bmatrix}\boldsymbol{\beta}_+^{(1)} \\ \mathbf{0}\end{bmatrix}$.

We have shown that at iteration $t = 1$, all conditions in Lemma 14 are satisfied. Consequently, all positively labeled examples are active, while all negatively labeled examples are inactive. We now complete the proof by induction and characterize the gradient descent dynamics for all subsequent iterations. By Lemma 14, since the conditions hold at $t = 1$, they also hold at $t = 2$. More generally, the same lemma implies that if the conditions hold at $t = \tilde{t}$ then they continue to hold at $t = \tilde{t} + 1$. This completes the induction argument.

As a result, for all $t \geq 1$, the activation pattern remains fixed, i.e., $\boldsymbol{D}(\boldsymbol{\beta}^{(t)}) = \boldsymbol{D}(\boldsymbol{\beta}^{(1)})$. By Lemma 1, the gradient descent dynamics from this point onward are equivalent to those of linear regression trained on the positively labeled examples, with initialization $\boldsymbol{w}^{(1)}$. Finally, by Lemma 2, the $\boldsymbol{w}^{(\infty)}$ satisfies

$$\boldsymbol{w}^{(\infty)} = \underset{\boldsymbol{w}\in\{\boldsymbol{w}:\boldsymbol{X}_+\boldsymbol{w}=\boldsymbol{y}_+\}}{\arg\min}\left\|\boldsymbol{w} - \boldsymbol{w}^{(1)}\right\|_2,$$

where we have $\boldsymbol{w}^{(1)} = \eta\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{\epsilon} + \frac{1}{\eta}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{\epsilon}\right)$. This completes the proof of Theorem 4. ■

D.3. Proof of Theorem 5 (Implicit Bias Approximation to $w^\star$)

In this section, we present the proof of implicit bias approximation to $\boldsymbol{w}^\star$ for single ReLU models.
Proof (Theorem 5) We restate the definition of $\boldsymbol{w}^\star$ in Equation (6) below.

$$\begin{aligned} \boldsymbol{w}^\star &= \arg\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ \text{s.t. } \boldsymbol{w}^\top \boldsymbol{x}_i &= y_i, \text{ for all } y_i > 0 \\ \boldsymbol{w}^\top \boldsymbol{x}_j &\le 0, \text{ for all } y_j \le 0. \end{aligned}$$

Recall that the gradient descent limit $\boldsymbol{w}^{(\infty)}$ satisfies the same set of constraints: it interpolates all positively labeled examples and produces negative predictions for negatively labeled examples. Consequently, both $\boldsymbol{w}^{(\infty)}$ and $\boldsymbol{w}^\star$ are feasible solutions achieving the minimum empirical risk.

We start with showing the upper bound on $\|\boldsymbol{w}^{(\infty)} - \boldsymbol{w}^\star\|_2$. We first relate the distance between the predictors $\boldsymbol{w}^{(\infty)}$ and $\boldsymbol{w}^\star$ to the distance in their predictions, i.e., $\|\boldsymbol{X}\boldsymbol{w}^{(\infty)} - \boldsymbol{X}\boldsymbol{w}^\star\|_2$. Since both vectors lie in the span of the data $\{\boldsymbol{x}_i\}_{i=1}^n$, their difference has no component in the null space corresponding to the smallest $d-n$ eigenvalues of $\boldsymbol{X}^\top\boldsymbol{X}$. Therefore, we have

$$\left\|\boldsymbol{X}\boldsymbol{w}^{(\infty)} - \boldsymbol{X}\boldsymbol{w}^\star\right\|_2^2 = \left\|\boldsymbol{X}\left(\boldsymbol{w}^{(\infty)} - \boldsymbol{w}^\star\right)\right\|_2^2 \ge \mu_n(\boldsymbol{X}^\top\boldsymbol{X})\left\|\boldsymbol{w}^{(\infty)} - \boldsymbol{w}^\star\right\|_2^2 = \mu_n(\boldsymbol{X}\boldsymbol{X}^\top)\left\|\boldsymbol{w}^{(\infty)} - \boldsymbol{w}^\star\right\|_2^2. \tag{19}$$

As a result, to derive an upper bound for $\|\boldsymbol{w}^{(\infty)} - \boldsymbol{w}^\star\|_2$, it suffices to upper bound the distance between their prediction $\|\boldsymbol{X}\boldsymbol{w}^{(\infty)} - \boldsymbol{X}\boldsymbol{w}^\star\|_2$. We begin with analyzing $\boldsymbol{w}^{(\infty)}$. By Theorem 4, $\boldsymbol{w}^{(\infty)}$ satisfies the following:

$$\begin{aligned} \boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i &= y_i && \text{for all } y_i > 0, \\ \alpha_j^{(\infty)} = \alpha_j^{(1)} &= \eta\left(y_j - \epsilon_j + \frac{1}{\eta}\boldsymbol{e}_j^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{\epsilon}\right) && \text{for all } y_j < 0, \end{aligned}$$

and also all the conditions in Lemma 14. On the other hand, according to the necessary conditions in Equation (14) in Lemma 3, $\boldsymbol{w}^\star$ satisfies

$$\begin{aligned} \boldsymbol{w}^\star = -\sum_{i\in S_1}\lambda_i^\star \boldsymbol{x}_i - \sum_{j\in\tilde{S}_2}\mu_j^\star \boldsymbol{x}_j&, \text{ with } \lambda_i^\star \in \mathbb{R} \text{ and } \mu_j^\star > 0, \\ \boldsymbol{w}^{\star\top}\boldsymbol{x}_i = y_i&, \text{ for all } i \in S_1, \\ \boldsymbol{w}^{\star\top}\boldsymbol{x}_j = 0&, \text{ for all } j \in \tilde{S}_2, \end{aligned}$$

where we have denoted $S_1 = \{i : y_i > 0, \text{ for all } i \in [n]\}$, $S_2 = \{j : y_j \le 0, \text{ for all } j \in [n]\}$, $\tilde{S}_2 \subseteq S_2$ (note that $\tilde{S}_2$ can be empty) and $S = S_1 \cup \tilde{S}_2$. Based on these necessary conditions, we can define $\boldsymbol{w}^\star = \boldsymbol{X}^\top\boldsymbol{\alpha}^\star$ where

$$\alpha_i^\star = \begin{cases} -\lambda_i^\star & \text{for all } i \in S_1 \\ -\mu_i^\star & \text{for all } i \in \tilde{S}_2 \\ 0 & \text{for all } i \in S_2 \cup \tilde{S}_2^{\mathsf{c}} =: S_3 \end{cases}.$$

Let $\boldsymbol{X}_S \in \mathbb{R}^{|S| \times d}$ denote the submatrix of $\boldsymbol{X}$ consisting of the rows indexed by S (taken in increasing order), and let $\boldsymbol{y}_S \in \mathbb{R}^{|S|}$ denote the corresponding label subvector with all negative entries replaced by zero. We have

$$\boldsymbol{y}_S = \boldsymbol{X}_S \boldsymbol{X}_S^\top \boldsymbol{\alpha}_S^\star,$$

and similarly, by taking the norm and using the matrix norm lower bound of the smallest eigenvalue of $\boldsymbol{X}_S \boldsymbol{X}_S^\top$, we have

$$\begin{aligned} \|\boldsymbol{y}_S\|_2 &= \left\| \boldsymbol{X}_S \boldsymbol{X}_S^\top \boldsymbol{\alpha}_S^\star \right\|_2 \\ &\geq \mu_{|S|}(\boldsymbol{X}_S \boldsymbol{X}_S^\top) \, \|\boldsymbol{\alpha}_S^\star\|_2 \, . \end{aligned}$$

Consequently, we have

$$\|\boldsymbol{\alpha}^\star\|_2 = \|\boldsymbol{\alpha}_S^\star\|_2 \leq \frac{\|\boldsymbol{y}_S\|_2}{\mu_{|S|}(\boldsymbol{X}_S \boldsymbol{X}_S^\top)} \leq \frac{\sqrt{n} y_{\max}}{\mu_n(\boldsymbol{X}\boldsymbol{X}^\top)} \leq \frac{C_g \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1}, \tag{20}$$

where the second inequality follows from the variational formulation of submatrix, and the last inequality follows from Lemma 11 with probability at least $1 - 2e^{-n/C_g}$.

We know that for all $i \in S_1$, $\boldsymbol{w}^{(\infty)\top} \boldsymbol{x}_i = \boldsymbol{w}^{\star\top} \boldsymbol{x}_i = y_i$, and $\boldsymbol{w}^{\star\top} \boldsymbol{x}_j = 0$ for all $j \in \tilde{S}_2$. Therefore, we can write

$$\begin{aligned} \left\| \boldsymbol{X}\boldsymbol{w}^{(\infty)} - \boldsymbol{X}\boldsymbol{w}^\star \right\|_2^2 &= \sum_{i=1}^n \left(\boldsymbol{w}^{(\infty)\top} \boldsymbol{x}_i - \boldsymbol{w}^{\star\top} \boldsymbol{x}_i \right)^2 \\ &= \sum_{i \in \tilde{S}_2} \left(\boldsymbol{w}^{(\infty)\top} \boldsymbol{x}_i \right)^2 + \sum_{i \in S_3} \left(\boldsymbol{w}^{(\infty)\top} \boldsymbol{x}_i - \boldsymbol{w}^{\star\top} \boldsymbol{x}_i \right)^2 . \end{aligned} \tag{21}$$

We start with upper bounding the term $(\boldsymbol{w}^{(\infty)\top} \boldsymbol{x}_i)^2$ for all $i \in \tilde{S}_2$. Since $\boldsymbol{w}^{(\infty)\top} \boldsymbol{x}_i \leq 0$ by the conditions in Lemma 14, it suffices to lower bound $\boldsymbol{w}^{(\infty)\top} \boldsymbol{x}_i$. We have

$$\begin{aligned} \boldsymbol{w}^{(\infty)\top} \boldsymbol{x}_i &= \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\alpha}^{(\infty)} \\ &= \boldsymbol{e}_i^\top \left[\|\boldsymbol{\lambda}\|_1 \boldsymbol{I} + (\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}) \right] \boldsymbol{\alpha}^{(\infty)} \\ &= \|\boldsymbol{\lambda}\|_1 \alpha_i^{(\infty)} + \boldsymbol{e}_i^\top (\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}) \boldsymbol{\alpha}^{(\infty)} \\ &\geq \|\boldsymbol{\lambda}\|_1 \alpha_i^{(\infty)} - \left\| \boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I} \right\|_2 \left\| \boldsymbol{\alpha}^{(\infty)} \right\|_2 \\ &\geq \|\boldsymbol{\lambda}\|_1 \left[\alpha_i^{(\infty)} - C \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty} \right) \left\| \boldsymbol{\alpha}^{(\infty)} \right\|_2 \right], \end{aligned}$$

where the last inequality applies Corollary 13. Substituting the bounds of $\alpha_i^{(\infty)}$ and $\left\| \boldsymbol{\alpha}^{(\infty)} \right\|_2$ from Lemma 14, we have

$$\begin{aligned} \boldsymbol{w}^{(\infty)\top} \boldsymbol{x}_i &\geq \|\boldsymbol{\lambda}\|_1 \left[-\frac{3 y_{\max}}{C_g \|\boldsymbol{\lambda}\|_1} - C \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty} \right) \frac{C_\alpha \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1} \right] \\ &\geq \|\boldsymbol{\lambda}\|_1 \left[-\frac{3 y_{\max}}{C_g \|\boldsymbol{\lambda}\|_1} - C \cdot \frac{y_{\min}}{C_0 y_{\max}} \frac{C_\alpha y_{\max}}{\|\boldsymbol{\lambda}\|_1} \right] \\ &\geq -\frac{4}{C_g} y_{\max}, \end{aligned}$$

where the inequalities above substitute $d_2 \geq C_0^2 \frac{n^2 y_{\max}^2}{y_{\min}^2}$ and $d_\infty \geq C_0 \frac{n^{1.5} y_{\max}}{y_{\min}}$ in Assumption 2 with $C_0 \gtrsim C_\alpha^2$ and $C_\alpha \gtrsim \max\{C_g^2, C_y C_g\}$. Therefore, we have $(\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i)^2 \leq \frac{16}{C_g^2} y_{\max}^2$ for all $i \in \tilde{S}_2$. Next, we upper bound the term $(\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}^{\star\top}\boldsymbol{x}_i)^2$ for all $i \in S_3$. We use the key idea that $\alpha_i^\star = 0$ for all $i \in S_3$. We have

$$
\begin{aligned}
\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}^{\star\top}\boldsymbol{x}_i &= \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\alpha}^{(\infty)} - \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\alpha}^\star \\
&= \boldsymbol{e}_i^\top \left[\|\boldsymbol{\lambda}\|_1 \boldsymbol{I} + (\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}) \right] (\boldsymbol{\alpha}^{(\infty)} - \boldsymbol{\alpha}^\star) \\
&= \|\boldsymbol{\lambda}\|_1 \alpha_i^{(\infty)} + \boldsymbol{e}_i^\top (\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I})(\boldsymbol{\alpha}^{(\infty)} - \boldsymbol{\alpha}^\star) \\
&\geq \|\boldsymbol{\lambda}\|_1 \alpha_i^{(\infty)} - \left\| \boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I} \right\|_2 \left(\left\| \boldsymbol{\alpha}^{(\infty)} \right\|_2 + \|\boldsymbol{\alpha}^\star\|_2 \right) \\
&\geq \|\boldsymbol{\lambda}\|_1 \left[-\frac{3 y_{\max}}{C_g \|\boldsymbol{\lambda}\|_1} - C \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty} \right) \left(\frac{C_\alpha \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1} + \frac{C_g \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1} \right) \right] \\
&\geq \|\boldsymbol{\lambda}\|_1 \left[-\frac{3 y_{\max}}{C_g \|\boldsymbol{\lambda}\|_1} - C \cdot \frac{y_{\min}}{C_0 y_{\max}} \left(\frac{C_\alpha y_{\max}}{\|\boldsymbol{\lambda}\|_1} + \frac{C_g y_{\max}}{\|\boldsymbol{\lambda}\|_1} \right) \right] \\
&\geq -\frac{4}{C_g} y_{\max},
\end{aligned}
$$

by applying the same argument and noting from Equation (20) that $\|\boldsymbol{\alpha}^\star\|_2 \leq \frac{C_g \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1}$. Substituting the upper bounds into Equation (21) gives us

$$
\begin{aligned}
\left\| \boldsymbol{X}\boldsymbol{w}^{(\infty)} - \boldsymbol{X}\boldsymbol{w}^\star \right\|_2^2 &= \sum_{i \in \tilde{S}_2} \left(\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i \right)^2 + \sum_{i \in S_3} \left(\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}^{\star\top}\boldsymbol{x}_i \right)^2 \\
&\leq \sum_{i \in \tilde{S}_2} \frac{16}{C_g^2} y_{\max}^2 + \sum_{i \in S_3} \frac{16}{C_g^2} y_{\max}^2 \\
&= \frac{16}{C_g^2} n_- y_{\max}^2.
\end{aligned} \tag{22}
$$

Finally, putting together Equation (19) and (22), we have

$$
\left\| \boldsymbol{w}^{(\infty)} - \boldsymbol{w}^\star \right\|_2^2 \leq \frac{\left\| \boldsymbol{X}\boldsymbol{w}^{(\infty)} - \boldsymbol{X}\boldsymbol{w}^\star \right\|_2^2}{\mu_n(\boldsymbol{X}\boldsymbol{X}^\top)} \leq \frac{16 n_- y_{\max}^2}{C_g \|\boldsymbol{\lambda}\|_1},
$$

which completes the proof od the upper bound. Next, we derive the lower bound of $\|\boldsymbol{w}^{(\infty)} - \boldsymbol{w}^\star\|_2$ in a similar approach. We again start with the prediction distance, given by

$$
\left\| \boldsymbol{X}\boldsymbol{w}^{(\infty)} - \boldsymbol{X}\boldsymbol{w}^\star \right\|_2^2 = \left\| \boldsymbol{X} \left(\boldsymbol{w}^{(\infty)} - \boldsymbol{w}^\star \right) \right\|_2^2 \leq \mu_1(\boldsymbol{X}^\top \boldsymbol{X}) \left\| \boldsymbol{w}^{(\infty)} - \boldsymbol{w}^\star \right\|_2^2 = \mu_1(\boldsymbol{X}\boldsymbol{X}^\top) \left\| \boldsymbol{w}^{(\infty)} - \boldsymbol{w}^\star \right\|_2^2. \tag{23}
$$

It suffices to lower bound $\|\boldsymbol{X}\boldsymbol{w}^{(\infty)} - \boldsymbol{X}\boldsymbol{w}^\star\|_2$ to get the lower bound of $\|\boldsymbol{w}^{(\infty)} - \boldsymbol{w}^\star\|_2$. By Equation (21), we have

$$
\left\| \boldsymbol{X}\boldsymbol{w}^{(\infty)} - \boldsymbol{X}\boldsymbol{w}^\star \right\|_2^2 = \sum_{i \in \tilde{S}_2} \left(\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i \right)^2 + \sum_{i \in S_3} \left(\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}^{\star\top}\boldsymbol{x}_i \right)^2.
$$

Therefore, we need to lower bound $(\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i)^2$ for $i \in \tilde{S}_2$, and $(\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}^{\star\top}\boldsymbol{x}_i)^2$ for $i \in S_3$. For $\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i$, since $\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i < 0$, we have

$$\begin{aligned}
\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i &= \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\alpha}^{(\infty)} \\
&= \boldsymbol{e}_i^\top \left[\|\boldsymbol{\lambda}\|_1 \boldsymbol{I} + (\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I})\right] \boldsymbol{\alpha}^{(\infty)} \\
&= \|\boldsymbol{\lambda}\|_1 \alpha_i^{(\infty)} + \boldsymbol{e}_i^\top (\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I})\boldsymbol{\alpha}^{(\infty)} \\
&\le \|\boldsymbol{\lambda}\|_1 \alpha_i^{(\infty)} + \left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right\|_2 \left\|\boldsymbol{\alpha}^{(\infty)}\right\|_2 \\
&\le \|\boldsymbol{\lambda}\|_1 \left[\alpha_i^{(\infty)} + C \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right) \left\|\boldsymbol{\alpha}^{(\infty)}\right\|_2\right],
\end{aligned}$$

where the last inequality applies Corollary 13. Substituting the bounds of $\alpha_i^{(\infty)}$ and $\left\|\boldsymbol{\alpha}^{(\infty)}\right\|_2$ from Lemma 14, we have

$$\begin{aligned}
\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i &\le \|\boldsymbol{\lambda}\|_1 \left[-\frac{y_{\min}}{C_\alpha \|\boldsymbol{\lambda}\|_1} + C \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right) \frac{C_\alpha \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right] \\
&\le \|\boldsymbol{\lambda}\|_1 \left[-\frac{y_{\min}}{C_\alpha \|\boldsymbol{\lambda}\|_1} + C \cdot \frac{y_{\min}}{C_0 y_{\max}} \frac{C_\alpha y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right] \\
&\le -(1 - \frac{C \cdot C_\alpha^2}{C_0}) \frac{y_{\min}}{C_\alpha},
\end{aligned}$$

where the inequalities above substitute $d_2 \ge C_0^2 \frac{n^2 y_{\max}^2}{y_{\min}^2}$ and $d_\infty \ge C_0 \frac{n^{1.5} y_{\max}}{y_{\min}}$ in Assumption 2 with $C_0 \gtrsim C_\alpha^2$. Similarly, for $\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}^{\star\top}\boldsymbol{x}_i$, we have

$$\begin{aligned}
\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}^{\star\top}\boldsymbol{x}_i &= \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\alpha}^{(\infty)} - \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\alpha}^\star \\
&= \boldsymbol{e}_i^\top \left[\|\boldsymbol{\lambda}\|_1 \boldsymbol{I} + (\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I})\right] (\boldsymbol{\alpha}^{(\infty)} - \boldsymbol{\alpha}^\star) \\
&= \|\boldsymbol{\lambda}\|_1 \alpha_i^{(\infty)} + \boldsymbol{e}_i^\top (\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I})(\boldsymbol{\alpha}^{(\infty)} - \boldsymbol{\alpha}^\star) \\
&\le \|\boldsymbol{\lambda}\|_1 \alpha_i^{(\infty)} + \left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right\|_2 (\left\|\boldsymbol{\alpha}^{(\infty)}\right\|_2 + \|\boldsymbol{\alpha}^\star\|_2) \\
&\le \|\boldsymbol{\lambda}\|_1 \left[-\frac{y_{\min}}{C_\alpha \|\boldsymbol{\lambda}\|_1} + C \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right) \left(\frac{C_\alpha \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1} + \frac{C_g \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right)\right] \\
&\le \|\boldsymbol{\lambda}\|_1 \left[-\frac{y_{\min}}{C_\alpha \|\boldsymbol{\lambda}\|_1} + C \cdot \frac{y_{\min}}{C_0 y_{\max}} \left(\frac{C_\alpha y_{\max}}{\|\boldsymbol{\lambda}\|_1} + \frac{C_g y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right)\right] \\
&\le -(1 - \frac{2C \cdot C_\alpha^2}{C_0}) \frac{y_{\min}}{C_\alpha},
\end{aligned}$$

by applying the same argument and noting from Equation (20) that $\|\boldsymbol{\alpha}^\star\|_2 \leq \frac{C_g\sqrt{n}y_{\max}}{\|\boldsymbol{\lambda}\|_1}$. Substituting the lower bounds into Equation (21) gives us

$$\begin{aligned}
\left\|\boldsymbol{X}\boldsymbol{w}^{(\infty)} - \boldsymbol{X}\boldsymbol{w}^\star\right\|_2^2 &= \sum_{i\in\tilde{S}_2}\left(\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i\right)^2 + \sum_{i\in S_3}\left(\boldsymbol{w}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}^{\star\top}\boldsymbol{x}_i\right)^2 \\
&\geq \sum_{i\in\tilde{S}_2}(1-\frac{C\cdot C_\alpha^2}{C_0})^2\frac{y_{\min}^2}{C_\alpha^2} + \sum_{i\in S_3}(1-\frac{2C\cdot C_\alpha^2}{C_0})^2\frac{y_{\min}^2}{C_\alpha^2} \\
&\geq \left(1-\frac{2C\cdot C_\alpha^2}{C_0}\right)^2\frac{n_- y_{\min}^2}{C_\alpha^2} \\
&= \frac{n_- y_{\min}^2}{\tilde{C}},
\end{aligned} \tag{24}$$

where we let $\tilde{C} := \frac{C_0^2 C_\alpha^2}{(C_0 - 2C\cdot C_\alpha^2)^2} > 1$. Finally, putting together Equations (23) and (24), we have

$$\left\|\boldsymbol{w}^{(\infty)} - \boldsymbol{w}^\star\right\|_2^2 \geq \frac{\left\|\boldsymbol{X}\boldsymbol{w}^{(\infty)} - \boldsymbol{X}\boldsymbol{w}^\star\right\|_2^2}{\mu_1(\boldsymbol{X}\boldsymbol{X}^\top)} \geq \frac{n_- y_{\min}^2}{\tilde{C}C_g\|\boldsymbol{\lambda}\|_1}.$$

This completes the proof of the lower bound. ■

Appendix E. Proofs for the Two ReLU Model ($m = 2$) Trained with Gradient Descent

In this section, we present the proofs concerning the behavior of the 2-ReLU model trained with gradient descent.

E.1. Proof of Lemma 6 (Characterization of $w^\star$)

Proof (Lemma 6) We first show that the feasible set of (7) is nonempty. Define $\tilde{\boldsymbol{w}}_\oplus := \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}_\oplus$ and $\tilde{\boldsymbol{w}}_\ominus := \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}_\ominus$, where we define $y_{\oplus,i} := \max\{y_i, 0\}$ and $y_{\ominus,i} := -\min\{y_i, 0\}$. Then for all $i \in [n]$, we have $\sigma(\tilde{\boldsymbol{w}}_\oplus^\top \boldsymbol{x}_i) - \sigma(\tilde{\boldsymbol{w}}_\ominus^\top \boldsymbol{x}_i) = \sigma(y_{\oplus,i}) - \sigma(y_{\ominus,i}) = y_i$. Thus, $\{\tilde{\boldsymbol{w}}_\oplus, \tilde{\boldsymbol{w}}_\ominus\}$ is feasible, and the feasible set is nonempty.

Next, we show that any optimal solution of (7) corresponds to an optimal solution of (8). Let $\{\boldsymbol{w}_\oplus^\star, \boldsymbol{w}_\ominus^\star\}$ be an optimal solution of (7).

Case 1: $i \in S_+$ **(positive labels)**
For $i \in S_+$, since $\sigma(\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i) - \sigma(\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i) = y_i > 0$, we have

$$\sigma(\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i) = y_i + \sigma(\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i) \geq y_i > 0.$$

Hence, $\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i > 0$ and $\sigma(\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i) = \boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i$. There are two possible activation patterns:

- If $\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i \leq 0$, then we have $\sigma(\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i) - \sigma(\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i) = \boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i = y_i$.
- If $\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i \geq 0$, then we have $\sigma(\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i) - \sigma(\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i) = \boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i - \boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i = y_i$.

(Note that $\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i = 0$ is covered by both cases.)

Case 2: $i \in S_-$ **(negative labels)**
For $i \in S_-$, since $\sigma(\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i) - \sigma(\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i) = y_i < 0$, we obtain

$$\sigma(\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i) = -y_i + \sigma(\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i) \geq -y_i > 0,$$

which implies $\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i > 0$ and $\sigma(\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i) = \boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i$. Again, two activation patterns are possible:

- If $\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i \leq 0$, then we have $\sigma(\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i) - \sigma(\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i) = -\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i = y_i$.
- If $\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i \geq 0$, then we have $\sigma(\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i) - \sigma(\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i) = \boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i - \boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i = y_i$.

(Note that $\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i = 0$ is covered by both cases.)

Combining the two cases (in total four patterns), there exist disjoint partitions

$$S_1 \cup S_2 = S_+,\ S_1 \cap S_2 = \varnothing,\ \text{and}\ S_3 \cup S_4 = S_-,\ S_3 \cap S_4 = \varnothing,$$

such that the optimal solution $\{\boldsymbol{w}_\oplus^\star, \boldsymbol{w}_\ominus^\star\}$ satisfies

$$\begin{aligned}
&\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i = y_i, & \boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i \leq 0, \quad & \text{for all } i \in S_1,\\
&\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i - \boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i = y_i, & -\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i \leq 0, \quad & \text{for all } i \in S_2,\\
&-\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i = y_i, \quad \boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i \leq 0, & & \text{for all } i \in S_3,\\
&\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i - \boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i = y_i, -\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i \leq 0, & & \text{for all } i \in S_4.
\end{aligned}$$

These constraints are exactly those in (8). Moreover, the feasible set of (8) is a subset of the feasible set of (7), since every feasible solution of (8) also satisfies the constraints of (7) (the converse need not hold). Since $\{\boldsymbol{w}_\oplus^\star, \boldsymbol{w}_\ominus^\star\}$ is feasible for both problems and is optimal for the larger feasible set (7), it must also be optimal for the restricted problem (8). ■

E.2. Proof of Theorem 7 (High-dimensional Implicit Bias)

In this section, we present the proof of Theorem 7. For the 2-ReLU model ($m = 2$), the primal–dual gradient update in (4) simplifies to

$$\text{(Primal)} \qquad \boldsymbol{\beta}_\oplus^{(t+1)} = \boldsymbol{\beta}_\oplus^{(t)} - \eta \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\beta}_\oplus^{(t)})(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}), \tag{25a}$$

$$\text{(Dual)} \qquad \boldsymbol{\alpha}_\oplus^{(t+1)} = \boldsymbol{\alpha}_\oplus^{(t)} - \eta \boldsymbol{D}(\boldsymbol{\beta}_\oplus^{(t)})(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}), \tag{25b}$$

and

$$\text{(Primal)} \qquad \boldsymbol{\beta}_\ominus^{(t+1)} = \boldsymbol{\beta}_\ominus^{(t)} + \eta \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\beta}_\ominus^{(t)})(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}), \tag{26a}$$

$$\text{(Dual)} \qquad \boldsymbol{\alpha}_\ominus^{(t+1)} = \boldsymbol{\alpha}_\ominus^{(t)} + \eta \boldsymbol{D}(\boldsymbol{\beta}_\ominus^{(t)})(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}). \tag{26b}$$

Before proceeding to the proof, we again introduce a set of sufficient conditions under which the signs of the primal variables agree with the signs of the labels times the sign of the ReLU neuron at iteration t, and moreover, these conditions are preserved at iteration $t + 1$. We use the results of Lemma 9 and Lemma 10 again to prove Lemma 15.

Lemma 15 *Under Assumptions 1 and 2, suppose the gradient descent step size satisfies* $\eta \le \frac{1}{C_g \|\boldsymbol{\lambda}\|_1}$. *For a 2-ReLU model, if the following eight conditions hold at some iteration* $t \ge 0$, *then they also hold at iteration* $t + 1$.

a. $\beta_{\oplus,i}^{(t)} > 0$ *for all* $i \in [n]$ *with* $y_i > 0$.

b. $\beta_{\ominus,j}^{(t)} > 0$ *for all* $j \in [n]$ *with* $y_j < 0$.

c. $-\frac{3y_{\max}}{C_g\|\boldsymbol{\lambda}\|_1} \le \alpha_{\oplus,j}^{(t)} \le -\frac{y_{\min}}{C_\alpha\|\boldsymbol{\lambda}\|_1}$ *for all* $j \in [n]$ *with* $y_j < 0$.

d. $-\frac{3y_{\max}}{C_g\|\boldsymbol{\lambda}\|_1} \le \alpha_{\ominus,i}^{(t)} \le -\frac{y_{\min}}{C_\alpha\|\boldsymbol{\lambda}\|_1}$ *for all* $i \in [n]$ *with* $y_i > 0$.

e. $\left\|\boldsymbol{\beta}_{\oplus,+}^{(t)} - \boldsymbol{y}_+\right\|_2 \le C_y \left\|\boldsymbol{y}_+\right\|_2$, *and* $\left\|\boldsymbol{\beta}_{\ominus,-}^{(t)} + \boldsymbol{y}_-\right\|_2 \le C_y \left\|\boldsymbol{y}_-\right\|_2$.

f. $\left\|\boldsymbol{\alpha}_\oplus^{(t)}\right\|_2 \le \frac{C_\alpha\sqrt{n}y_{\max}}{\|\boldsymbol{\lambda}\|_1}$ *and* $\left\|\boldsymbol{\alpha}_\ominus^{(t)}\right\|_2 \le \frac{C_\alpha\sqrt{n}y_{\max}}{\|\boldsymbol{\lambda}\|_1}$.

g. $\beta_{\oplus,j}^{(t)} \le 0$ *for all* $j \in [n]$ *with* $y_j < 0$.

h. $\beta_{\ominus,i}^{(t)} \le 0$ *for all* $i \in [n]$ *with* $y_i > 0$.

Consequently, the set of active examples consists exactly of the positively labeled examples for the positive neuron, and the activation pattern remains unchanged, i.e., $\boldsymbol{D}(\boldsymbol{\beta}_\oplus^{(t)}) = \boldsymbol{D}(\boldsymbol{\beta}_\oplus^{(t+1)})$. *The set of active examples consists exactly of the negatively labeled examples for the negative neuron, and the activation pattern remains unchanged, i.e.,* $\boldsymbol{D}(\boldsymbol{\beta}_\ominus^{(t)}) = \boldsymbol{D}(\boldsymbol{\beta}_\ominus^{(t+1)})$.

Proof (Lemma 15) We now verify that these conditions are preserved from iteration t to $t+1$.

Part (a): By conditions (a), (b), (e), (g) and (h) at iteration t, we have

$$\begin{aligned}\left\|h_{\mathbf{\Theta}^{(t)}}(\boldsymbol{X})-\boldsymbol{y}\right\|_2^2 &= \left\|\sigma(\boldsymbol{\beta}_{\oplus}^{(t)})-\sigma(\boldsymbol{\beta}_{\ominus}^{(t)})-\boldsymbol{y}\right\|_2^2 \\ &= \left\|\sigma(\boldsymbol{\beta}_{\oplus}^{(t)})-\begin{bmatrix}\boldsymbol{y}_+\\ \mathbf{0}\end{bmatrix}-(\sigma(\boldsymbol{\beta}_{\ominus}^{(t)})+\begin{bmatrix}\mathbf{0}\\ \boldsymbol{y}_-\end{bmatrix})\right\|_2^2 \\ &= \left\|\boldsymbol{\beta}_{\oplus,+}^{(t)}-\boldsymbol{y}_+\right\|_2^2+\left\|\boldsymbol{\beta}_{\ominus,-}^{(t)}+\boldsymbol{y}_-\right\|_2^2 \le C_y^2\left\|\boldsymbol{y}\right\|_2^2,\end{aligned}$$

and therefore, $\left\|h_{\mathbf{\Theta}^{(t)}}(\boldsymbol{X})-\boldsymbol{y}\right\|_2 \le C_y\left\|\boldsymbol{y}\right\|_2$. Together with $h_{\mathbf{\Theta}^{(t)}}(\boldsymbol{x}_i)=\beta_{\oplus,i}^{(t)}$ and condition (a), the assumptions of Lemma 9 are satisfied for all i with $y_i>0$. Consequently, $\beta_{\oplus,i}^{(t+1)}>0$ for all $i\in[n]$ with $y_i>0$.

Part (b): According to the proof of part (a), we have $\left\|h_{\mathbf{\Theta}^{(t)}}(\boldsymbol{X})-\boldsymbol{y}\right\|_2 \le C_y\left\|\boldsymbol{y}\right\|_2$ and $-h_{\mathbf{\Theta}^{(t)}}(\boldsymbol{x}_j)=\beta_{\ominus,j}^{(t)}$. Together with condition (b), the assumptions of Lemma 9 are satisfied for all j with $y_j<0$. Consequently, we have $\beta_{\ominus,j}^{(t+1)}>0$ for all $j\in[n]$ with $y_j<0$.

Part (c): According to the dual gradient update in Equation (25b), and using condition (g) at iteration t, we have:

$$\alpha_{\oplus,j}^{(t+1)}=\alpha_{\oplus,j}^{(t)} \quad \text{for all } j\in[n] \text{ with } y_j<0.$$

Therefore, condition (c) continues to hold at iteration $t+1$.

Part (d): According to the dual gradient update in Equation (26b), and using conditions (h) at iteration t, we have:

$$\alpha_{\ominus,i}^{(t+1)}=\alpha_{\ominus,i}^{(t)} \quad \text{for all } i\in[n] \text{ with } y_i>0.$$

Therefore, condition (d) continues to hold at iteration $t+1$.

Part (e): By conditions (a), (b), (g), and (h), the gradient update at iteration t for $\boldsymbol{\beta}_{\oplus}^{(t)}$ depends only on the positively labeled examples, and the update for $\boldsymbol{\beta}_{\ominus}^{(t)}$ depends only on the negatively labeled examples. Hence, the gradient update for an individual neuron is equivalent to gradient descent on a certain linear regression problem. Since the step size satisfies $\eta\le\frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$, the linear regression squared loss is monotonically nonincreasing (as in the proof of Lemma 2), and by condition (e) at iteration t, we obtain

$$\begin{aligned}\left\|\boldsymbol{\beta}_{\oplus,+}^{(t+1)}-\boldsymbol{y}_+\right\|_2 &\le \left\|\boldsymbol{\beta}_{\oplus,+}^{(t)}-\boldsymbol{y}_+\right\|_2 \le C_y\left\|\boldsymbol{y}_+\right\|_2, \\ \left\|-\boldsymbol{\beta}_{\ominus,-}^{(t+1)}-\boldsymbol{y}_-\right\|_2 &\le \left\|-\boldsymbol{\beta}_{\ominus,-}^{(t)}-\boldsymbol{y}_-\right\|_2 \le C_y\left\|\boldsymbol{y}_-\right\|_2,\end{aligned}$$

where we use $\boldsymbol{y}_+$ ($\boldsymbol{y}_-$) to denote the vector of positively labeled (negatively labeled) examples. Therefore, condition (e) holds at iteration $t+1$.

Part (f): Following the same argument as in Part (d) of Lemma 14, using conditions (c), (d), and (e) at iteration $t+1$, together with the eigenvalue bounds from Lemma 11, we have

$$\left\|\boldsymbol{\alpha}_{\oplus}^{(t+1)}\right\|_2 \le \frac{C_\alpha \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1}, \quad \left\|\boldsymbol{\alpha}_{\ominus}^{(t+1)}\right\|_2 \le \frac{C_\alpha \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1}$$

with probability at least $1-2e^{-n/C_g}$. Thus, condition (f) holds at iteration $t+1$.

Part (g): By Lemma 10, since conditions (c) and (f) hold at iteration $t+1$, we conclude that $\beta_{\oplus,j}^{(t+1)} \le 0$ for all $j \in [n]$ with $y_j < 0$. Thus, condition (g) holds at iteration $t+1$.

Part (h): Similarly, since conditions (d) and (f) hold at iteration $t+1$, we have $\beta_{\ominus,i}^{(t+1)} \le 0$ for all $i \in [n]$ with $y_i > 0$. Thus, condition (h) holds at iteration $t+1$.

■

Equipped with Lemma 15, we are ready to prove Theorem 7.

Proof (Theorem 7) The proof follows a similar structure to that of Theorem 4 for single ReLU models, but now we must track the dynamics for both $\boldsymbol{w}_{\oplus}$ and $\boldsymbol{w}_{\ominus}$ simultaneously. Equipped with sufficient conditions under which the activation patterns are preserved in Lemma 15, we verify these conditions hold after the first gradient step, and use induction to characterize the full gradient descent dynamics.

We first verify that the iterate at $t=1$ satisfies all the sufficient conditions. With the initialization

$$\boldsymbol{w}_{\oplus}^{(0)} = \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} \boldsymbol{\epsilon}_{\oplus}, \quad \boldsymbol{w}_{\ominus}^{(0)} = \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} \boldsymbol{\epsilon}_{\ominus},$$

we have $\boldsymbol{\beta}_{\oplus}^{(0)} = \boldsymbol{\epsilon}_{\oplus}$ and $\boldsymbol{\beta}_{\ominus}^{(0)} = \boldsymbol{\epsilon}_{\ominus}$. By the theorem assumptions, $0 < \epsilon_{\oplus,i} \le \frac{1}{2C_\alpha} y_{\min}$ and $0 < \epsilon_{\ominus,i} \le \frac{1}{2C_\alpha} y_{\min}$ for all $i \in [n]$. Using the primal gradient updates in Equations (25a) and (26a), we have

$$\begin{aligned}
\boldsymbol{\beta}_{\oplus}^{(1)} &= \boldsymbol{\epsilon}_{\oplus} - \eta \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\epsilon}_{\oplus})(h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) - \boldsymbol{y}) \\
&= \boldsymbol{\epsilon}_{\oplus} - \eta \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\epsilon}_{\oplus})(\sigma(\boldsymbol{\epsilon}_{\oplus}) - \sigma(\boldsymbol{\epsilon}_{\ominus}) - \boldsymbol{y}) \\
&= \boldsymbol{\epsilon}_{\oplus} - \eta \boldsymbol{X}\boldsymbol{X}^\top (\boldsymbol{\epsilon}_{\oplus} - \boldsymbol{\epsilon}_{\ominus} - \boldsymbol{y}), \qquad (27)
\end{aligned}$$

where the last equality uses the fact that $\boldsymbol{\epsilon}_{\oplus} > \boldsymbol{0}$ and $\boldsymbol{\epsilon}_{\ominus} > \boldsymbol{0}$ componentwise, so $\boldsymbol{D}(\boldsymbol{\epsilon}_{\oplus}) = \boldsymbol{I}$, $\sigma(\boldsymbol{\epsilon}_{\oplus}) = \boldsymbol{\epsilon}_{\oplus}$, and $\sigma(\boldsymbol{\epsilon}_{\ominus}) = \boldsymbol{\epsilon}_{\ominus}$. Similarly, we have

$$\begin{aligned}
\boldsymbol{\beta}_{\ominus}^{(1)} &= \boldsymbol{\epsilon}_{\ominus} + \eta \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\epsilon}_{\ominus})(\sigma(\boldsymbol{\epsilon}_{\oplus}) - \sigma(\boldsymbol{\epsilon}_{\ominus}) - \boldsymbol{y}) \\
&= \boldsymbol{\epsilon}_{\ominus} + \eta \boldsymbol{X}\boldsymbol{X}^\top (\boldsymbol{\epsilon}_{\oplus} - \boldsymbol{\epsilon}_{\ominus} - \boldsymbol{y}). \qquad (28)
\end{aligned}$$

For the dual variables, we have $\boldsymbol{\alpha}_{\oplus}^{(0)} = (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{\epsilon}_{\oplus}$ and $\boldsymbol{\alpha}_{\ominus}^{(0)} = (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{\epsilon}_{\ominus}$. The dual updates give:

$$\begin{aligned}
\boldsymbol{\alpha}_{\oplus}^{(1)} &= \boldsymbol{\alpha}_{\oplus}^{(0)} - \eta \boldsymbol{D}(\boldsymbol{\epsilon}_{\oplus})(\boldsymbol{\epsilon}_{\oplus} - \boldsymbol{\epsilon}_{\ominus} - \boldsymbol{y}) \\
&= \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} \boldsymbol{\epsilon}_{\oplus} - \eta(\boldsymbol{\epsilon}_{\oplus} - \boldsymbol{\epsilon}_{\ominus} - \boldsymbol{y}) \\
&= \eta \left(\boldsymbol{y} - \boldsymbol{\epsilon}_{\oplus} + \boldsymbol{\epsilon}_{\ominus} + \frac{1}{\eta}\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} \boldsymbol{\epsilon}_{\oplus}\right),
\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\alpha}_{\ominus}^{(1)} &= \boldsymbol{\alpha}_{\ominus}^{(0)} + \eta \boldsymbol{D}(\boldsymbol{\epsilon}_{\ominus})(\boldsymbol{\epsilon}_{\oplus} - \boldsymbol{\epsilon}_{\ominus} - \boldsymbol{y}) \\ &= \left(\boldsymbol{X}\boldsymbol{X}^{\top}\right)^{-1} \boldsymbol{\epsilon}_{\ominus} + \eta(\boldsymbol{\epsilon}_{\oplus} - \boldsymbol{\epsilon}_{\ominus} - \boldsymbol{y}) \\ &= \eta \left(-\boldsymbol{y} + \boldsymbol{\epsilon}_{\oplus} - \boldsymbol{\epsilon}_{\ominus} + \frac{1}{\eta}\left(\boldsymbol{X}\boldsymbol{X}^{\top}\right)^{-1} \boldsymbol{\epsilon}_{\ominus}\right).\end{aligned}$$

We now verify each condition at $t = 1$.

Part (a): For all $i \in [n]$ with $y_i > 0$, we apply Lemma 9. Since $\beta_{\oplus,i}^{(0)} = \epsilon_{\oplus,i} > 0$, $h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{x}_i) = \beta_{\oplus,i}^{(0)} - \beta_{\ominus,i}^{(0)} \le \beta_{\oplus,i}^{(0)}$, and

$$\begin{aligned}\left\|h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) - \boldsymbol{y}\right\|_2 &= \|\boldsymbol{\epsilon}_{\oplus} - \boldsymbol{\epsilon}_{\ominus} - \boldsymbol{y}\|_2 \\ &\le \|\boldsymbol{\epsilon}_{\oplus}\|_2 + \|\boldsymbol{\epsilon}_{\ominus}\|_2 + \|\boldsymbol{y}\|_2 \le \frac{\sqrt{n}}{C_\alpha} y_{\min} + \|\boldsymbol{y}\|_2 \le C_y \|\boldsymbol{y}\|_2 \,, \end{aligned} \tag{29}$$

where we have used $C_y \ge \frac{1}{C_\alpha} + 1$. We conclude that $\beta_{\oplus,i}^{(1)} > 0$ for all i with $y_i > 0$.

Part (b): For all $j \in [n]$ with $y_j < 0$, we apply Lemma 9. Since $\beta_{\ominus,j}^{(0)} = \epsilon_{\ominus,j} > 0$, $-h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{x}_j) = -\beta_{\oplus,j}^{(0)} + \beta_{\ominus,j}^{(0)} \le \beta_{\ominus,j}^{(0)}$, and $\left\|h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) - \boldsymbol{y}\right\|_2 \le C_y \|\boldsymbol{y}\|_2$ by Equation (29), we conclude that $\beta_{\ominus,j}^{(1)} > 0$ for all $j \in [n]$ with $y_j < 0$.

Part (c): For all $j \in [n]$ with $y_j < 0$, we verify that $\alpha_{\oplus,j}^{(1)}$ satisfies the required upper and lower bounds. For the upper bound, we have

$$\begin{aligned}\alpha_{\oplus,j}^{(1)} &= \eta \left(y_j - \epsilon_{\oplus,j} + \epsilon_{\ominus,j} + \frac{1}{\eta}\boldsymbol{e}_j^{\top}\left(\boldsymbol{X}\boldsymbol{X}^{\top}\right)^{-1}\boldsymbol{\epsilon}_{\oplus}\right) \\ &= \eta \left(y_j - \epsilon_{\oplus,j} + \epsilon_{\ominus,j} + \frac{1}{\eta}\boldsymbol{e}_j^{\top}\left[\frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I} + \left(\left(\boldsymbol{X}\boldsymbol{X}^{\top}\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I}\right)\right]\boldsymbol{\epsilon}_{\oplus}\right) \\ &= \eta \left(y_j - \epsilon_{\oplus,j} + \epsilon_{\ominus,j} + \frac{\epsilon_{\oplus,j}}{\eta\|\boldsymbol{\lambda}\|_1} + \frac{1}{\eta}\boldsymbol{e}_j^{\top}\left(\left(\boldsymbol{X}\boldsymbol{X}^{\top}\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I}\right)\boldsymbol{\epsilon}_{\oplus}\right) \\ &\le \eta \left(y_j + \epsilon_{\ominus,j} + \frac{\epsilon_{\oplus,j}}{\eta\|\boldsymbol{\lambda}\|_1} + \frac{1}{\eta}\left\|\left(\boldsymbol{X}\boldsymbol{X}^{\top}\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I}\right\|_2 \|\boldsymbol{\epsilon}_{\oplus}\|_2\right),\end{aligned}$$

where the inequality drops the negative term $-\epsilon_{\oplus,j}$. By Corollary 13, we have

$$\left\|\left(\boldsymbol{X}\boldsymbol{X}^{\top}\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I}\right\|_2 \le \frac{C_g C}{\|\boldsymbol{\lambda}\|_1} \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right),$$

with probability at least $1 - 2\exp(-n(Cc - \ln 9))$. Moreover, by the theorem assumptions, $\epsilon_{\oplus,j} \le \frac{1}{2C_\alpha} y_{\min}$, $\epsilon_{\ominus,j} \le \frac{1}{2C_\alpha} y_{\min}$, and $\frac{1}{\eta} \le C C_g \|\boldsymbol{\lambda}\|_1$. Combining these bounds yields

$$
\begin{aligned}
\alpha_{\oplus,j}^{(1)} &\le \frac{1}{CC_g \left\|\boldsymbol{\lambda}\right\|_1}\left(-y_{\min} + \frac{1}{2C_\alpha}y_{\min} + \frac{CC_g}{2C_\alpha}y_{\min} + C^2C_g^2\cdot\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\cdot\frac{\sqrt{n}}{2C_\alpha}y_{\min}\right)\\
&\le \frac{1}{CC_g \left\|\boldsymbol{\lambda}\right\|_1}\left(-y_{\min} + \frac{1}{2C_\alpha}y_{\min} + \frac{CC_g}{2C_\alpha}y_{\min} + C^2C_g^2\cdot\frac{y_{\min}}{C_0 y_{\max}}\cdot\frac{1}{2C_\alpha}y_{\min}\right)\\
&= -\frac{y_{\min}}{C_\alpha \left\|\boldsymbol{\lambda}\right\|_1}\left(\frac{C_\alpha}{CC_g} - \frac{1}{2CC_g} - \frac{1}{2} - \frac{CC_g y_{\min}}{2C_0 y_{\max}}\right)\\
&\le -\frac{y_{\min}}{C_\alpha \left\|\boldsymbol{\lambda}\right\|_1}. \qquad (30)
\end{aligned}
$$

The second inequality follows from $d_2 \ge C_0^2\frac{n^2 y_{\max}^2}{y_{\min}^2}$ and $d_\infty \ge C_0\frac{n^{1.5}y_{\max}}{y_{\min}}$ in Assumption 2, and the last inequality follows the relationship between constants $C_0 \gtrsim C_\alpha^2$ and $C_\alpha \gtrsim \max\{C_g^2, C_y C_g\}$. For the lower bound, we have

$$
\begin{aligned}
\alpha_{\oplus,j}^{(1)} &= \eta\left(y_j - \epsilon_{\oplus,j} + \epsilon_{\ominus,j} + \frac{\epsilon_{\oplus,j}}{\eta\left\|\boldsymbol{\lambda}\right\|_1} + \frac{1}{\eta}\boldsymbol{e}_j^\top\left(\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\left\|\boldsymbol{\lambda}\right\|_1}\boldsymbol{I}\right)\boldsymbol{\epsilon}_\oplus\right)\\
&\ge \eta\left(-y_{\max} - \epsilon_{\oplus,j} - \frac{1}{\eta}\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\left\|\boldsymbol{\lambda}\right\|_1}\boldsymbol{I}\right\|_2\left\|\boldsymbol{\epsilon}_\oplus\right\|_2\right)\\
&\ge \frac{1}{C_g\left\|\boldsymbol{\lambda}\right\|_1}\left(-y_{\max} - \frac{1}{2C_\alpha}y_{\min} - C^2C_g^2\cdot\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\cdot\frac{\sqrt{n}}{2C_\alpha}y_{\min}\right)\\
&\ge \frac{1}{C_g\left\|\boldsymbol{\lambda}\right\|_1}\left(-y_{\max} - \frac{1}{2C_\alpha}y_{\min} - C^2C_g^2\cdot\frac{y_{\min}}{C_0 y_{\max}}\cdot\frac{1}{2C_\alpha}y_{\min}\right)\\
&\ge -\frac{3y_{\max}}{C_g\left\|\boldsymbol{\lambda}\right\|_1}, \qquad (31)
\end{aligned}
$$

by the same arguments. Thus, $\alpha_{\oplus,j}^{(1)}$ satisfies both the required upper and lower bounds for all j with $y_j < 0$.

Part (d): For all $i \in [n]$ with $y_i > 0$, we verify that $\alpha_{\ominus,i}^{(1)}$ satisfies the required bounds in the approach analogous to Part (c). For the upper bound, we have

$$
\begin{aligned}
\alpha_{\ominus,i}^{(1)} &= \eta\left(-y_i + \epsilon_{\oplus,i} - \epsilon_{\ominus,i} + \frac{1}{\eta}\boldsymbol{e}_i^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{\epsilon}_\ominus\right)\\
&= \eta\left(-y_i + \epsilon_{\oplus,i} - \epsilon_{\ominus,i} + \frac{1}{\eta}\boldsymbol{e}_i^\top\left[\frac{1}{\left\|\boldsymbol{\lambda}\right\|_1}\boldsymbol{I} + \left(\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\left\|\boldsymbol{\lambda}\right\|_1}\boldsymbol{I}\right)\right]\boldsymbol{\epsilon}_\ominus\right)\\
&= \eta\left(-y_i + \epsilon_{\oplus,i} - \epsilon_{\ominus,i} + \frac{\epsilon_{\ominus,i}}{\eta\left\|\boldsymbol{\lambda}\right\|_1} + \frac{1}{\eta}\boldsymbol{e}_i^\top\left(\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\left\|\boldsymbol{\lambda}\right\|_1}\boldsymbol{I}\right)\boldsymbol{\epsilon}_\ominus\right)\\
&\le \eta\left(-y_i + \epsilon_{\oplus,i} + \frac{\epsilon_{\ominus,i}}{\eta\left\|\boldsymbol{\lambda}\right\|_1} + \frac{1}{\eta}\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\left\|\boldsymbol{\lambda}\right\|_1}\boldsymbol{I}\right\|_2\left\|\boldsymbol{\epsilon}_\ominus\right\|_2\right),
\end{aligned}
$$

where the inequality drops the negative term $-\epsilon_{\ominus,i}$. Applying the upper bound in Corollary 13 and the theorem assumptions $\epsilon_{\oplus,i} \le \frac{1}{2C_\alpha}y_{\min}$, $\epsilon_{\ominus,i} \le \frac{1}{2C_\alpha}y_{\min}$, and

$\frac{1}{\eta} \leq CC_g \|\boldsymbol{\lambda}\|_1$, we have

$$\alpha_{\ominus,i}^{(1)} \leq \frac{1}{CC_g \|\boldsymbol{\lambda}\|_1} \left(-y_{\min} + \frac{1}{2C_\alpha} y_{\min} + \frac{CC_g}{2C_\alpha} y_{\min} + C^2 C_g^2 \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty} \right) \cdot \frac{\sqrt{n}}{2C_\alpha} y_{\min} \right)$$
$$\leq -\frac{y_{\min}}{C_\alpha \|\boldsymbol{\lambda}\|_1},$$

where the second inequality follows the argument used in Equation (30). For the lower bound, following the same argument as in Part (c), we have

$$\begin{aligned} \alpha_{\ominus,i}^{(1)} &= \eta \left(-y_i + \epsilon_{\oplus,i} - \epsilon_{\ominus,i} + \frac{\epsilon_{\ominus,i}}{\eta \|\boldsymbol{\lambda}\|_1} + \frac{1}{\eta} \boldsymbol{e}_i^\top \left(\left(\boldsymbol{X}\boldsymbol{X}^\top \right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1} \boldsymbol{I} \right) \boldsymbol{\epsilon}_\ominus \right) \\ &\geq \eta \left(-y_{\max} - \epsilon_{\ominus,i} - \frac{1}{\eta} \left\| \left(\boldsymbol{X}\boldsymbol{X}^\top \right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1} \boldsymbol{I} \right\|_2 \|\boldsymbol{\epsilon}_\ominus\|_2 \right) \\ &\geq \frac{1}{C_g \|\boldsymbol{\lambda}\|_1} \left(-y_{\max} - \frac{1}{2C_\alpha} y_{\min} - C^2 C_g^2 \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty} \right) \cdot \frac{\sqrt{n}}{2C_\alpha} y_{\min} \right) \\ &\geq -\frac{3y_{\max}}{C_g \|\boldsymbol{\lambda}\|_1}, \end{aligned}$$

where the last inequality follows follows the argument used in Equation (31). Thus, $\alpha_{\ominus,i}^{(1)}$ satisfies both bounds for all i with $y_i > 0$.

Part (e): We verify that the primal variables $\boldsymbol{\beta}_\oplus^{(1)}$ corresponding to positively labeled examples minus $\boldsymbol{y}_+$ satisfy the norm bound. Specifically, we show that $\left\| \boldsymbol{\beta}_{\oplus,+}^{(1)} - \boldsymbol{y}_+ \right\|_2^2 \leq C_y^2 \|\boldsymbol{y}_+\|_2^2$. According to Equation (27), we have

$$\begin{aligned} \left\| \boldsymbol{\beta}_{\oplus,+}^{(1)} - \boldsymbol{y}_+ \right\|_2^2 &= \sum_{i:y_i>0} \left(\beta_{\oplus,i}^{(1)} - y_i \right)^2 \\ &= \sum_{i:y_i>0} \Bigg(\underbrace{\epsilon_{\oplus,i} - \eta \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \left(\boldsymbol{\epsilon}_\oplus - \boldsymbol{\epsilon}_\ominus - \boldsymbol{y} \right) - y_i}_{=:T_i} \Bigg)^2 . \end{aligned} \tag{32}$$

Next, we bound the term $T_i := \epsilon_{\oplus,i} - \eta \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \left(\boldsymbol{\epsilon}_\oplus - \boldsymbol{\epsilon}_\ominus - \boldsymbol{y} \right) - y_i$ for all $i \in [n]$ with $y_i > 0$. We have

$$\begin{aligned} T_i &= \epsilon_{\oplus,i} - \eta \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \left(\boldsymbol{\epsilon}_\oplus - \boldsymbol{\epsilon}_\ominus - \boldsymbol{y} \right) - y_i \\ &= (\epsilon_{\oplus,i} - y_i) - \eta \boldsymbol{e}_i^\top \left[\|\boldsymbol{\lambda}\|_1 \boldsymbol{I} + \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I} \right) \right] \left(\boldsymbol{\epsilon}_\oplus - \boldsymbol{\epsilon}_\ominus - \boldsymbol{y} \right) \\ &= (1 - \eta \|\boldsymbol{\lambda}\|_1) \epsilon_{\oplus,i} + \eta \|\boldsymbol{\lambda}\|_1 \epsilon_{\ominus,i} - (1 - \eta \|\boldsymbol{\lambda}\|_1) y_i - \eta \boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I} \right) \left(\boldsymbol{\epsilon}_\oplus - \boldsymbol{\epsilon}_\ominus - \boldsymbol{y} \right). \end{aligned}$$

Since the step size assumption guarantees that $\frac{1}{CC_g\|\boldsymbol{\lambda}\|_1} \leq \eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$, and $\epsilon_{\oplus,i}, \epsilon_{\ominus,i} \leq \frac{1}{2C_\alpha} y_{\min}$, we have

$$\begin{aligned} (1 - \eta \|\boldsymbol{\lambda}\|_1) \epsilon_{\oplus,i} + \eta \|\boldsymbol{\lambda}\|_1 \epsilon_{\ominus,i} - (1 - \eta \|\boldsymbol{\lambda}\|_1) y_i &\leq \epsilon_{\oplus,i} + \eta \|\boldsymbol{\lambda}\|_1 \epsilon_{\ominus,i} - (1 - \eta \|\boldsymbol{\lambda}\|_1) y_i \\ &\leq \left(1 + \frac{1}{C_g} \right) \frac{1}{2C_\alpha} y_{\min} - \left(1 - \frac{1}{C_g} \right) y_{\min} \\ &< 0, \end{aligned}$$

with $C_\alpha \gtrsim C_g^2$. Hence, in order to upper bound T_i^2, it suffices to find the lower bound for T_i. We have

$$\begin{aligned} T_i &= (1-\eta\|\boldsymbol{\lambda}\|_1)\epsilon_{\oplus,i} + \eta\|\boldsymbol{\lambda}\|_1\epsilon_{\ominus,i} - (1-\eta\|\boldsymbol{\lambda}\|_1)y_i - \eta\boldsymbol{e}_i^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1\boldsymbol{I}\right)(\boldsymbol{\epsilon}_\oplus - \boldsymbol{\epsilon}_\ominus - \boldsymbol{y}) \\ &\geq -y_i - \eta\left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1\boldsymbol{I}\right\|_2\|\boldsymbol{\epsilon}_\oplus - \boldsymbol{\epsilon}_\ominus - \boldsymbol{y}\|_2, \end{aligned}$$

where the inequality drops the positive terms $(1-\eta\|\boldsymbol{\lambda}\|_1)\epsilon_{\oplus,i}$, $\eta\|\boldsymbol{\lambda}\|_1\epsilon_{\ominus,i}$, and $\eta\|\boldsymbol{\lambda}\|_1 y_i$. We again upper bound $\left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1\boldsymbol{I}\right\|_2$ by Corollary 13. With probability at least $1-2\exp(-n(Cc-\ln 9))$, we have

$$\begin{aligned} T_i &\geq -y_i - \eta\cdot C\|\boldsymbol{\lambda}\|_1\cdot\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\|\boldsymbol{\epsilon}_\oplus - \boldsymbol{\epsilon}_\ominus - \boldsymbol{y}\|_2 \\ &\geq -y_i - \frac{C}{C_g}\cdot\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\|\boldsymbol{\epsilon}_\oplus - \boldsymbol{\epsilon}_\ominus - \boldsymbol{y}\|_2, \end{aligned}$$

by applying $\eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$. Finally, we apply the upper bounds for $\|\boldsymbol{\epsilon}_\oplus\|_2$, $\|\boldsymbol{\epsilon}_\ominus\|_2$ and $\|\boldsymbol{y}\|_2$, and Assumption 2 ensures that $d_2 \geq C_0^2\frac{n^2 y_{\max}^2}{y_{\min}^2}$ and $d_\infty \geq C_0\frac{n^{1.5}y_{\max}}{y_{\min}}$. We have

$$\begin{aligned} T_i &\geq -y_i - \frac{C}{C_g}\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\left(\frac{\sqrt{n}}{C_\alpha}y_{\min} + \sqrt{n}y_{\max}\right) \\ &\geq -y_i - \frac{Cy_{\min}}{C_gC_0y_{\max}}\left(\frac{1}{C_\alpha}y_{\min} + y_{\max}\right) \\ &\geq -y_i\left(1 + \frac{2C}{C_gC_0}\right) \\ &\geq -C_y y_i, \end{aligned}$$

with the choice of $C_y \geq 2$. Substituting $T_i^2 \leq C_y^2 y_i^2$ into Equation (32), we have

$$\left\|\boldsymbol{\beta}_{\oplus,+}^{(1)} - \boldsymbol{y}_+\right\|_2^2 \leq \sum_{i:y_i>0} C_y^2 y_i^2 = C_y^2\|\boldsymbol{y}_+\|_2^2.$$

As a result, we conclude that $\left\|\boldsymbol{\beta}_{\oplus,+}^{(1)} - \boldsymbol{y}_+\right\|_2 \leq C_y\|\boldsymbol{y}_+\|_2$ as required. The same derivation holds for $\left\|\boldsymbol{\beta}_{\ominus,-}^{(1)} + \boldsymbol{y}_-\right\|_2 \leq C_y\|\boldsymbol{y}_-\|_2$ by an analogous argument. Therefore, condition (e) holds at $t=1$.

Part (f): We verify the norm bounds on the dual variables. By the triangle inequality, we have

$$\begin{aligned} \left\|\boldsymbol{\alpha}_\oplus^{(1)}\right\|_2 &= \left\|\eta\left(\boldsymbol{y} - \boldsymbol{\epsilon}_\oplus + \boldsymbol{\epsilon}_\ominus + \frac{1}{\eta}\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{\epsilon}_\oplus\right)\right\|_2 \\ &\leq \eta\left(\|\boldsymbol{y}\|_2 + \|\boldsymbol{\epsilon}_\oplus\|_2 + \|\boldsymbol{\epsilon}_\ominus\|_2 + \frac{1}{\eta}\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\right\|_2\|\boldsymbol{\epsilon}_\oplus\|_2\right). \end{aligned}$$

Using $\|\boldsymbol{y}\|_2 \leq \sqrt{n}y_{\max}$, $\|\boldsymbol{\epsilon}_\oplus\|_2, \|\boldsymbol{\epsilon}_\ominus\|_2 \leq \frac{\sqrt{n}}{2C_\alpha}y_{\min}$, $\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\right\|_2 \leq \frac{C_g}{\|\boldsymbol{\lambda}\|_1}$, and $\frac{1}{CC_g\|\boldsymbol{\lambda}\|_1} \leq \eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$, we have

$$\begin{aligned}\left\|\boldsymbol{\alpha}_\oplus^{(1)}\right\|_2 &\leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}\left(\sqrt{n}y_{\max} + \frac{\sqrt{n}}{C_\alpha}y_{\min} + CC_g\|\boldsymbol{\lambda}\|_1 \cdot \frac{C_g}{\|\boldsymbol{\lambda}\|_1} \cdot \frac{\sqrt{n}}{C_\alpha}y_{\min}\right) \\ &\leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}\left(3\sqrt{n}y_{\max}\right) \\ &\leq \frac{C_\alpha\sqrt{n}y_{\max}}{\|\boldsymbol{\lambda}\|_1},\end{aligned}$$

with $C_\alpha \gtrsim \max\{C_g^2, C_yC_g\}$. The same bound holds for $\left\|\boldsymbol{\alpha}_\ominus^{(1)}\right\|_2$. Thus, condition (f) holds at $t = 1$.

Part (g): Since we have shown that $\alpha_{\oplus,j}^{(1)} \leq -\frac{y_{\min}}{C_\alpha\|\boldsymbol{\lambda}\|_1}$ and $\left\|\boldsymbol{\alpha}_\oplus^{(1)}\right\|_2 \leq \frac{C_\alpha\sqrt{n}y_{\max}}{\|\boldsymbol{\lambda}\|_1}$ for all $j \in [n]$ with $y_j < 0$, it follows from Lemma 10 that $\beta_{\oplus,j}^{(1)} \leq 0$ for all $j \in [n]$ with $y_j < 0$.

Part (h): Similarly, since we have shown that $\alpha_{\ominus,i}^{(1)} \leq -\frac{y_{\min}}{C_\alpha\|\boldsymbol{\lambda}\|_1}$ and $\left\|\boldsymbol{\alpha}_\ominus^{(1)}\right\|_2 \leq \frac{C_\alpha\sqrt{n}y_{\max}}{\|\boldsymbol{\lambda}\|_1}$ for all $i \in [n]$ with $y_i > 0$, it follows from Lemma 10 that $\beta_{\ominus,i}^{(1)} \leq 0$ for all $i \in [n]$ with $y_i > 0$.

We have shown that at iteration $t = 1$ the conditions in Lemma 15 are satisfied, and by induction, these conditions will also hold for $t \geq 1$. As a result, the positive neuron $\boldsymbol{w}_\oplus$ is trained with only positive examples starting from the iteration $t = 1$, and it is equivalent to linear regression using only positive examples with initialization $\boldsymbol{w}_\oplus^{(1)} = \eta\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{\epsilon}_\oplus + \boldsymbol{\epsilon}_\ominus + \frac{1}{\eta}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{\epsilon}_\oplus\right)$. Finally, since $\boldsymbol{w}_\oplus$ and $\boldsymbol{w}_\ominus$ are trained on disjoint subsets of examples, $\boldsymbol{w}_\oplus^{(\infty)}$ satisfies

$$\boldsymbol{w}_\oplus^{(\infty)} = \underset{\boldsymbol{w}\in\{\boldsymbol{w}:\boldsymbol{X}_+\boldsymbol{w}=\boldsymbol{y}_+\}}{\arg\min}\left\|\boldsymbol{w} - \boldsymbol{w}_\oplus^{(1)}\right\|_2,$$

by Lemma 2. The same arguments apply to the negative neuron $\boldsymbol{w}_\ominus$ as well. This completes the proof of Theorem 7.

■

E.3. Proof of Theorem 8 (Implicit Bias Approximation to $w^\star$)

Proof (Theorem 8) We divide the proof into four steps, and formally show the result for $\boldsymbol{w}_\oplus^\star$. The result for $\boldsymbol{w}_\ominus^\star$ follows an identical series of steps. In the first step, we derive an upper bound for $\|\boldsymbol{\alpha}_\oplus^\star\|_2$ and $\|\boldsymbol{\alpha}_\ominus^\star\|_2$ where $\boldsymbol{w}_\oplus^\star := \boldsymbol{X}^\top\boldsymbol{\alpha}_\oplus^\star$ and $\boldsymbol{w}_\ominus^\star := \boldsymbol{X}^\top\boldsymbol{\alpha}_\ominus^\star$, by using the optimality of the objective function in (7). In the second step, we use the KKT conditions of (8) to find a representation of $\{\boldsymbol{w}_\oplus^\star, \boldsymbol{w}_\ominus^\star\}$. In Steps 3 and 4, we derive the corresponding upper bound and lower bound.

Step 1: Upper bounds for $\|\boldsymbol{\alpha}_\oplus^\star\|_2$ and $\|\boldsymbol{\alpha}_\ominus^\star\|_2$.
$\{\boldsymbol{w}_\oplus^\star, \boldsymbol{w}_\ominus^\star\}$ is the optimal solution to (7) and it achieves the minimum objective in (7). In the proof of Lemma 6, we introduce $\{\tilde{\boldsymbol{w}}_\oplus, \tilde{\boldsymbol{w}}_\ominus\}$ which is also a feasible solution to (7), where $\tilde{\boldsymbol{w}}_\oplus :=$

$\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}_\oplus$ and $\tilde{\boldsymbol{w}}_\ominus := \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}_\ominus$, with $y_{\oplus,i} = \max\{y_i, 0\}$ and $y_{\ominus,i} = -\min\{y_i, 0\}$. Therefore, by the optimality of $\{\boldsymbol{w}_\oplus^\star, \boldsymbol{w}_\ominus^\star\}$ in the objective, we have

$$\begin{aligned}\left\|\boldsymbol{w}_\oplus^\star\right\|_2^2 + \left\|\boldsymbol{w}_\ominus^\star\right\|_2^2 &= \boldsymbol{\alpha}_\oplus^{\star\top}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{\alpha}_\oplus^\star + \boldsymbol{\alpha}_\ominus^{\star\top}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{\alpha}_\ominus^\star \\ &\le \|\tilde{\boldsymbol{w}}_\oplus\|_2^2 + \|\tilde{\boldsymbol{w}}_\ominus\|_2^2 \\ &= \boldsymbol{y}_\oplus^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}_\oplus + \boldsymbol{y}_\ominus^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}_\ominus \\ &\le \left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\right\|_2 \|\boldsymbol{y}_\oplus\|_2^2 + \left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\right\|_2 \|\boldsymbol{y}_\ominus\|_2^2 \\ &\le \frac{2C_g n y_{\max}^2}{\|\boldsymbol{\lambda}\|_1},\end{aligned}$$

where the last inequality uses Lemma 11 with probability at least $1 - 2e^{-n/C_g}$, and we have $\max\{\left\|\boldsymbol{y}_\oplus\right\|_2^2, \left\|\boldsymbol{y}_\ominus\right\|_2^2\} \le \|\boldsymbol{y}\|_2^2 \le n y_{\max}^2$. Therefore, we have

$$\begin{aligned}\lambda_n(\boldsymbol{X}\boldsymbol{X}^\top)\left\|\boldsymbol{\alpha}_\oplus^\star\right\|_2^2 &\le \boldsymbol{\alpha}_\oplus^{\star\top}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{\alpha}_\oplus^\star \\ &\le \boldsymbol{\alpha}_\oplus^{\star\top}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{\alpha}_\oplus^\star + \boldsymbol{\alpha}_\ominus^{\star\top}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{\alpha}_\ominus^\star \\ &\le \frac{2C_g n y_{\max}^2}{\|\boldsymbol{\lambda}\|_1}.\end{aligned}$$

As a result, we have $\left\|\boldsymbol{\alpha}_\oplus^\star\right\|_2 \le \frac{\sqrt{2}C_g\sqrt{n}y_{\max}}{\|\boldsymbol{\lambda}\|_1}$. This bound applies to $\left\|\boldsymbol{\alpha}_\ominus^\star\right\|_2$ as well via an identical argument.

Step 2: KKT conditions of $\{\boldsymbol{w}_\oplus^\star, \boldsymbol{w}_\ominus^\star\}$ by Lemma 6.
Based on Lemma 6, the optimal solution $\{\boldsymbol{w}_\oplus^\star, \boldsymbol{w}_\ominus^\star\}$ of (7) is also the optimal solution of a convex program (8). Hence, we restate the convex program in (8) below

$$\begin{aligned} \boldsymbol{w}_\oplus^\star, \boldsymbol{w}_\ominus^\star \quad &= \underset{\boldsymbol{w}_\oplus, \boldsymbol{w}_\ominus}{\arg\min}\ \frac{1}{2}\|\boldsymbol{w}_\oplus\|_2^2 + \frac{1}{2}\|\boldsymbol{w}_\ominus\|_2^2 \\ \text{s.t. } \boldsymbol{w}_\oplus^\top\boldsymbol{x}_i \quad &= y_i, \qquad \boldsymbol{w}_\ominus^\top\boldsymbol{x}_i \le 0, \quad \text{for all } i \in S_1, \\ \boldsymbol{w}_\oplus^\top\boldsymbol{x}_i - \boldsymbol{w}_\ominus^\top\boldsymbol{x}_i &= y_i, \qquad -\boldsymbol{w}_\ominus^\top\boldsymbol{x}_i \le 0, \quad \text{for all } i \in S_2, \\ -\boldsymbol{w}_\ominus^\top\boldsymbol{x}_i &= y_i, \quad \boldsymbol{w}_\oplus^\top\boldsymbol{x}_i \le 0, \qquad \text{for all } i \in S_3, \\ \boldsymbol{w}_\oplus^\top\boldsymbol{x}_i - \boldsymbol{w}_\ominus^\top\boldsymbol{x}_i &= y_i, -\boldsymbol{w}_\oplus^\top\boldsymbol{x}_i \le 0, \qquad \text{for all } i \in S_4. \end{aligned}$$

The Lagrange function in terms of $\boldsymbol{\delta} \in \mathbb{R}^n$ and non-negative $\boldsymbol{\mu} \in \mathbb{R}_+^n$ is given by

$$\begin{aligned}\mathcal{L} = \frac{1}{2}\|\boldsymbol{w}_\oplus\|_2^2 + \frac{1}{2}\|\boldsymbol{w}_\ominus\|_2^2 &+ \sum_{i\in S_1}\delta_i\left(\boldsymbol{w}_\oplus^\top\boldsymbol{x}_i - y_i\right) + \sum_{i\in S_1}\mu_i\left(\boldsymbol{w}_\ominus^\top\boldsymbol{x}_i\right) \\ &+ \sum_{i\in S_2}\delta_i\left(\boldsymbol{w}_\oplus^\top\boldsymbol{x}_i - \boldsymbol{w}_\ominus^\top\boldsymbol{x}_i - y_i\right) - \sum_{i\in S_2}\mu_i\left(\boldsymbol{w}_\ominus^\top\boldsymbol{x}_i\right) \\ &+ \sum_{i\in S_3}\delta_i\left(-\boldsymbol{w}_\ominus^\top\boldsymbol{x}_i - y_i\right) + \sum_{i\in S_3}\mu_i\left(\boldsymbol{w}_\oplus^\top\boldsymbol{x}_i\right) \\ &+ \sum_{i\in S_4}\delta_i\left(\boldsymbol{w}_\oplus^\top\boldsymbol{x}_i - \boldsymbol{w}_\ominus^\top\boldsymbol{x}_i - y_i\right) - \sum_{i\in S_4}\mu_i\left(\boldsymbol{w}_\oplus^\top\boldsymbol{x}_i\right).\end{aligned}$$

The KKT conditions are given below.
Stationarity:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}_\oplus} = \boldsymbol{w}_\oplus^\star + \sum_{i \in S_1} \delta_i^\star \boldsymbol{x}_i + \sum_{i \in S_2} \delta_i^\star \boldsymbol{x}_i + \sum_{i \in S_3} \mu_i^\star \boldsymbol{x}_i + \sum_{i \in S_4} (\delta_i^\star - \mu_i^\star) \boldsymbol{x}_i = 0,$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}_\ominus} = \boldsymbol{w}_\ominus^\star + \sum_{i \in S_1} \mu_i^\star \boldsymbol{x}_i - \sum_{i \in S_2} (\delta_i^\star + \mu_i^\star) \boldsymbol{x}_i - \sum_{i \in S_3} \delta_i^\star \boldsymbol{x}_i - \sum_{i \in S_4} \delta_i^\star \boldsymbol{x}_i = 0,$$

$$\Leftrightarrow \boldsymbol{w}_\oplus^\star = - \sum_{i \in S_1} \delta_i^\star \boldsymbol{x}_i - \sum_{i \in S_2} \delta_i^\star \boldsymbol{x}_i - \sum_{i \in S_3} \mu_i^\star \boldsymbol{x}_i + \sum_{i \in S_4} (-\delta_i^\star + \mu_i^\star) \boldsymbol{x}_i, \tag{33}$$

$$\boldsymbol{w}_\ominus^\star = - \sum_{i \in S_1} \mu_i^\star \boldsymbol{x}_i + \sum_{i \in S_2} (\delta_i^\star + \mu_i^\star) \boldsymbol{x}_i + \sum_{i \in S_3} \delta_i^\star \boldsymbol{x}_i + \sum_{i \in S_4} \delta_i^\star \boldsymbol{x}_i. \tag{34}$$

Primal feasibility:

$$\begin{aligned}
&\boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i = y_i, && \boldsymbol{w}_\ominus^{\star\top} \boldsymbol{x}_i \le 0, && \text{for all } i \in S_1,\\
&\boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i - \boldsymbol{w}_\ominus^{\star\top} \boldsymbol{x}_i = y_i, && -\boldsymbol{w}_\ominus^{\star\top} \boldsymbol{x}_i \le 0, && \text{for all } i \in S_2,\\
&-\boldsymbol{w}_\ominus^{\star\top} \boldsymbol{x}_i = y_i, \quad \boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i \le 0, && && \text{for all } i \in S_3,\\
&\boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i - \boldsymbol{w}_\ominus^{\star\top} \boldsymbol{x}_i = y_i, -\boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i \le 0, && && \text{for all } i \in S_4.
\end{aligned}$$

Dual feasibility:

$$\begin{aligned}
\delta_i^\star &\in \mathbb{R}, \text{ for all } i \in [n],\\
\mu_i^\star &\ge 0, \text{ for all } i \in [n].
\end{aligned}$$

Complementary slackness:

$$\sum_{i \in S_1} \mu_i^\star \left(\boldsymbol{w}_\ominus^{\star\top} \boldsymbol{x}_i \right) + \sum_{i \in S_2} \mu_i^\star \left(-\boldsymbol{w}_\ominus^{\star\top} \boldsymbol{x}_i \right) + \sum_{i \in S_3} \mu_i^\star \left(\boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i \right) + \sum_{i \in S_4} \mu_i^\star \left(-\boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i \right) = 0.$$

Note that the representation of $\boldsymbol{w}_\oplus^\star$ and $\boldsymbol{w}_\ominus^\star$ shares the parameters $\{\delta_i^\star : i \in S_2 \cup S_4\}$. As a result, since we define $\boldsymbol{w}_\oplus^\star = \boldsymbol{X}^\top \boldsymbol{\alpha}_\oplus^\star$ and $\boldsymbol{w}_\ominus^\star = \boldsymbol{X}^\top \boldsymbol{\alpha}_\ominus^\star$, we can write $\alpha_{\oplus,i}^\star$ and $\alpha_{\ominus,i}^\star$ in terms of δ_i and μ_i for all $i \in [n]$ by Equations (33) and (34) as

$$\alpha_{\oplus,i}^\star = \begin{cases} -\delta_i^\star & \forall i \in S_1 \\ -\delta_i^\star & \forall i \in S_2 \\ -\mu_i^\star & \forall i \in S_3 \\ -\delta_i^\star + \mu_i^\star & \forall i \in S_4 \end{cases}, \text{ and } \alpha_{\ominus,i}^\star = \begin{cases} -\mu_i^\star & \forall i \in S_1 \\ \delta_i^\star + \mu_i^\star & \forall i \in S_2 \\ \delta_i^\star & \forall i \in S_3 \\ \delta_i^\star & \forall i \in S_4 \end{cases}.$$

Step 3: Upper bound for $\left\| \boldsymbol{w}_\oplus^{(\infty)} - \boldsymbol{w}_\oplus^\star \right\|_2$.

We now generalize the proof of Theorem 5. We first relate the distance between the predictors $\boldsymbol{w}_\oplus^{(\infty)}$ and $\boldsymbol{w}_\oplus^\star$ to the distance in their predictions, i.e., $\left\| \boldsymbol{X}\boldsymbol{w}_\oplus^{(\infty)} - \boldsymbol{X}\boldsymbol{w}_\oplus^\star \right\|_2$. Since both vectors lie in the span of the data $\{\boldsymbol{x}_i\}_{i=1}^n$, their difference has no component in the null space corresponding to the smallest $d-n$ eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$. Therefore, we have

$$\left\| \boldsymbol{X}\boldsymbol{w}_\oplus^{(\infty)} - \boldsymbol{X}\boldsymbol{w}_\oplus^\star \right\|_2^2 = \left\| \boldsymbol{X} \left(\boldsymbol{w}_\oplus^{(\infty)} - \boldsymbol{w}_\oplus^\star \right) \right\|_2^2 \ge \mu_n(\boldsymbol{X}^\top \boldsymbol{X}) \left\| \boldsymbol{w}_\oplus^{(\infty)} - \boldsymbol{w}_\oplus^\star \right\|_2^2 = \mu_n(\boldsymbol{X}\boldsymbol{X}^\top) \left\| \boldsymbol{w}_\oplus^{(\infty)} - \boldsymbol{w}_\oplus^\star \right\|_2^2. \tag{35}$$

As a result, to derive an upper bound for $\left\|\boldsymbol{w}_{\oplus}^{(\infty)}-\boldsymbol{w}_{\oplus}^{\star}\right\|_2$, it suffices to upper bound the distance between their predictions $\left\|\boldsymbol{X}\boldsymbol{w}_{\oplus}^{(\infty)}-\boldsymbol{X}\boldsymbol{w}_{\oplus}^{\star}\right\|_2$. We begin with analyzing $\boldsymbol{w}_{\oplus}^{(\infty)}$. By Theorem 7, $\boldsymbol{w}_{\oplus}^{(\infty)}$ satisfies the following

$$\boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i = y_i \qquad \text{for all } y_i > 0, \tag{36a}$$

$$\alpha_{\oplus,j}^{(\infty)} = \alpha_{\oplus,j}^{(1)} = \eta(y_j - \epsilon_{\oplus,j} + \epsilon_{\ominus,j} + \frac{1}{\eta}\boldsymbol{e}_j^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{\epsilon}_{\oplus}) \qquad \text{for all } y_j < 0, \tag{36b}$$

and also all the conditions in Lemma 15.

We know that $\boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i = \boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i = y_i$ for all $i \in S_1$, and $\boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i = y_i$ and $\boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i = y_i + \boldsymbol{w}_{\ominus}^{\star\top}\boldsymbol{x}_i$ for all $i \in S_2$. Therefore, we can write

$$\begin{aligned}\left\|\boldsymbol{X}\boldsymbol{w}_{\oplus}^{(\infty)}-\boldsymbol{X}\boldsymbol{w}_{\oplus}^{\star}\right\|_2^2 &= \sum_{i=1}^n\left(\boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i\right)^2\\ &= \sum_{i\in S_2}\left(-\boldsymbol{w}_{\ominus}^{\star\top}\boldsymbol{x}_i\right)^2 + \sum_{i\in S_3}\left(\boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i\right)^2 + \sum_{i\in S_4}\left(\boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i\right)^2.\end{aligned} \tag{37}$$

We start with upper bounding the term $(-\boldsymbol{w}_{\ominus}^{\star\top}\boldsymbol{x}_i)^2$ for all $i \in S_2$. For $i \in S_2$, by the complementary slackness, we either have $-\boldsymbol{w}_{\ominus}^{\star\top}\boldsymbol{x}_i = 0$ with $\mu_i^\star \geq 0$ or $-\boldsymbol{w}_{\ominus}^{\star\top}\boldsymbol{x}_i \leq 0$ with $\mu_i^\star = 0$. In the first case, we have $(-\boldsymbol{w}_{\ominus}^{\star\top}\boldsymbol{x}_i)^2 = 0$. In the second case, we define $\tilde{S}_2 \subseteq S_2$ such that $\mu_i^\star = 0$ and $-\boldsymbol{w}_{\ominus}^{\star\top}\boldsymbol{x}_i \leq 0$ for all $i \in \tilde{S}_2$, and we will show that $\tilde{S}_2$ is empty with probability at least $1 - 2\exp(-n(Cc - \ln 9))$. Since $\boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i - \boldsymbol{w}_{\ominus}^{\star\top}\boldsymbol{x}_i = y_i$ for all $i \in \tilde{S}_2 \subseteq S_2$, we have

$$\begin{aligned}y_i = \boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i - \boldsymbol{w}_{\ominus}^{\star\top}\boldsymbol{x}_i &= \boldsymbol{e}_i^\top\boldsymbol{X}\boldsymbol{X}^\top\left(\boldsymbol{\alpha}_{\oplus}^\star - \boldsymbol{\alpha}_{\ominus}^\star\right)\\ &= \|\boldsymbol{\lambda}\|_1\left(-2\delta_i^\star - \underbrace{\mu_i^\star}_{=0}\right) + \boldsymbol{e}_i^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1\boldsymbol{I}\right)\left(\boldsymbol{\alpha}_{\oplus}^\star - \boldsymbol{\alpha}_{\ominus}^\star\right).\end{aligned}$$

The, for all $i \in \tilde{S}_2$, we have

$$\delta_i^\star = \frac{y_i - \boldsymbol{e}_i^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1\boldsymbol{I}\right)\left(\boldsymbol{\alpha}_{\oplus}^\star - \boldsymbol{\alpha}_{\ominus}^\star\right)}{-2\|\boldsymbol{\lambda}\|_1}.$$

Based on this representation of $\delta_i^\star$, for all $i \in \tilde{S}_2$, we have

$$
\begin{aligned}
\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i &= \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\alpha}_\ominus^\star \\
&= \|\boldsymbol{\lambda}\|_1 (\delta_i^\star) + \boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right) \boldsymbol{\alpha}_\ominus^\star \\
&= -\frac{y_i}{2} + \frac{1}{2}\boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right)(\boldsymbol{\alpha}_\oplus^\star + \boldsymbol{\alpha}_\ominus^\star) \\
&\le -\frac{y_i}{2} + \frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right\|_2 \left(\|\boldsymbol{\alpha}_\oplus^\star\|_2 + \|\boldsymbol{\alpha}_\ominus^\star\|_2\right) \\
&\le \|\boldsymbol{\lambda}\|_1 \left[-\frac{y_i}{2\|\boldsymbol{\lambda}\|_1} + \frac{1}{2}C\cdot\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\cdot\frac{2\sqrt{2}C_g\sqrt{n}y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right] \\
&\le \|\boldsymbol{\lambda}\|_1 \left[-\frac{y_{\min}}{2\|\boldsymbol{\lambda}\|_1} + \frac{1}{2}C\cdot\frac{y_{\min}}{C_0 y_{\max}}\cdot\frac{2\sqrt{2}C_g y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right] \\
&< 0,
\end{aligned}
$$

where the inequalities above apply Corollary 13 and the upper bound of $\|\boldsymbol{\alpha}_\oplus^\star\|_2$, $\|\boldsymbol{\alpha}_\ominus^\star\|_2$ from Step 1, and substitute $d_2 \ge C_0^2 \frac{n^2 y_{\max}^2}{y_{\min}^2}$ and $d_\infty \ge C_0 \frac{n^{1.5} y_{\max}}{y_{\min}}$ in Assumption 2 with $C_0 \gtrsim C_\alpha^2$ and $C_\alpha \gtrsim \max\{C_g^2, C_y C_g\}$. However, $\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i < 0$ contradicts with the condition $-\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i \le 0$ for $i \in \tilde{S}_2$. Therefore, $\tilde{S}_2 = \varnothing$. By combining these two cases, we conclude that $\sum_{i\in S_2}\left(-\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i\right)^2 = 0$.

Next, we upper bound the term $\left(\boldsymbol{w}_\oplus^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i\right)^2$ for all $i \in S_3$. We know that $\boldsymbol{w}_\oplus^{(\infty)\top}\boldsymbol{x}_i < 0$ in Theorem 7. For $\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i$ with $i \in S_3$, by the complementary slackness, we either have $\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i = 0$ with $\mu_i^\star \ge 0$ or $\mu_i^\star = 0$ with $\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i \le 0$. In the first case, $\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i = 0$, we have

$$
\begin{aligned}
\boldsymbol{w}_\oplus^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i &= \boldsymbol{w}_\oplus^{(\infty)\top}\boldsymbol{x}_i \\
&= \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\alpha}_\oplus^{(\infty)} \\
&= \boldsymbol{e}_i^\top \left[\|\boldsymbol{\lambda}\|_1 \boldsymbol{I} + \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right)\right] \boldsymbol{\alpha}_\oplus^{(\infty)} \\
&= \|\boldsymbol{\lambda}\|_1 \alpha_{\oplus,i}^{(\infty)} + \boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right) \boldsymbol{\alpha}_\oplus^{(\infty)} \\
&\ge \|\boldsymbol{\lambda}\|_1 \alpha_{\oplus,i}^{(\infty)} - \left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right\|_2 \left\|\boldsymbol{\alpha}_\oplus^{(\infty)}\right\|_2 \\
&\ge \|\boldsymbol{\lambda}\|_1 \left[\alpha_{\oplus,i}^{(\infty)} - C\cdot\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\left\|\boldsymbol{\alpha}_\oplus^{(\infty)}\right\|_2\right],
\end{aligned}
$$

where the last inequality applies Corollary 13. Substituting the bounds of $\alpha_{\oplus,i}^{(\infty)}$ and $\left\|\boldsymbol{\alpha}_\oplus^{(\infty)}\right\|_2$ from Lemma 15, we have

$$
\begin{aligned}
\boldsymbol{w}_\oplus^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i &\ge \|\boldsymbol{\lambda}\|_1 \left[-\frac{3y_{\max}}{C_g\|\boldsymbol{\lambda}\|_1} - C\cdot\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\frac{C_\alpha y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right] \\
&\ge \|\boldsymbol{\lambda}\|_1 \left[-\frac{3y_{\max}}{C_g\|\boldsymbol{\lambda}\|_1} - C\cdot\frac{y_{\min}}{C_0 y_{\max}}\frac{C_\alpha y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right] \\
&\ge -\frac{4}{C_g}y_{\max},
\end{aligned}
$$

where the inequalities substitute $d_2 \geq C_0^2 \frac{n^2 y_{\max}^2}{y_{\min}^2}$ and $d_\infty \geq C_0 \frac{n^{1.5} y_{\max}}{y_{\min}}$ in Assumption 2 with $C_0 \gtrsim C_\alpha^2$ and $C_\alpha \gtrsim \max\{C_g^2, C_y C_g\}$. In the second case, $\alpha_{\oplus,i}^\star = -\mu_i^\star = 0$ for $i \in S_3$, we have

$$\begin{aligned}
\boldsymbol{w}_\oplus^{(\infty)\top} \boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i &= \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \left(\boldsymbol{\alpha}_\oplus^{(\infty)} - \boldsymbol{\alpha}_\oplus^\star\right) \\
&= \boldsymbol{e}_i^\top \left[\|\boldsymbol{\lambda}\|_1 \boldsymbol{I} + \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right)\right] \left(\boldsymbol{\alpha}_\oplus^{(\infty)} - \boldsymbol{\alpha}_\oplus^\star\right) \\
&= \|\boldsymbol{\lambda}\|_1 \alpha_{\oplus,i}^{(\infty)} + \boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right) \left(\boldsymbol{\alpha}_\oplus^{(\infty)} - \boldsymbol{\alpha}_\oplus^\star\right) \\
&\geq \|\boldsymbol{\lambda}\|_1 \alpha_{\oplus,i}^{(\infty)} - \left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right\|_2 \left(\left\|\boldsymbol{\alpha}_\oplus^{(\infty)}\right\|_2 + \left\|\boldsymbol{\alpha}_\oplus^\star\right\|_2\right) \\
&\geq \|\boldsymbol{\lambda}\|_1 \left[-\frac{3 y_{\max}}{C_g \|\boldsymbol{\lambda}\|_1} - C \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right) \left(\frac{C_\alpha \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1} + \frac{\sqrt{2} C_g \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right)\right] \\
&\geq \|\boldsymbol{\lambda}\|_1 \left[-\frac{3 y_{\max}}{C_g \|\boldsymbol{\lambda}\|_1} - C \cdot \frac{y_{\min}}{C_0 y_{\max}} \left(\frac{C_\alpha y_{\max}}{\|\boldsymbol{\lambda}\|_1} + \frac{\sqrt{2} C_g y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right)\right] \\
&\geq -\frac{4}{C_g} y_{\max},
\end{aligned}$$

by applying the same argument and the upper bound from Step 1 that $\left\|\boldsymbol{\alpha}_\oplus^\star\right\|_2 \leq \frac{\sqrt{2} C_g \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1}$. Therefore, we have $\left(\boldsymbol{w}_\oplus^{(\infty)\top} \boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i\right)^2 \leq \frac{16}{C_g^2} y_{\max}^2$ for all $i \in S_3$.

Next, we upper bound the term $\left(\boldsymbol{w}_\oplus^{(\infty)\top} \boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i\right)^2$ for all $i \in S_4$ in a similar way compared to S_2. For $i \in S_4$, by the complementary slackness, we either have $-\boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i = 0$ with $\mu_i^\star \geq 0$ or $-\boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i \leq 0$ with $\mu_i^\star = 0$. In the first case, $(-\boldsymbol{w}_\ominus^{\star\top} \boldsymbol{x}_i)^2 = 0$, and we have $\left(\boldsymbol{w}_\oplus^{(\infty)\top} \boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i\right)^2 = \left(\boldsymbol{w}_\oplus^{(\infty)\top} \boldsymbol{x}_i\right)^2$. Therefore, we can reuse the upper bound we derived in S_3 such that $0 \geq \boldsymbol{w}_\oplus^{(\infty)\top} \boldsymbol{x}_i \geq -\frac{4}{C_g} y_{\max}$. In the second case, we define $\tilde{S}_4 \subseteq S_4$ such that $\mu_i^\star = 0$ and $-\boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i \leq 0$ for all $i \in \tilde{S}_4$, and we will show that $\tilde{S}_4$ is empty with probability at least $1 - 2\exp(-n(Cc - \ln 9))$. Since $\boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i - \boldsymbol{w}_\ominus^{\star\top} \boldsymbol{x}_i = y_i$ for all $i \in \tilde{S}_4 \subseteq S_4$, we have

$$\begin{aligned}
y_i = \boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i - \boldsymbol{w}_\ominus^{\star\top} \boldsymbol{x}_i &= \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \left(\boldsymbol{\alpha}_\oplus^\star - \boldsymbol{\alpha}_\ominus^\star\right) \\
&= \|\boldsymbol{\lambda}\|_1 (-2\delta_i^\star + \underbrace{\mu_i^\star}_{=0}) + \boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right) \left(\boldsymbol{\alpha}_\oplus^\star - \boldsymbol{\alpha}_\ominus^\star\right).
\end{aligned}$$

For all $i \in \tilde{S}_4$, we have

$$\delta_i^\star = \frac{y_i - \boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right) \left(\boldsymbol{\alpha}_\oplus^\star - \boldsymbol{\alpha}_\ominus^\star\right)}{-2 \|\boldsymbol{\lambda}\|_1}.$$

Based on this representation of $\delta_i^\star$, for all $i \in \tilde{S}_4$, we have

$$\begin{aligned}
\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i &= \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\alpha}_\oplus^\star \\
&= \|\boldsymbol{\lambda}\|_1 (-\delta_i^\star) + \boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right)\boldsymbol{\alpha}_\oplus^\star \\
&= \frac{y_i}{2} + \frac{1}{2}\boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right)(\boldsymbol{\alpha}_\oplus^\star + \boldsymbol{\alpha}_\ominus^\star) \\
&\le \frac{y_i}{2} + \frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right\|_2 \left(\|\boldsymbol{\alpha}_\oplus^\star\|_2 + \|\boldsymbol{\alpha}_\ominus^\star\|_2\right) \\
&\le \|\boldsymbol{\lambda}\|_1 \left[\frac{y_i}{2\|\boldsymbol{\lambda}\|_1} + \frac{1}{2}C \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right) \cdot \frac{2\sqrt{2}C_g\sqrt{n}y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right] \\
&\le \|\boldsymbol{\lambda}\|_1 \left[-\frac{y_{\min}}{2\|\boldsymbol{\lambda}\|_1} + \frac{1}{2}C \cdot \frac{y_{\min}}{C_0 y_{\max}} \cdot \frac{2\sqrt{2}C_g y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right] \\
&< 0,
\end{aligned}$$

where the inequalities apply Corollary 13 and the upper bound of $\left\|\boldsymbol{\alpha}_\oplus^\star\right\|_2$ in Step 1, and substitute $d_2 \ge C_0^2 \frac{n^2 y_{\max}^2}{y_{\min}^2}$ and $d_\infty \ge C_0 \frac{n^{1.5} y_{\max}}{y_{\min}}$ in Assumption 2 with $C_0 \gtrsim C_\alpha^2$ and $C_\alpha \gtrsim \max\{C_g^2, C_y C_g\}$. However, $\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i < 0$ contradicts with the condition $-\boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i \le 0$ for $i \in \tilde{S}_4$. Therefore, $\tilde{S}_4 = \varnothing$. By combining these two cases, we have $\sum_{i\in S_4}\left(\boldsymbol{w}_\oplus^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i\right)^2 = \sum_{i\in S_4}\left(\boldsymbol{w}_\oplus^{(\infty)\top}\boldsymbol{x}_i\right)^2 \le \sum_{i\in S_4}\frac{16}{C_g^2}y_{\max}^2$.

Substituting the upper bounds into Equation (37) gives us

$$\begin{aligned}
\left\|\boldsymbol{X}\boldsymbol{w}^{(\infty)} - \boldsymbol{X}\boldsymbol{w}^\star\right\|_2^2 &= \sum_{i\in S_2}\left(-\boldsymbol{w}_\ominus^{\star\top}\boldsymbol{x}_i\right)^2 + \sum_{i\in S_3}\left(\boldsymbol{w}_\oplus^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i\right)^2 + \sum_{i\in S_4}\left(\boldsymbol{w}_\oplus^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top}\boldsymbol{x}_i\right)^2. \\
&\le \sum_{i\in S_3}\frac{16}{C_g^2}y_{\max}^2 + \sum_{i\in S_4}\frac{16}{C_g^2}y_{\max}^2 \\
&= \frac{16}{C_g^2}n_- y_{\max}^2.
\end{aligned} \tag{38}$$

Finally, putting together Equation (35) and (38), we have

$$\left\|\boldsymbol{w}^{(\infty)} - \boldsymbol{w}^\star\right\|_2^2 \le \frac{\left\|\boldsymbol{X}\boldsymbol{w}^{(\infty)} - \boldsymbol{X}\boldsymbol{w}^\star\right\|_2^2}{\mu_n(\boldsymbol{X}\boldsymbol{X}^\top)} \le \frac{16 n_- y_{\max}^2}{C_g\|\boldsymbol{\lambda}\|_1},$$

which completes the proof of the upper bound.

Step 4: Lower bound for $\left\|\boldsymbol{w}_\oplus^{(\infty)} - \boldsymbol{w}_\oplus^\star\right\|_2$.

Now, we derive the lower bound of $\left\|\boldsymbol{w}_\oplus^{(\infty)} - \boldsymbol{w}_\oplus^\star\right\|_2$ in a similar approach. We again start with the prediction distance

$$\left\|\boldsymbol{X}\boldsymbol{w}_\oplus^{(\infty)} - \boldsymbol{X}\boldsymbol{w}_\oplus^\star\right\|_2^2 = \left\|\boldsymbol{X}\left(\boldsymbol{w}_\oplus^{(\infty)} - \boldsymbol{w}_\oplus^\star\right)\right\|_2^2 \le \mu_1(\boldsymbol{X}^\top\boldsymbol{X})\left\|\boldsymbol{w}_\oplus^{(\infty)} - \boldsymbol{w}_\oplus^\star\right\|_2^2 = \mu_1(\boldsymbol{X}\boldsymbol{X}^\top)\left\|\boldsymbol{w}_\oplus^{(\infty)} - \boldsymbol{w}_\oplus^\star\right\|_2^2. \tag{39}$$

Therefore, it suffices to lower bound $\left\|\boldsymbol{X}\boldsymbol{w}_{\oplus}^{(\infty)} - \boldsymbol{X}\boldsymbol{w}_{\oplus}^{\star}\right\|_2$ to get the lower bound of $\left\|\boldsymbol{w}_{\oplus}^{(\infty)} - \boldsymbol{w}_{\oplus}^{\star}\right\|_2$. By Equation (37), we have

$$\begin{aligned}\left\|\boldsymbol{X}\boldsymbol{w}_{\oplus}^{(\infty)} - \boldsymbol{X}\boldsymbol{w}_{\oplus}^{\star}\right\|_2^2 &= \sum_{i\in S_2}\left(-\boldsymbol{w}_{\ominus}^{\star\top}\boldsymbol{x}_i\right)^2 + \sum_{i\in S_3}\left(\boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i\right)^2 + \sum_{i\in S_4}\left(\boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i\right)^2 \\ &\geq \sum_{i\in S_3}\left(\boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i\right)^2 + \sum_{i\in S_4}\left(\boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i\right)^2. \end{aligned} \tag{40}$$

We omit the partition in S_2 because we have shown that $\sum_{i\in S_2}\left(-\boldsymbol{w}_{\ominus}^{\star\top}\boldsymbol{x}_i\right)^2 = 0$ with probability at least $1 - 2\exp(-n(Cc - \ln 9))$. Therefore, we need to lower bound $\left(\boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i\right)^2$ for $i \in S_3$ and $\left(\boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i\right)^2$ for $i \in S_4$.

We start with lower bounding $\left(\boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i\right)^2$ for $i \in S_3$. We know that $\boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i < 0$ in Theorem 7. For $\boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i$ with $i \in S_3$, by the complementary slackness, we either have $\boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i = 0$ with $\mu_i^\star \geq 0$ or $\mu_i^\star = 0$ with $\boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i \leq 0$. In the first case, $\boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i = 0$, we have

$$\begin{aligned}\boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i &= \boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i \\ &= \boldsymbol{e}_i^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{\alpha}_{\oplus}^{(\infty)} \\ &= \boldsymbol{e}_i^\top\left[\|\boldsymbol{\lambda}\|_1\boldsymbol{I} + \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1\boldsymbol{I}\right)\right]\boldsymbol{\alpha}_{\oplus}^{(\infty)} \\ &= \|\boldsymbol{\lambda}\|_1\alpha_{\oplus,i}^{(\infty)} + \boldsymbol{e}_i^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1\boldsymbol{I}\right)\boldsymbol{\alpha}_{\oplus}^{(\infty)} \\ &\leq \|\boldsymbol{\lambda}\|_1\alpha_{\oplus,i}^{(\infty)} + \left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1\boldsymbol{I}\right\|_2\left\|\boldsymbol{\alpha}_{\oplus}^{(\infty)}\right\|_2 \\ &\leq \|\boldsymbol{\lambda}\|_1\left[\alpha_{\oplus,i}^{(\infty)} + C\cdot\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\left\|\boldsymbol{\alpha}_{\oplus}^{(\infty)}\right\|_2\right],\end{aligned}$$

where the last inequality applies Corollary 13. Substituting the bounds of $\alpha_{\oplus,i}^{(\infty)}$ and $\left\|\boldsymbol{\alpha}_{\oplus}^{(\infty)}\right\|_2$ from Lemma 15, we have

$$\begin{aligned}\boldsymbol{w}_{\oplus}^{(\infty)\top}\boldsymbol{x}_i - \boldsymbol{w}_{\oplus}^{\star\top}\boldsymbol{x}_i &\leq \|\boldsymbol{\lambda}\|_1\left[-\frac{y_{\min}}{C_\alpha\|\boldsymbol{\lambda}\|_1} + C\cdot\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\frac{C_\alpha\sqrt{n}y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right] \\ &\leq \|\boldsymbol{\lambda}\|_1\left[-\frac{y_{\min}}{C_\alpha\|\boldsymbol{\lambda}\|_1} + C\cdot\frac{y_{\min}}{C_0 y_{\max}}\frac{C_\alpha y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right] \\ &\leq -(1 - \frac{C\cdot C_\alpha^2}{C_0})\frac{y_{\min}}{C_\alpha},\end{aligned}$$

where the inequalities substitute $d_2 \geq C_0^2 \frac{n^2 y_{\max}^2}{y_{\min}^2}$ and $d_\infty \geq C_0 \frac{n^{1.5} y_{\max}}{y_{\min}}$ in Assumption 2 with $C_0 \gtrsim C_\alpha^2$. In the second case, $\alpha_{\oplus,i}^\star = -\mu_i^\star = 0$ for $i \in S_3$, we have

$$\begin{aligned}
\boldsymbol{w}_\oplus^{(\infty)\top} \boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i &= \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \left(\boldsymbol{\alpha}_\oplus^{(\infty)} - \boldsymbol{\alpha}_\oplus^\star\right) \\
&= \boldsymbol{e}_i^\top \left[\|\boldsymbol{\lambda}\|_1 \boldsymbol{I} + \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right)\right] \left(\boldsymbol{\alpha}_\oplus^{(\infty)} - \boldsymbol{\alpha}_\oplus^\star\right) \\
&= \|\boldsymbol{\lambda}\|_1 \alpha_{\oplus,i}^{(\infty)} + \boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right) \left(\boldsymbol{\alpha}_\oplus^{(\infty)} - \boldsymbol{\alpha}_\oplus^\star\right) \\
&\leq \|\boldsymbol{\lambda}\|_1 \alpha_{\oplus,i}^{(\infty)} + \left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right\|_2 \left(\left\|\boldsymbol{\alpha}_\oplus^{(\infty)}\right\|_2 + \|\boldsymbol{\alpha}_\oplus^\star\|_2\right) \\
&\leq \|\boldsymbol{\lambda}\|_1 \left[-\frac{y_{\min}}{C_\alpha \|\boldsymbol{\lambda}\|_1} + C \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right) \left(\frac{C_\alpha \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1} + \frac{\sqrt{2} C_g \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right)\right] \\
&\leq \|\boldsymbol{\lambda}\|_1 \left[-\frac{y_{\min}}{C_\alpha \|\boldsymbol{\lambda}\|_1} + C \cdot \frac{y_{\min}}{C_0 y_{\max}} \left(\frac{C_\alpha y_{\max}}{\|\boldsymbol{\lambda}\|_1} + \frac{\sqrt{2} C_g y_{\max}}{\|\boldsymbol{\lambda}\|_1}\right)\right] \\
&\leq -(1 - \frac{2C \cdot C_\alpha^2}{C_0}) \frac{y_{\min}}{C_\alpha},
\end{aligned}$$

by applying the same argument and the upper bound from Step 1 that $\left\|\boldsymbol{\alpha}_\oplus^\star\right\|_2 \leq \frac{\sqrt{2} C_g \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1}$. Therefore, we have $\left(\boldsymbol{w}_\oplus^{(\infty)\top} \boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i\right)^2 \geq \left(1 - \frac{2C \cdot C_\alpha^2}{C_0}\right)^2 \frac{y_{\min}^2}{C_\alpha^2}$, for all $i \in S_3$.

Next, we lower bound the term $\left(\boldsymbol{w}_\oplus^{(\infty)\top} \boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i\right)^2$ for all $i \in S_4$. In Step 3, we already showed the two cases in $\boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i$ with $i \in S_4$ by the complementary slackness. In the first case, $(-\boldsymbol{w}_\ominus^{\star\top} \boldsymbol{x}_i)^2 = 0$, and we have $\left(\boldsymbol{w}_\oplus^{(\infty)\top} \boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i\right)^2 = \left(\boldsymbol{w}_\oplus^{(\infty)\top} \boldsymbol{x}_i\right)^2$. Therefore, we can reuse the lower bound we derived in S_3 such that $\boldsymbol{w}_\oplus^{(\infty)\top} \boldsymbol{x}_i \leq -(1 - \frac{C \cdot C_\alpha^2}{C_0}) \frac{y_{\min}}{C_\alpha}$. In the second case, we have shown that $\tilde{S}_4 = \varnothing$. By concluding two cases, we have $\sum_{i \in S_4} \left(\boldsymbol{w}_\oplus^{(\infty)\top} \boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i\right)^2 = \sum_{i \in S_4} \left(\boldsymbol{w}_\oplus^{(\infty)\top} \boldsymbol{x}_i\right)^2 \geq \sum_{i \in S_4} \left(1 - \frac{2C \cdot C_\alpha^2}{C_0}\right)^2 \frac{y_{\min}^2}{C_\alpha^2}$.

Substituting the lower bounds into Equation (40) gives us

$$\begin{aligned}
\left\|\boldsymbol{X}\boldsymbol{w}_\oplus^{(\infty)} - \boldsymbol{X}\boldsymbol{w}_\oplus^\star\right\|_2^2 &\geq \sum_{i \in S_3} \left(\boldsymbol{w}_\oplus^{(\infty)\top} \boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i\right)^2 + \sum_{i \in S_4} \left(\boldsymbol{w}_\oplus^{(\infty)\top} \boldsymbol{x}_i - \boldsymbol{w}_\oplus^{\star\top} \boldsymbol{x}_i\right)^2 \\
&\geq \sum_{i \in S_3} \left(1 - \frac{2C \cdot C_\alpha^2}{C_0}\right)^2 \frac{y_{\min}^2}{C_\alpha^2} + \sum_{i \in S_4} \left(1 - \frac{2C \cdot C_\alpha^2}{C_0}\right)^2 \frac{y_{\min}^2}{C_\alpha^2} \\
&= \frac{n_- y_{\min}^2}{\tilde{C}},
\end{aligned} \tag{41}$$

where we let $\tilde{C} := \frac{C_0^2 C_\alpha^2}{(C_0 - 2C \cdot C_\alpha^2)^2} > 1$. Finally, putting together Equation (39) and (41), we have

$$\left\|\boldsymbol{w}_\oplus^{(\infty)} - \boldsymbol{w}_\oplus^\star\right\|_2^2 \geq \frac{\left\|\boldsymbol{X}\boldsymbol{w}_\oplus^{(\infty)} - \boldsymbol{X}\boldsymbol{w}_\oplus^\star\right\|_2^2}{\mu_1(\boldsymbol{X}\boldsymbol{X}^\top)} \geq \frac{n_- y_{\min}^2}{\tilde{C} C_g \|\boldsymbol{\lambda}\|_1}.$$

This completes the proof of the lower bound. ■

Appendix F. Implicit Bias of Multiple ReLU Models ($m > 2$) Under Gradient Descent

In this section, we extend our analysis to multiple ReLU models trained with $m > 2$ neurons under stronger assumptions on the initialization. We consider models of the form: $h_{\boldsymbol{\Theta}}(\boldsymbol{x}) := h_{\{\boldsymbol{w}_k\}_{k=1}^m}(\boldsymbol{x}) = \sum_{k=1}^m s_k \sigma(\boldsymbol{w}_k^\top \boldsymbol{x})$, where $\boldsymbol{w}_k \in \mathbb{R}^d$ are the model weights and there are at least one positive neuron and one negative neuron. The parameter set is hence denoted by $\boldsymbol{\Theta} = \{\boldsymbol{w}_k\}_{k=1}^m$. The empirical risk is defined in (1) as

$$\mathcal{R}(\boldsymbol{\Theta}) = \frac{1}{2}\sum_{i=1}^n \big(h_{\boldsymbol{\Theta}}(\boldsymbol{x}_i) - y_i\big)^2 = \frac{1}{2}\sum_{i=1}^n \left(\sum_{k=1}^m s_k \sigma(\boldsymbol{w}_k^\top \boldsymbol{x}_i) - y_i\right)^2 .$$

Here, we fix $s_k \in \{\pm 1\}$ and only train the hidden weights $\{\boldsymbol{w}_k\}_{k=1}^m$.

F.1. Gradient Descent Updates and Convergence

The gradient of the empirical risk in (1) with respect to $\boldsymbol{w}_k$ is given in (2) as

$$\boldsymbol{w}_k^{(t+1)} = \boldsymbol{w}_k^{(t)} - \eta \nabla_{\boldsymbol{w}_k} \mathcal{R}(\boldsymbol{\Theta}^{(t)}) = \boldsymbol{w}_k^{(t)} - \eta s_k \boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}_k^{(t)})\big(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}\big). \tag{42}$$

The primal–dual gradient update in (4) is given by

$$\text{(Primal)} \qquad \boldsymbol{\beta}_k^{(t+1)} = \boldsymbol{\beta}_k^{(t)} - \eta s_k \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\beta}_k^{(t)})(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}), \tag{43a}$$

$$\text{(Dual)} \qquad \boldsymbol{\alpha}_k^{(t+1)} = \boldsymbol{\alpha}_k^{(t)} - \eta s_k \boldsymbol{D}(\boldsymbol{\beta}^{(t)})(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}). \tag{43b}$$

Next, we consider a regime in which, after some time t_0, each neuron activates on a fixed subset of training examples, and this activation pattern remains unchanged throughout the subsequent dynamics. Moreover, these active subsets are disjoint across different neurons. That is, for every training example, at most one neuron is active, while each neuron may be active on a subset of examples. In this regime, each neuron effectively reduces to a linear model trained only on its own active examples.

Lemma 16 *Consider a multiple ReLU model $h_{\boldsymbol{\Theta}}$. For each neuron $k \in [m]$, suppose there exists iteration $t_0 \geq 0$ such that*

1. *At time t_0, the subset of examples on which the k-th neuron is active is disjoint from the subsets activated by all other neurons, i.e., $\boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}_k^{(t_0)})\boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}_\ell^{(t_0)}) = \boldsymbol{0}_{n\times n}$ for any $\ell \neq k$.*

2. *The activation pattern of the k-th remains unchanged after time t_0, i.e., $\boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}_k^{(t_0)}) = \boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}_k^{(t)})$ for all $t \geq t_0$.*

Then, for all $t \geq t_0$, and each $k \in [m]$, the gradient descent dynamics of the k-th neuron are equivalent to gradient descent applied to a linear model, initialized at $\boldsymbol{w}_k^{(t_0)}$, and trained using only the subset of samples satisfying $\boldsymbol{x}_i^\top \boldsymbol{w}_k^{(t_0)} > 0$.

The proof of Lemma 16 is provided in Appendix G.1.

F.2. Minimum-ℓ_2-norm Solution of Multiple ReLU Models

The minimum-ℓ_2-norm solution for the multiple ReLU regression in (5) is given by

$$\{\boldsymbol{w}_k^\star\}_{k=1}^m = \arg\min_{\{\boldsymbol{w}_k\}_{k=1}^m} \frac{1}{2}\sum_{k=1}^m \|\boldsymbol{w}_k\|_2^2 \tag{44}$$
$$\text{s.t.} \sum_{k=1}^m s_k \sigma(\boldsymbol{w}_k^\top \boldsymbol{x}_i) = y_i, \text{ for all } i \in [n].$$

F.3. High-dimensional Implicit Bias of Multiple ReLU Models

In this section, we characterize the implicit bias of multiple ReLU models trained by gradient descent in the high-dimensional regime. We identify a setup in which each neuron is only active toward a fixed and disjoint subset of training examples, where the labels y_i of these examples have the same sign as the neuron's sign s_k.

To formalize this setup, we introduce an assignment vector $\boldsymbol{a} \in [m]^n$, where each entry $a_i \in [m]$ indicates which neuron is responsible for example i.

Assumption 3 *For a multiple ReLU model, we assume that there exists an assignment vector $\boldsymbol{a} \in [m]^n$ such that for each example $i \in [n]$, $a_i = k$, for some neuron k satisfying $s_k \cdot y_i > 0$. For each neuron $k \in [m]$, define a diagonal matrix $\boldsymbol{A}_k \in \mathbb{R}^{n\times n}$ with diagonal entries*

$$(\boldsymbol{A}_k)_{ii} = \begin{cases} 0, & \text{if } a_i = k, \text{ or } s_k \cdot y_i < 0 \\ -\operatorname{sign}(y_i), & \text{otherwise} \end{cases}.$$

Assumption 3 is used to design a proper initialization which ensures that the gradient descent can converge to the desired regime. In this regime, we show that if a neuron's primal variable $\beta_{k,i}$ is positive and the sign of the neuron agrees with the label (i.e., $s_k \cdot y_i > 0$), then the corresponding example remains active throughout training. Conversely, if the associated dual variable $\alpha_{k,j}$ stays sufficiently negative, it remains frozen and is no longer updated.

Theorem 17 *Under Assumptions 1, 2 and 3, suppose we choose $\boldsymbol{w}_k^{(0)} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\left(\frac{1}{C_g}\boldsymbol{A}_k\boldsymbol{y} + \boldsymbol{\epsilon}_k\right)$, where $0 < \epsilon_{k,i} \leq \frac{1}{C\alpha m} y_{\min}$ for all $k \in [m]$ and $i \in [n]$, and the gradient descent step size satisfies $\frac{1}{CC_g\|\boldsymbol{\lambda}\|_1} \leq \eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$. Then, the gradient descent limit $\boldsymbol{w}_k^{(\infty)}$ for multiple ReLU models coincides with the solution obtained by training a linear model on disjoint subsets of examples, initialized at $\boldsymbol{w}_k^{(1)}$ with probability at least $1 - 2\exp(-cn)$. Formally, we have $\boldsymbol{w}_k^{(\infty)} = \arg\min_{\boldsymbol{w}\in\{\boldsymbol{w}:\boldsymbol{X}_{S_k}\boldsymbol{w}=\boldsymbol{y}_{S_k}\}} \left\|\boldsymbol{w} - \boldsymbol{w}_k^{(1)}\right\|_2$ and $\boldsymbol{X}_{S_k^c}\boldsymbol{w}_k^{(\infty)} \preceq \boldsymbol{0}$, where $S_k \coloneqq \{i \in [n] : a_i = k\}$.*

The full proof is provided in Appendix G.2. Note that the initialization, constructed by the matrices $\boldsymbol{A}_k$, ensures that each training example i is activated by exactly one neuron that matches its sign—namely, the a_i-th neuron. All other neurons with the same sign remain inactive on this example.

F.4. Approximation to Minimum-ℓ_2-norm Solution in High Dimensions

In this section, we show that in high dimensions, the implicit bias solution for multiple ReLU models derived in Theorem 17 is close to the corresponding minimum-ℓ_2-norm solution $\{\boldsymbol{w}_k^\star\}_{k=1}^m$ defined in (44).

Theorem 18 *Under Assumptions 1, 2 and 3, suppose we choose* $\boldsymbol{w}_k^{(0)} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\left(\frac{1}{C_g}\boldsymbol{A}_k\boldsymbol{y} + \boldsymbol{\epsilon}_k\right)$, *where* $0 < \epsilon_{k,i} \leq \frac{1}{C_\alpha m} y_{\min}$ *for all* $k \in [m]$ *and* $i \in [n]$, *and the gradient descent step size satisfies* $\frac{1}{CC_g\|\boldsymbol{\lambda}\|_1} \leq \eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$. *Then, we have* $\sqrt{\sum_{k=1}^m \left\|\boldsymbol{w}_k^{(\infty)} - \boldsymbol{w}_k^\star\right\|_2^2} \leq \sqrt{\frac{4C_g C_\alpha^2 mny_{\max}^2}{\|\boldsymbol{\lambda}\|_1}}$ *with probability at least* $1 - 2\exp(-cn)$.

The proof is deferred to Appendix G.3. Note that since the minimum-ℓ_2-norm solution $\{\boldsymbol{w}_k^\star\}_{k=1}^m$ is more involved to characterize, Theorem 18 only provides an upper bound for the approximation of the implicit bias to $\{\boldsymbol{w}_k^\star\}_{k=1}^m$. A more fine-grained characterization, as well as a deeper understanding of the role of overparameterization, is left for future work.

Appendix G. Proofs for Multiple ReLU Models ($m > 2$) Trained with Gradient Descent

In this section, we present the proofs concerning the behavior of the multiple ReLU model trained with gradient descent.

G.1. Proof of Lemma 16 (Gradient Descent Convergence)

Proof (Lemma 16) This proof is analogous to Lemma 1. The key idea is to show that once the activation pattern becomes fixed after some iteration $t_0 \geq 0$, the gradient descent dynamics of each neuron are equivalent to those of a linear model trained on a fixed subset of examples.

Fix a neuron $k \in [m]$. Consider the linear model

$$h(\boldsymbol{x}) = s_k \boldsymbol{w}^\top \boldsymbol{x},$$

where $\boldsymbol{w} \in \mathbb{R}^d$ is the linear model parameter (also called weight). Let $S_k^{(t_0)} \subseteq [n]$ denote the active set of the k-th neuron at iteration t_0, defined by $S_k^{(t_0)} := \{i \in [n] : \boldsymbol{x}_i^\top \boldsymbol{w}_k^{(t_0)} > 0\}$. We define the empirical risk with the linear model using only the examples in $S_k^{(t_0)}$ as

$$\mathcal{R}_{S_k^{(t_0)}}(\boldsymbol{w}) = \frac{1}{2} \sum_{i \in S_k^{(t_0)}} (s_k \boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2.$$

The gradient descent update for this linear model is then given by

$$\begin{aligned} \boldsymbol{w}^{(t+1)} &= \boldsymbol{w}^{(t)} - \eta \nabla \mathcal{R}_{S_k^{(t_0)}}(\boldsymbol{w}^{(t)}) \\ &= \boldsymbol{w}^{(t)} - \eta s_k \sum_{i \in S_k^{(t_0)}} (s_k \boldsymbol{w}^{(t)\top} \boldsymbol{x}_i - y_i) \boldsymbol{x}_i. \end{aligned} \tag{45}$$

On the other hand, the original gradient descent update of the multiple ReLU model in Equation (42) tells us that

$$\boldsymbol{w}_k^{(t+1)} = \boldsymbol{w}_k^{(t)} - \eta s_k \boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}_k^{(t)}) \big(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}\big).$$

Under the first assumption in the lemma, $\boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}_k^{(t_0)})\boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}_\ell^{(t_0)}) = \boldsymbol{0}_{n\times n}$ for any $\ell \neq k$, the activation patterns of different neurons are disjoint at iteration t_0. Consequently, for any $i \in S_k^{(t_0)}$, only the k-th neuron is active, and therefore we have

$$h_{\boldsymbol{\Theta}^{(t_0)}}(\boldsymbol{x}_i) = \sum_{k=1}^m s_k \sigma(\boldsymbol{w}_k^\top \boldsymbol{x}_i) = s_k \boldsymbol{w}_k^\top \boldsymbol{x}_i.$$

Moreover, by the second assumption of the lemma, the activation pattern of the k-th neuron remains unchanged after iteration t_0, i.e., $\boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}_k^{(t_0)}) = \boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}_k^{(t)})$ for all $t \geq t_0$. Hence, for all $t \geq t_0$, the diagonal entries of $\boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}_k^{(t)})$ satisfy $D_{ii} = \mathbb{1}_{i \in S_k^{(t_0)}}$ for all $i \in [n]$. Therefore, for all $t \geq t_0$, the gradient update of the k-th neuron in the multiple ReLU model is given by

$$\begin{aligned} \boldsymbol{w}_k^{(t+1)} &= \boldsymbol{w}_k^{(t)} - \eta s_k \boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{X}\boldsymbol{w}_k^{(t)})(h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}) \\ &= \boldsymbol{w}^{(t)} - \eta s_k \sum_{i \in S_k^{(t_0)}} (s_k \boldsymbol{w}_k^{(t)\top} \boldsymbol{x}_i - y_i) \boldsymbol{x}_i. \end{aligned}$$

This update is identical to the gradient descent update of the linear model in Equation (45). Hence, for all $t \geq t_0$, the gradient descent dynamics of the k-th neuron in the multiple ReLU model are equivalent to those of a linear model trained using only the examples in $S_k^{(t_0)}$. This completes the proof of the lemma. ■

G.2. Proof of Theorem 17 (High-dimensional Implicit Bias)

In this section, we present the proof of Theorem 17. Before proceeding to the proof, we again introduce a set of sufficient conditions under which the active pattern for a neuron at iteration t will be preserved at iteration $t+1$. Similar to the single ReLU model and 2-ReLU model cases, our analysis relies on Lemma 9 and Lemma 10 to characterize the dynamics of primal and dual variables. Using these results, we establish Lemma 19, which characterizes that the active sets of all neurons remain unchanged across gradient descent iterations.

Lemma 19 *Under Assumption 1, 2 and 3, suppose the gradient descent step size satisfies* $\eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$. *For a multiple ReLU model, if the following five conditions hold at some iteration* $t \geq 0$, *then they also hold at iteration* $t+1$.

a. $\beta^{(t)}_{a_i,i} > 0$, *for all* $i \in [n]$.

b. $-\frac{3y_{\max}}{C_g\|\boldsymbol{\lambda}\|_1} \leq \alpha^{(t)}_{k,j} \leq -\frac{y_{\min}}{C_\alpha\|\boldsymbol{\lambda}\|_1}$, *for all* $j \in [n]$ *with* $k \neq a_j$.

c. $\left\|\boldsymbol{\beta}^{(t)}_{k,S_k} - s_k\boldsymbol{y}_{S_k}\right\|_2 \leq C_y\left\|\boldsymbol{y}_{S_k}\right\|_2$, *for all* $k \in [m]$.

d. $\left\|\boldsymbol{\alpha}^{(t)}_k\right\|_2 \leq \frac{C_\alpha\sqrt{n}y_{\max}}{\|\boldsymbol{\lambda}\|_1}$, *for all* $k \in [m]$.

e. $\beta^{(t)}_{k,j} \leq 0$, *for all* $j \in [n]$ *with* $k \neq a_j$.

Consequently, the activation pattern of each neuron remains unchanged from iteration t *to* $t+1$. *In the above, we define* $S_k \coloneqq \{i \in [n] : a_i = k\}$, *and for any vector* $\boldsymbol{v} \in \mathbb{R}^n$, *we use* $\boldsymbol{v}_{S_k}$ *to denote the subvector of entries indexed by* S_k.

Proof (Lemma 19) We now verify that these conditions are preserved from iteration t to $t+1$.

Part (a): By conditions (a), (c) and (e) at iteration t, we have

$$\left\|h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{X}) - \boldsymbol{y}\right\|_2^2 = \left\|\sum_{k=1}^m s_k\sigma(\boldsymbol{\beta}^{(t)}_k) - \boldsymbol{y}\right\|_2^2 = \sum_{k=1}^m \left\|s_k\left(\boldsymbol{\beta}^{(t)}_{k,S_k} - s_k\boldsymbol{y}_{S_k}\right)\right\|_2^2 \leq C_y^2\left\|\boldsymbol{y}\right\|_2^2,$$

where the last inequality uses the fact that the sets $\{S_k\}_{k=1}^m$ are disjoint. Also, we have $s_{a_i}h_{\boldsymbol{\Theta}^{(t)}}(\boldsymbol{x}_i) = \beta^{(t)}_{a_i,i}$ from conditions (a) and (e). Together with condition (a), the assumptions of Lemma 9 are satisfied for all $i \in [n]$. Consequently, $\beta^{(t+1)}_{a_i,i} > 0$ for all $i \in [n]$.

Part (b): According to the dual gradient update in Equation (43b), and using condition (e) at iteration t, we have:

$$\alpha^{(t+1)}_{k,j} = \alpha^{(t)}_{k,j} \quad \text{for all } j \in [n] \text{ with } k \neq a_j.$$

Therefore, condition (b) continues to hold at iteration $t+1$.

Part (c): By conditions (a) and (e), the gradient update at iteration t for $\boldsymbol{\beta}_k^{(t)}$ depends only on the examples in the subset S_k. Hence, the gradient update for an individual neuron is equivalent to a linear regression gradient descent. As similarly argued in the proof of Lemma 2, since the step size satisfies $\eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$, the linear regression squared loss is monotonically nonincreasing, and by condition (c) at iteration t, we obtain

$$\left\|\boldsymbol{\beta}_{k,S_k}^{(t+1)} - s_k \boldsymbol{y}_{S_k}\right\|_2 \leq \left\|\boldsymbol{\beta}_{k,S_k}^{(t)} - s_k \boldsymbol{y}_{S_k}\right\|_2 \leq C_y \left\|\boldsymbol{y}_{S_k}\right\|_2 .$$

Therefore, condition (c) holds at iteration $t+1$.

Part (d): Following the same argument as in Part (d) of Lemma 14, using conditions (b) and (c) at iteration $t+1$, together with the eigenvalue bounds from Lemma 11, we can establish that

$$\left\|\boldsymbol{\alpha}_k^{(t+1)}\right\|_2 \leq \frac{C_\alpha \sqrt{n} y_{\max}}{\|\boldsymbol{\lambda}\|_1} \text{ for all } k \in [m],$$

with probability at least $1 - 2e^{-n/C_g}$. Thus, condition (d) holds at iteration $t+1$.

Part (e): By Lemma 10, since conditions (b) and (d) hold at iteration $t+1$, we conclude that $\beta_{k,j}^{(t+1)} \leq 0$ for all $j \in [n]$ with $k \neq a_j$. Thus, condition (e) holds at iteration $t+1$.

■

Equipped with Lemma 19, we are ready to prove Theorem 17.

Proof (Theorem 17) The proof follows a similar structure to that of Theorem 4 for single ReLU models, but now we must track the dynamics of all the neurons $\{\boldsymbol{w}_k\}_{k=1}^m$ simultaneously. Equipped with sufficient conditions under which the activation patterns are preserved in Lemma 19, we verify these conditions hold after the first gradient step, and use induction to characterize the full gradient descent dynamics.

We first verify that the iterate at $t=1$ satisfies all the sufficient conditions. With the initialization

$$\boldsymbol{w}_k^{(0)} = \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} \left(\frac{1}{C_g}\boldsymbol{A}_k \boldsymbol{y} + \boldsymbol{\epsilon}_k\right),$$

we have $\boldsymbol{\beta}_k^{(0)} = \frac{1}{C_g}\boldsymbol{A}_k\boldsymbol{y} + \boldsymbol{\epsilon}_k$. Recalling the definition of $\boldsymbol{A}_k$ in Assumption 3, we have

$$\boldsymbol{\beta}_{k,i}^{(0)} = \begin{cases} \epsilon_{a_i,i}, & \text{if } a_i = k, \text{ or } s_k \cdot y_i < 0 \\ -\frac{|y_i|}{C_g} + \epsilon_{k,i}, & \text{otherwise} \end{cases}, \tag{46}$$

for all $k \in [m]$ and $i \in [n]$. Since the theorem assumption ensures $\epsilon_{k,i} \leq \frac{1}{C_\alpha m} y_{\min}$ and $C_\alpha \gtrsim C_g^2$, we have $-\frac{|y_i|}{C_g} + \epsilon_{k,i} < 0$. Therefore, we obtain

$$h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{x}_i) = \sum_{k=1}^m s_k \sigma\left(\boldsymbol{\beta}_{k,i}^{(0)}\right) = s_{a_i}\epsilon_{a_i,i} - s_{a_i} \sum_{k: s_k \cdot y_i < 0} \epsilon_{k,i}, \tag{47}$$

for all $i \in [n]$. Therefore, using the primal gradient update in Equation (43a), we obtain

$$\begin{aligned}\boldsymbol{\beta}_k^{(1)} &= \boldsymbol{\beta}_k^{(0)} - \eta s_k \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\beta}_k^{(0)})(h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) - \boldsymbol{y}) \\ &= \boldsymbol{X}\boldsymbol{X}^\top \left[\underbrace{\eta \left(s_k \boldsymbol{D}(\boldsymbol{\beta}_k^{(0)}) \left(\boldsymbol{y} - h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) \right) + \frac{1}{\eta} \left(\boldsymbol{X}\boldsymbol{X}^\top \right)^{-1} \boldsymbol{\beta}_k^{(0)} \right)}_{=:\boldsymbol{\alpha}_k^{(1)}} \right], \end{aligned} \tag{48}$$

according to the primal-dual formulation $\boldsymbol{\beta}_k^{(1)} = \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{\alpha}_k^{(1)}$ in Equation (3). In the below, we show that at iteration $t = 1$, the variables $\boldsymbol{\beta}_k^{(1)}$ and $\boldsymbol{\alpha}_k^{(1)}$ satisfy all the conditions in Lemma 19.

Part (a): For all $i \in [n]$, we show that $\beta_{a_i,i}^{(1)} > 0$ by applying Lemma 9. According to Equation (46) and Equation (47), we have $\beta_{a_i,i}^{(0)} = \epsilon_{a_i,i} > 0$ and $s_{a_i} \cdot h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{x}_i) = \epsilon_{a_i,i} - \sum_{k:s_k \cdot y_i < 0} \epsilon_{k,i} \leq \beta_{a_i,i}^{(0)}$. Moreover, we have

$$\left\| h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) - \boldsymbol{y} \right\|_2 \leq \left\| h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) \right\|_2 + \|\boldsymbol{y}\|_2 \leq \sum_{k=1}^m \|\boldsymbol{\epsilon}_k\|_2 + \|\boldsymbol{y}\|_2 \leq \frac{\sqrt{n}}{C_\alpha} y_{\min} + \|\boldsymbol{y}\|_2 \leq C_y \|\boldsymbol{y}\|_2 ,$$

with $C_y \geq 1 + \frac{1}{C_\alpha}$. All the conditions of Lemma 9 are satisfied, and therefore, $\beta_{a_i,i}^{(1)} > 0$ for all $i \in [n]$.

Part (b): For all $j \in [n]$ with $k \neq a_j$, we verify that $\alpha_{k,j}^{(1)}$ satisfies the required upper and lower bounds. We need to discuss two cases: 1) $\beta_{k,j}^{(0)} = \epsilon_{k,j} > 0$ with $s_k \cdot y_j < 0$, and 2) $\beta_{k,j}^{(0)} = -\frac{|y_j|}{C_g} + \epsilon_{k,j} < 0$ with $s_k \cdot y_j > 0$.

Case 1): For $\beta_{k,j}^{(0)} = \epsilon_{k,j} > 0$ with $s_k \cdot y_j < 0$, we work from Equation (48) to get

$$\begin{aligned}\alpha_{k,j}^{(1)} &= \eta \boldsymbol{e}_j^\top \left(s_k \boldsymbol{D}(\boldsymbol{\beta}_k^{(0)}) \left(\boldsymbol{y} - h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) \right) + \frac{1}{\eta} \left(\boldsymbol{X}\boldsymbol{X}^\top \right)^{-1} \boldsymbol{\beta}_k^{(0)} \right) \\ &\overset{\text{(i)}}{=} \eta \left(-|y_j| - s_k \left(s_{a_j} \epsilon_{a_j,j} - s_{a_j} \sum_{k:s_k \cdot y_j < 0} \epsilon_{k,j} \right) + \frac{1}{\eta} \boldsymbol{e}_j^\top \left(\boldsymbol{X}\boldsymbol{X}^\top \right)^{-1} \left(\frac{1}{C_g} \boldsymbol{A}_k \boldsymbol{y} + \boldsymbol{\epsilon}_k \right) \right) \\ &= \eta \left(-|y_j| + \epsilon_{a_j,j} - \sum_{k:s_k \cdot y_j < 0} \epsilon_{k,j} + \frac{1}{\eta} \boldsymbol{e}_j^\top \left(\boldsymbol{X}\boldsymbol{X}^\top \right)^{-1} \left(\frac{1}{C_g} \boldsymbol{A}_k \boldsymbol{y} + \boldsymbol{\epsilon}_k \right) \right) \\ &= \eta \left(-|y_j| + \epsilon_{a_j,j} - \sum_{k:s_k \cdot y_j < 0} \epsilon_{k,j} + \frac{1}{\eta} \boldsymbol{e}_j^\top \left[\frac{1}{\|\boldsymbol{\lambda}\|_1} \boldsymbol{I} + \left(\left(\boldsymbol{X}\boldsymbol{X}^\top \right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1} \boldsymbol{I} \right) \right] \left(\frac{1}{C_g} \boldsymbol{A}_k \boldsymbol{y} + \boldsymbol{\epsilon}_k \right) \right) \\ &\overset{\text{(ii)}}{=} \eta \left(-|y_j| + \epsilon_{a_j,j} - \sum_{k:s_k \cdot y_j < 0} \epsilon_{k,j} + \frac{\epsilon_{k,j}}{\eta \|\boldsymbol{\lambda}\|_1} + \frac{1}{\eta} \boldsymbol{e}_j^\top \left(\left(\boldsymbol{X}\boldsymbol{X}^\top \right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1} \boldsymbol{I} \right) \left(\frac{1}{C_g} \boldsymbol{A}_k \boldsymbol{y} + \boldsymbol{\epsilon}_k \right) \right) \end{aligned} \tag{49}$$

where equality (i) substitutes $h_{\mathbf{\Theta}^{(0)}}(\boldsymbol{x}_j) = s_{a_j}\epsilon_{a_j,j} - s_{a_j}\sum_{k:s_k\cdot y_j<0}\epsilon_{k,j}$ from Equation (46) and $\boldsymbol{\beta}_k^{(0)} = \frac{1}{C_g}\boldsymbol{A}_k\boldsymbol{y} + \boldsymbol{\epsilon}_k$, and equality (ii) applies $(\boldsymbol{A}_k)_{jj} = 0$ for $k \neq a_j$ with $s_k \cdot y_j < 0$. For the upper bound, we have

$$\alpha_{k,j}^{(1)} \leq \eta\left(-|y_j| + \epsilon_{a_j,j} + \frac{\epsilon_{k,j}}{\eta\|\boldsymbol{\lambda}\|_1} + \frac{1}{\eta}\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I}\right\|_2\left\|\frac{1}{C_g}\boldsymbol{A}_k\boldsymbol{y} + \boldsymbol{\epsilon}_k\right\|_2\right), \tag{50}$$

by dropping negative terms $-\sum_{k:s_k\cdot y_j<0}\epsilon_{k,j}$. Next, by applying the upper bound in Corollary 13 and the upper bounds for $\|\boldsymbol{y}\|_2$ and $\|\boldsymbol{\epsilon}_k\|_2$, we have

$$\begin{aligned}\alpha_{k,j}^{(1)} &\leq \eta\left(-|y_j| + \epsilon_{a_j,j} + \frac{\epsilon_{k,j}}{\eta\|\boldsymbol{\lambda}\|_1} + \frac{C_g}{\eta\|\boldsymbol{\lambda}\|_1}C\cdot\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\left(\frac{\sqrt{n}y_{\max}}{C_g} + \frac{\sqrt{n}}{C_\alpha m}y_{\min}\right)\right)\\ &\leq \eta\left(-|y_j| + \epsilon_{a_j,j} + \frac{\epsilon_{k,j}}{\eta\|\boldsymbol{\lambda}\|_1} + \frac{C_g}{\eta\|\boldsymbol{\lambda}\|_1}C\cdot\frac{y_{\min}}{C_0 y_{\max}}\left(\frac{y_{\max}}{C_g} + \frac{y_{\min}}{C_\alpha m}\right)\right),\end{aligned}$$

where the second inequality substitutes $d_2 \geq C_0^2\frac{n^2y_{\max}^2}{y_{\min}^2}$ and $d_\infty \geq C_0\frac{n^{1.5}y_{\max}}{y_{\min}}$ in Assumption 2. Finally, by using the step size assumption $\frac{1}{\eta} \leq CC_g\|\boldsymbol{\lambda}\|_1$ and the theorem assumption $\epsilon_{k,j} \leq \frac{1}{C_\alpha m}y_{\min}$, we have

$$\begin{aligned}\alpha_{k,j}^{(1)} &\leq \frac{1}{CC_g\|\boldsymbol{\lambda}\|_1}\left(-y_{\min} + \frac{y_{\min}}{C_\alpha m} + \frac{CC_g y_{\min}}{C_\alpha m} + \frac{2C^2C_g^2}{C_0}y_{\min}\right)\\ &\leq -\frac{y_{\min}}{C_\alpha\|\boldsymbol{\lambda}\|_1},\end{aligned}$$

with $C_0 \gtrsim C_\alpha^2$ and $C_\alpha \gtrsim \max\{C_g^2, C_yC_g\}$. For the lower bound, starting from Equation (49), we have

$$\begin{aligned}\alpha_{k,j}^{(1)} &\geq \eta\left(-|y_j| - \sum_{k:s_k\cdot y_j<0}\epsilon_{k,j} - \frac{1}{\eta}\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I}\right\|_2\left\|\frac{1}{C_g}\boldsymbol{A}_k\boldsymbol{y} + \boldsymbol{\epsilon}_k\right\|_2\right)\\ &\overset{\text{(i)}}{\geq} \eta\left(-|y_j| - \sum_{k:s_k\cdot y_j<0}\epsilon_{k,j} - \frac{C_g}{\eta\|\boldsymbol{\lambda}\|_1}C\cdot\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\left(\frac{\sqrt{n}y_{\max}}{C_g} + \frac{\sqrt{n}}{C_\alpha m}y_{\min}\right)\right)\\ &\overset{\text{(ii)}}{\geq} \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}\left(-y_{\max} - \frac{y_{\min}}{C_\alpha} - C^2C_g^2\cdot\frac{y_{\min}}{C_0y_{\max}}\left(\frac{y_{\max}}{C_g} + \frac{y_{\min}}{C_\alpha m}\right)\right)\\ &\overset{\text{(iii)}}{\geq} -\frac{3y_{\max}}{C_g\|\boldsymbol{\lambda}\|_1},\end{aligned}$$

where inequality (i) applies the upper bound in Corollary 13 and the upper bounds for $\|\boldsymbol{y}\|_2$ and $\|\boldsymbol{\epsilon}_k\|_2$, inequalities (ii) applies $d_2 \geq C_0^2\frac{n^2y_{\max}^2}{y_{\min}^2}$ and $d_\infty \geq C_0\frac{n^{1.5}y_{\max}}{y_{\min}}$ in Assumption 2, and inequality (iii) follows by the constant relationship that $C_0 \gtrsim C_\alpha^2$ and $C_\alpha \gtrsim \max\{C_g^2, C_yC_g\}$. Thus, for $\beta_{k,j}^{(0)} = \epsilon_{k,j} > 0$ with $s_k \cdot y_j < 0$, $\alpha_{k,j}^{(1)}$ satisfies both the required upper and lower bounds.

Case 2): For $\beta_{k,j}^{(0)} = -\frac{|y_j|}{C_g} + \epsilon_{k,j} < 0$ with $s_k \cdot y_j > 0$, we work from Equation (48) to get

$$\begin{aligned}\alpha_{k,j}^{(1)} &= \eta \boldsymbol{e}_j^\top \left(s_k \boldsymbol{D}(\boldsymbol{\beta}_k^{(0)}) \left(\boldsymbol{y} - h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X})\right) + \frac{1}{\eta}\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} \boldsymbol{\beta}_k^{(0)} \right) \\ &= \boldsymbol{e}_j^\top \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} \left(\frac{1}{C_g}\boldsymbol{A}_k \boldsymbol{y} + \boldsymbol{\epsilon}_k \right),\end{aligned}$$

where we substitute $\boldsymbol{\beta}_k^{(0)} = \frac{1}{C_g}\boldsymbol{A}_k\boldsymbol{y} + \boldsymbol{\epsilon}_k$ and $\beta_{k,j}^{(0)} = -\frac{1}{C_g}|y_j| + \epsilon_{k,j} < 0$, and this eliminates the first term, since $D_{jj} = 0$. Then, $\alpha_{k,j}^{(1)}$ can further be written as

$$\begin{aligned}\alpha_{k,j}^{(1)} &= \boldsymbol{e}_j^\top \left[\frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I} + \left(\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I} \right)\right] \left(\frac{1}{C_g}\boldsymbol{A}_k \boldsymbol{y} + \boldsymbol{\epsilon}_k \right) \\ &= \frac{1}{\|\boldsymbol{\lambda}\|_1}\left(-\frac{|y_j|}{C_g} + \epsilon_{k,j} \right) + \boldsymbol{e}_j^\top \left(\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I} \right) \left(\frac{1}{C_g}\boldsymbol{A}_k \boldsymbol{y} + \boldsymbol{\epsilon}_k \right).\end{aligned} \tag{51}$$

For the upper bound, we have

$$\begin{aligned}\alpha_{k,j}^{(1)} &\le \frac{1}{\|\boldsymbol{\lambda}\|_1}\left(-\frac{1}{C_g}|y_j| + \epsilon_{k,j} \right) + \left\| \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I} \right\|_2 \left\| \frac{1}{C_g}\boldsymbol{A}_k \boldsymbol{y} + \boldsymbol{\epsilon}_k \right\|_2 \\ &\overset{\text{(i)}}{\le} \frac{1}{\|\boldsymbol{\lambda}\|_1}\left(-\frac{|y_j|}{C_g} + \epsilon_{k,j} + C_g C \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty} \right) \left(\frac{\sqrt{n} y_{\max}}{C_g} + \frac{\sqrt{n}}{C_\alpha} y_{\min} \right)\right) \\ &\overset{\text{(ii)}}{\le} \frac{1}{\|\boldsymbol{\lambda}\|_1}\left(-\frac{y_{\min}}{C_g} + \frac{y_{\min}}{C_\alpha m} + C_g C \cdot \frac{y_{\min}}{C_0 y_{\max}} \left(\frac{y_{\max}}{C_g} + \frac{1}{C_\alpha} y_{\min} \right)\right) \\ &\le -\frac{y_{\min}}{C_\alpha \|\boldsymbol{\lambda}\|_1},\end{aligned}$$

where inequality (i) applies the upper bound in Corollary 13 and the upper bounds for $\|\boldsymbol{y}\|_2$ and $\|\boldsymbol{\epsilon}_k\|_2$, inequalities (ii) substitutes $d_2 \ge C_0^2 \frac{n^2 y_{\max}^2}{y_{\min}^2}$ and $d_\infty \ge C_0 \frac{n^{1.5} y_{\max}}{y_{\min}}$ in Assumption 2. The last inequality follows by $C_0 \gtrsim C_\alpha^2$ and $C_\alpha \gtrsim \max\{C_g^2, C_y C_g\}$. For the lower bound, we work from Equation (51) to get

$$\begin{aligned}\alpha_{k,j}^{(1)} &\ge \frac{1}{\|\boldsymbol{\lambda}\|_1}\left(-\frac{|y_j|}{C_g} \right) - \left\| \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} - \frac{1}{\|\boldsymbol{\lambda}\|_1}\boldsymbol{I} \right\|_2 \left\| \frac{1}{C_g}\boldsymbol{A}_k \boldsymbol{y} + \boldsymbol{\epsilon}_k \right\|_2 \\ &\ge \frac{1}{\|\boldsymbol{\lambda}\|_1}\left(-\frac{|y_j|}{C_g} - C_g C \cdot \max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty} \right) \left(\frac{\sqrt{n} y_{\max}}{C_g} + \frac{\sqrt{n}}{C_\alpha} y_{\min} \right)\right) \\ &\ge \frac{1}{\|\boldsymbol{\lambda}\|_1}\left(-\frac{y_{\max}}{C_g} - C_g C \cdot \frac{y_{\min}}{C_0 y_{\max}} \left(\frac{y_{\max}}{C_g} + \frac{y_{\min}}{C_\alpha} \right)\right) \\ &\ge -\frac{3 y_{\max}}{C_g \|\boldsymbol{\lambda}\|_1},\end{aligned}$$

by the same argument. Thus, for $\beta_{k,j}^{(0)} = -\frac{|y_j|}{C_g} + \epsilon_{k,j} < 0$ with $s_k \cdot y_j > 0$, $\alpha_{k,j}^{(1)}$ satisfies both the required upper and lower bounds. This completes the proof of this part.

Part (c): We verify that the primal variables $\boldsymbol{\beta}_{k,S_k}^{(1)}$ corresponding to active examples minus $\boldsymbol{y}_{S_k}$ satisfy the norm bound. Specifically, we show that $\left\|\boldsymbol{\beta}_{k,S_k}^{(1)} - s_k \boldsymbol{y}_{S_k}\right\|_2^2 \leq C_y^2 \|\boldsymbol{y}\|_2^2$. According to Equation (48), we have

$$
\begin{aligned}
\left\|\boldsymbol{\beta}_{k,S_k}^{(1)} - s_k \boldsymbol{y}_{S_k}\right\|_2^2 &= \sum_{i:a_i=k} \left(\beta_{k,i}^{(1)} - s_k y_i\right)^2 \\
&= \sum_{i:a_i=k} \left(\beta_{k,i}^{(0)} - \eta s_k \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\beta}_k^{(0)})(h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) - \boldsymbol{y}) - s_k y_i\right)^2 \\
&= \sum_{i:a_i=k} \left(\underbrace{\epsilon_{a_i,i} - \eta s_{a_i} \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\beta}_{a_i}^{(0)})(h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) - \boldsymbol{y}) - |y_i|}_{=:T_i}\right)^2,
\end{aligned}
\tag{52}
$$

where we substitute $\beta_{k,i}^{(0)} = \beta_{a_i,i}^{(0)} = \epsilon_{a_i,i}$ for $k = a_i$, and $s_{a_i} \cdot y_i > 0$. Next, we bound the term
$T_i := \epsilon_{a_i,i} - \eta s_{a_i} \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\beta}_{a_i}^{(0)})(h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) - \boldsymbol{y}) - |y_i|$ for all $i \in [n]$. We have

$$
\begin{aligned}
T_i &= \epsilon_{a_i,i} - \eta s_{a_i} \boldsymbol{e}_i^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\beta}_{a_i}^{(0)})(h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) - \boldsymbol{y}) - |y_i| \\
&= (\epsilon_{a_i,i} - |y_i|) - \eta s_{a_i} \boldsymbol{e}_i^\top \left[\|\boldsymbol{\lambda}\|_1 \boldsymbol{I} + \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right)\right] \boldsymbol{D}(\boldsymbol{\beta}_{a_i}^{(0)}) \left(h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) - \boldsymbol{y}\right) \\
&= (\epsilon_{a_i,i} - |y_i|) - \eta s_{a_i} \|\boldsymbol{\lambda}\|_1 \left(s_{a_i}\epsilon_{a_i,i} - s_{a_i} \sum_{k:s_k \cdot y_i < 0} \epsilon_{k,i} - y_i\right) \\
&\qquad - \eta s_k \boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right) \boldsymbol{D}(\boldsymbol{\beta}_k^{(0)}) \left(h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) - \boldsymbol{y}\right) \\
&= (1 - \eta \|\boldsymbol{\lambda}\|_1)\epsilon_{a_i,i} - (1 - \eta \|\boldsymbol{\lambda}\|_1)|y_i| + \eta \|\boldsymbol{\lambda}\|_1 \sum_{k:s_k \cdot y_i < 0} \epsilon_{k,i} \\
&\qquad - \eta s_k \boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right) \boldsymbol{D}(\boldsymbol{\beta}_k^{(0)}) \left(h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) - \boldsymbol{y}\right),
\end{aligned}
$$

by applying $h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{x}_i) = s_{a_i}\epsilon_{a_i,i} - s_{a_i} \sum_{k:s_k \cdot y_i < 0} \epsilon_{k,i}$ from Equation (47). Since the step size assumption guarantees that $\frac{1}{CC_g\|\boldsymbol{\lambda}\|_1} \leq \eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$, and $\epsilon_{k,i} \leq \frac{1}{C_\alpha m} y_{\min}$, we have

$$
\begin{aligned}
&(1 - \eta \|\boldsymbol{\lambda}\|_1)\epsilon_{a_i,i} - (1 - \eta \|\boldsymbol{\lambda}\|_1)|y_i| + \eta \|\boldsymbol{\lambda}\|_1 \sum_{k:s_k \cdot y_i < 0} \epsilon_{k,i} \\
&\leq \epsilon_{a_i,i} - (1 - \eta \|\boldsymbol{\lambda}\|_1)|y_i| + \eta \|\boldsymbol{\lambda}\|_1 \sum_{k:s_k \cdot y_i < 0} \epsilon_{k,i} \\
&\leq \left(\frac{1}{m} + \frac{1}{C_g}\right) \frac{1}{C_\alpha} y_{\min} - \left(1 - \frac{1}{C_g}\right) y_{\min} \\
&< 0,
\end{aligned}
$$

with $C_\alpha \gtrsim C_g^2$. Hence, in order to upper bound T_i^2, it suffices to find the lower bound for T_i. We have

$$\begin{aligned} T_i &= (1-\eta\|\boldsymbol{\lambda}\|_1)\epsilon_{a_i,i} - (1-\eta\|\boldsymbol{\lambda}\|_1)|y_i| + \eta\|\boldsymbol{\lambda}\|_1 \sum_{k:s_k\cdot y_i<0}\epsilon_{k,i} \\ &\qquad - \eta s_k \boldsymbol{e}_i^\top \left(\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1 \boldsymbol{I}\right)\boldsymbol{D}(\boldsymbol{\beta}_k^{(0)})\left(h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) - \boldsymbol{y}\right) \\ &\geq -|y_i| - \eta\left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1\boldsymbol{I}\right\|_2 \left\|h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) - \boldsymbol{y}\right\|_2, \end{aligned}$$

where the inequality drops the positive terms $(1-\eta\|\boldsymbol{\lambda}\|_1)\epsilon_{a_i,i}$, $\eta\|\boldsymbol{\lambda}\|_1|y_i|$, and $\eta\|\boldsymbol{\lambda}\|_1\sum_{k:s_k\cdot y_i<0}\epsilon_{k,i}$. We again upper bound $\left\|\boldsymbol{X}\boldsymbol{X}^\top - \|\boldsymbol{\lambda}\|_1\boldsymbol{I}\right\|_2$ by Corollary 13. With probability at least $1-2\exp(-n(Cc-\ln 9))$, we have

$$\begin{aligned} T_i &\geq -|y_i| - \eta\cdot C\|\boldsymbol{\lambda}\|_1\cdot\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\left\|h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) - \boldsymbol{y}\right\|_2 \\ &\geq -|y_i| - \frac{C}{C_g}\cdot\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\left\|h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X}) - \boldsymbol{y}\right\|_2, \end{aligned}$$

by applying $\eta \leq \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$. Finally, we apply the upper bound that $\left\|h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X})\right\|_2 \leq \sum_{k=1}^m \|\boldsymbol{\epsilon}_k\|_2 \leq \frac{\sqrt{n}}{C_\alpha}y_{\min}$ and $\|\boldsymbol{y}\|_2 \leq \sqrt{n}y_{\max}$, and Assumption 2 ensures that $d_2 \geq C_0^2\frac{n^2y_{\max}^2}{y_{\min}^2}$ and $d_\infty \geq C_0\frac{n^{1.5}y_{\max}}{y_{\min}}$. We have

$$\begin{aligned} T_i &\geq -|y_i| - \frac{C}{C_g}\max\left(\sqrt{\frac{n}{d_2}}, \frac{n}{d_\infty}\right)\left(\frac{\sqrt{n}}{C_\alpha}y_{\min} + \sqrt{n}y_{\max}\right) \\ &\geq -|y_i| - \frac{Cy_{\min}}{C_gC_0y_{\max}}\left(\frac{1}{C_\alpha}y_{\min} + y_{\max}\right) \\ &\geq -|y_i|\left(1 + \frac{2C}{C_gC_0}\right) \\ &\geq -C_y|y_i|, \end{aligned}$$

with the choice of $C_y \geq 2$. Substituting $T_i^2 \leq C_y^2y_i^2$ into Equation (52), we have

$$\left\|\boldsymbol{\beta}_{k,S_k}^{(1)} - s_k\boldsymbol{y}_{S_k}\right\|_2^2 \leq \sum_{i:a_i=k} C_y^2y_i^2 = C_y^2\left\|\boldsymbol{y}_{S_k}\right\|_2^2.$$

As a result, we conclude that $\left\|\boldsymbol{\beta}_{k,S_k}^{(1)} - s_k\boldsymbol{y}_{S_k}\right\|_2 \leq C_y\left\|\boldsymbol{y}_{S_k}\right\|_2$ as required.

Part (d): We verify the norm bounds on the dual variables. By the triangle inequality, we work from Equation (48) to get

$$\begin{aligned} \left\|\boldsymbol{\alpha}_k^{(1)}\right\|_2 &= \left\|\eta\left(s_k\boldsymbol{D}(\boldsymbol{\beta}_k^{(0)})\left(\boldsymbol{y} - h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X})\right) + \frac{1}{\eta}\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{\beta}_k^{(0)}\right)\right\|_2 \\ &\leq \eta\left[\|\boldsymbol{y}\|_2 + \left\|h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X})\right\|_2 + \frac{1}{\eta}\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\right\|_2\left\|\boldsymbol{\beta}_k^{(0)}\right\|_2\right] \\ &= \eta\left[\|\boldsymbol{y}\|_2 + \left\|h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X})\right\|_2 + \frac{1}{\eta}\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\right\|_2\left\|\frac{1}{C_g}\boldsymbol{A}_k\boldsymbol{y} + \boldsymbol{\epsilon}_k\right\|_2\right], \end{aligned}$$

by substituting $\boldsymbol{\beta}_k^{(0)} = \frac{1}{C_g}\boldsymbol{A}_k\boldsymbol{y} + \boldsymbol{\epsilon}_k$. Using $\|\boldsymbol{y}\|_2 \le \sqrt{n}y_{\max}$, $\left\|h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X})\right\|_2 \le \sum_{k=1}^m \|\boldsymbol{\epsilon}_k\|_2 \le \frac{\sqrt{n}}{C_\alpha}y_{\min}$, $\left\|\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\right\|_2 \le \frac{C_g}{\|\boldsymbol{\lambda}\|_1}$, $\epsilon_{k,i} \le \frac{1}{C_\alpha m}y_{\min}$, and $\frac{1}{CC_g\|\boldsymbol{\lambda}\|_1} \le \eta \le \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}$, we have

$$\begin{aligned}\left\|\boldsymbol{\alpha}_k^{(1)}\right\|_2 &\le \frac{1}{C_g\|\boldsymbol{\lambda}\|_1}\left[\sqrt{n}y_{\max} + \frac{\sqrt{n}}{C_\alpha}y_{\min} + CC_g\|\boldsymbol{\lambda}\|_1 \cdot \frac{C_g}{\|\boldsymbol{\lambda}\|_1}\cdot\left(\frac{\sqrt{n}y_{\max}}{C_g} + \frac{\sqrt{n}}{C_\alpha m}y_{\min}\right)\right] \\ &\le \frac{1}{\|\boldsymbol{\lambda}\|_1}\left(3\sqrt{n}y_{\max}\right) \\ &\le \frac{C_\alpha\sqrt{n}y_{\max}}{\|\boldsymbol{\lambda}\|_1},\end{aligned}$$

with $C_\alpha \gtrsim \max\{C_g^2, C_yC_g\}$. Thus, condition (d) holds at $t = 1$.

Part (e): Since we have shown that $\alpha_{k,j}^{(1)} \le -\frac{y_{\min}}{C_\alpha\|\boldsymbol{\lambda}\|_1}$ and $\left\|\boldsymbol{\alpha}_k^{(1)}\right\|_2 \le \frac{C_\alpha\sqrt{n}y_{\max}}{\|\boldsymbol{\lambda}\|_1}$ for all $j \in [n]$ with $k \ne a_j$, by Lemma 10, it follows that $\beta_{k,j}^{(1)} \le 0$ for all $j \in [n]$ with $k \ne a_j$.

We have shown that at iteration $t = 1$ the conditions in Lemma 19 are satisfied, and by induction, these conditions will also hold for $t \ge 1$. As a result, $\boldsymbol{w}_k$ is trained with only predefined active examples starting from the iteration $t = 0$, and it is equivalent to linear regression using only active examples with initialization $\boldsymbol{w}_k^{(1)} = \eta\boldsymbol{X}^\top\left(s_k\boldsymbol{D}(\boldsymbol{\beta}_k^{(0)})\left(\boldsymbol{y} - h_{\boldsymbol{\Theta}^{(0)}}(\boldsymbol{X})\right) + \frac{1}{\eta}\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{\beta}_k^{(0)}\right)$. Finally, since $\boldsymbol{w}_k$ is trained on disjoint subset of examples by Assumption 3, by Lemma 2, $\boldsymbol{w}_k^{(\infty)}$ satisfies

$$\boldsymbol{w}_k^{(\infty)} = \underset{\boldsymbol{w}\in\{\boldsymbol{w}:\boldsymbol{X}_{S_k}\boldsymbol{w}=\boldsymbol{y}_{S_k}\}}{\arg\min}\left\|\boldsymbol{w} - \boldsymbol{w}_k^{(1)}\right\|_2.$$

This completes the proof of Theorem 17. ■

G.3. Proof of Theorem 18 (Implicit Bias Approximation to $w^\star$)

Proof (Theorem 18) We restate the definition of $\boldsymbol{w}^\star$ in Equation (44).

$$\begin{aligned}\{\boldsymbol{w}_k^\star\}_{k=1}^m = \underset{\{\boldsymbol{w}_k\}_{k=1}^m}{\arg\min}\ \frac{1}{2}\sum_{k=1}^m \|\boldsymbol{w}_k\|_2^2 \\ \text{s.t.} \sum_{k=1}^m s_k\sigma(\boldsymbol{w}_k^\top\boldsymbol{x}_i) = y_i, \text{ for all } i \in [n].\end{aligned} \tag{53}$$

Recall that the gradient descent limit $\{\boldsymbol{w}_k^{(\infty)}\}_{k=1}^m$ satisfies the same set of constraints: it interpolates all examples. Consequently, both $\{\boldsymbol{w}_k^{(\infty)}\}_{k=1}^m$ and $\{\boldsymbol{w}_k^\star\}_{k=1}^m$ are feasible solutions to (53). We show that the norm difference between $\boldsymbol{w}_k^{(\infty)}$ and $\boldsymbol{w}_k^\star$ can be upper bounded by 2 times the norm of $\boldsymbol{w}_k^{(\infty)}$.

$$\sum_{k=1}^m\left\|\boldsymbol{w}_k^{(\infty)} - \boldsymbol{w}_k^\star\right\|_2^2 \le 2\sum_{k=1}^m\left\|\boldsymbol{w}_k^{(\infty)}\right\|_2^2 + 2\sum_{k=1}^m\|\boldsymbol{w}_k^\star\|_2^2 \le 4\sum_{k=1}^m\left\|\boldsymbol{w}_k^{(\infty)}\right\|_2^2,$$

where it follows the definition of (53). By Lemma 19, we have the upper bound for $\left\|\boldsymbol{w}_k^{(\infty)}\right\|_2^2$ as

$$\left\|\boldsymbol{w}_k^{(\infty)}\right\|_2^2 = \boldsymbol{\alpha}_k^{(\infty)\top}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{\alpha}_k^{(\infty)} \le \mu_1(\boldsymbol{X}\boldsymbol{X}^\top)\left\|\boldsymbol{\alpha}_k^{(\infty)}\right\|_2^2 \le C_g\left\|\boldsymbol{\lambda}\right\|_1 \cdot \frac{C_\alpha^2 n y_{\max}^2}{\left\|\boldsymbol{\lambda}\right\|_1^2} = \frac{C_g C_\alpha^2 n y_{\max}^2}{\left\|\boldsymbol{\lambda}\right\|_1}.$$

As a result, we have

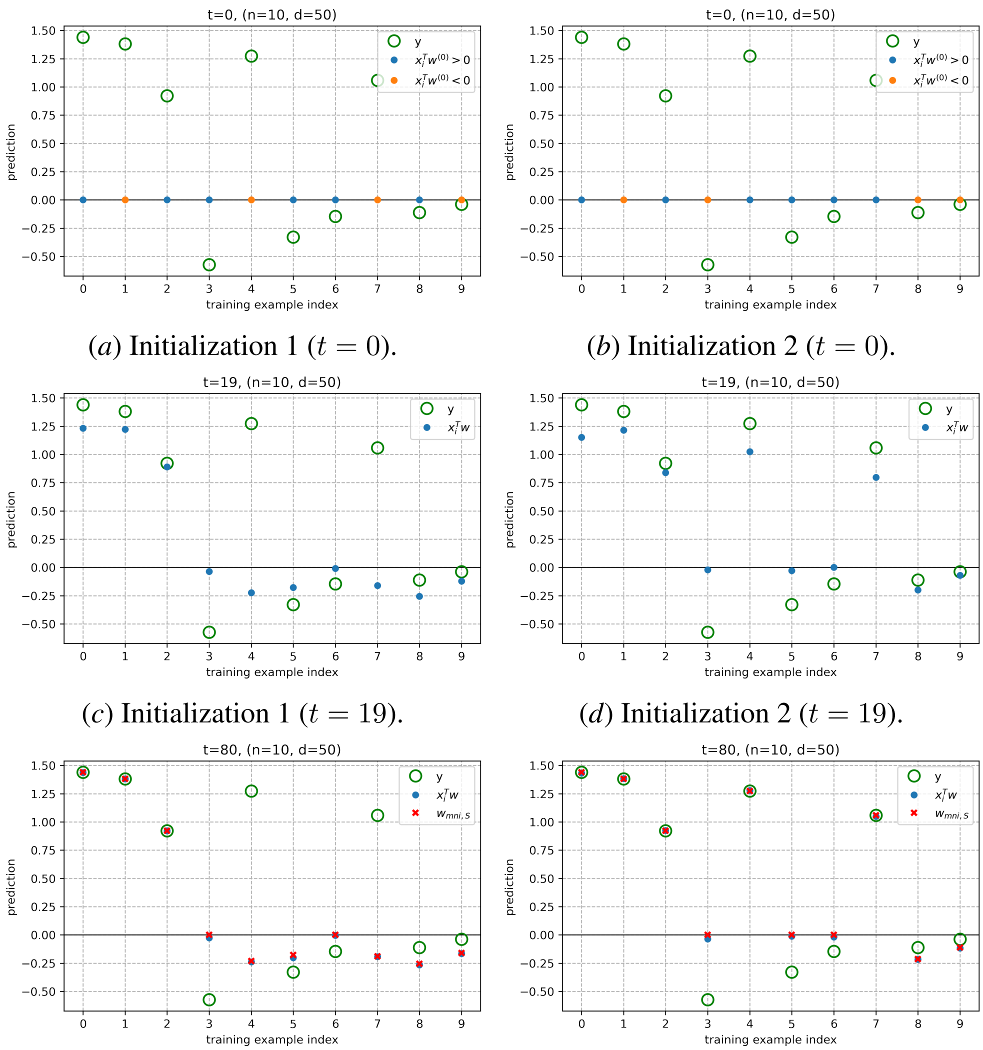
$$\sum_{k=1}^m \left\|\boldsymbol{w}_k^{(\infty)} - \boldsymbol{w}_k^\star\right\|_2^2 \le \frac{4C_g C_\alpha^2 m n y_{\max}^2}{\left\|\boldsymbol{\lambda}\right\|_1}.$$

■

Appendix H. Simulations

We present visualizations of an exploratory nature, of the evolution of the primal variables at iteration checkpoints in settings that violate the assumptions made in our theoretical results.

H.1. Moderate-Dimensional Data and Single ReLU Model



(*a*) Initialization 1 ($t = 0$).

(*b*) Initialization 2 ($t = 0$).

(*c*) Initialization 1 ($t = 19$).

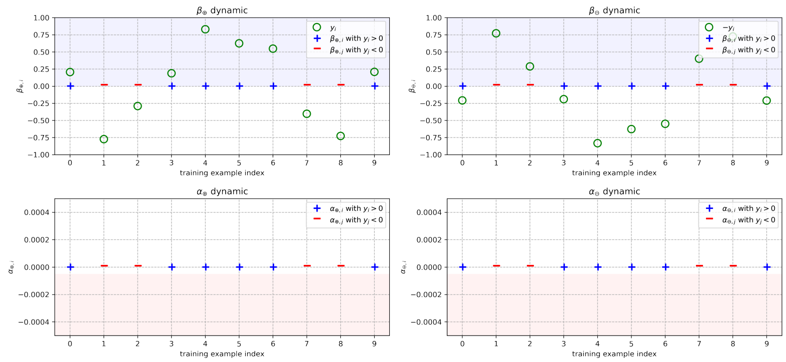
(*d*) Initialization 2 ($t = 19$).

(*e*) Initialization 1 ($t = 80$).

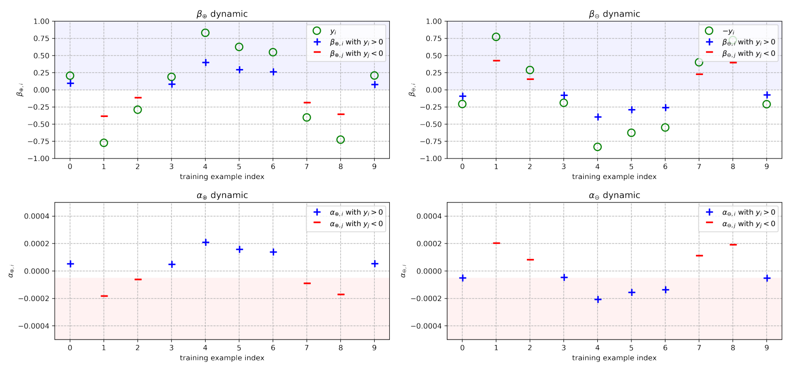
(*f*) Initialization 2 ($t = 80$).

Figure 3: We illustrate the prediction dynamics of gradient descent for a single ReLU model under different random initializations when d is comparable with n. In both cases, with sufficiently small step size, the final solution converges to a linear minimum-ℓ_2-norm interpolator on some subset of the training examples, i.e. of the form $\boldsymbol{w}_{\text{linear}-\text{MNI},S} = \boldsymbol{X}_S^\top(\boldsymbol{X}_S\boldsymbol{X}_S^\top)^{-1}\tilde{\boldsymbol{y}}_S$, where $\tilde{y}_{S,i} = \max\{y_i, 0\}$. In contrast to the high-dimensional regime, *different initializations lead to different subsets* S, indicating that ReLU training implicitly performs an example "selection" process, that is initialization-dependent, rather than fitting all positively-labeled samples. The experiment uses $n = 10$, $d = 50$, $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, $y \sim \mathcal{N}(0, 1)$, $\boldsymbol{w}^{(0)} \sim \mathcal{N}(\mathbf{0}, 2 \times 10^{-6}\boldsymbol{I})$, and $\eta = 10^{-4}$.

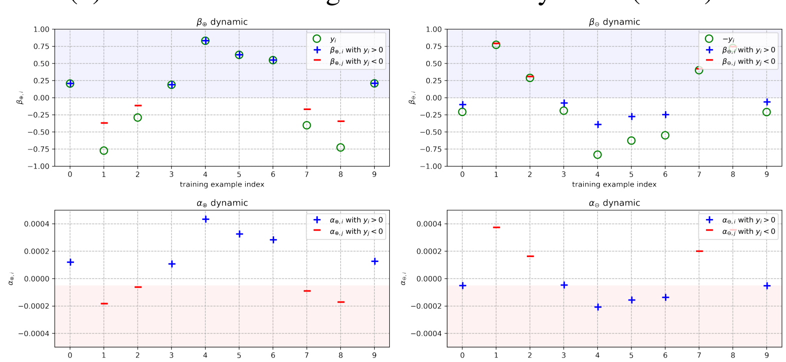
H.2. Gradient Descent Dynamics of Two ReLU Models



(*a*) Two ReLU model gradient descent dynamic ($t = 0$).

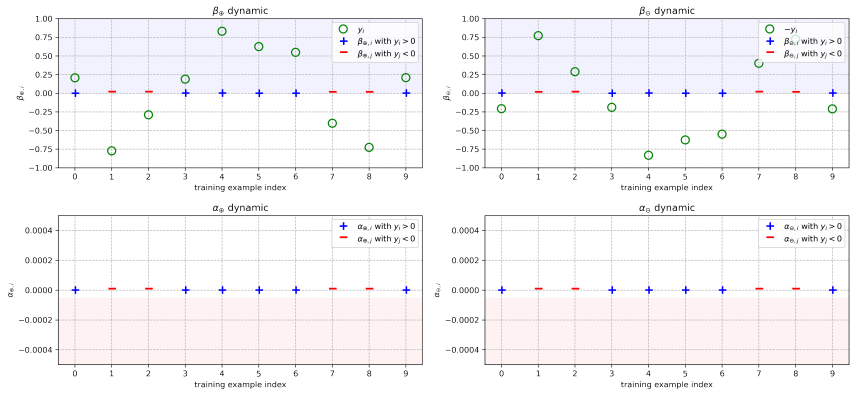


(*b*) Two ReLU model gradient descent dynamic ($t = 1$).

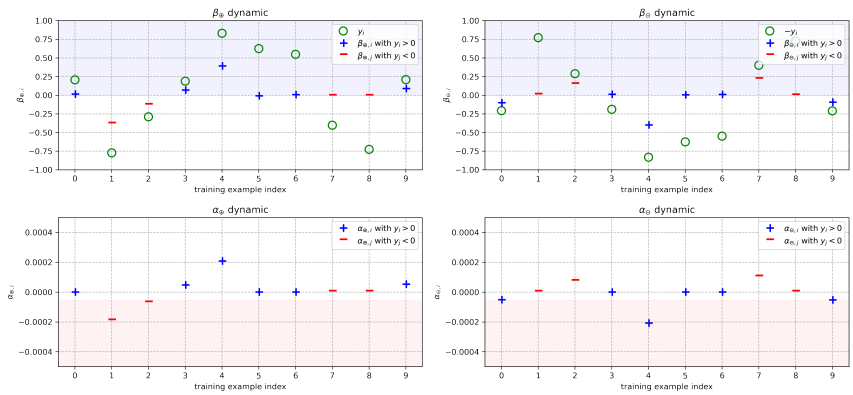


(*c*) Two ReLU model gradient descent dynamic ($t = 17$).

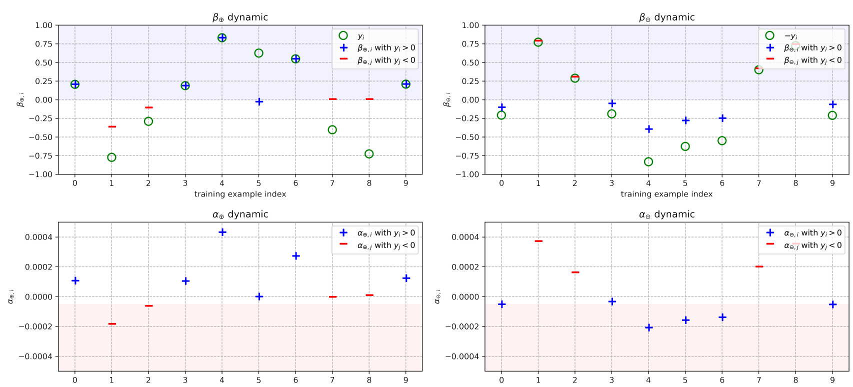
Figure 4: Simulation illustrating Theorem 7. In the high-dimensional regime and under our "all-positive" initialization, after the first gradient step, examples with positive labels remain active while examples with negative labels become inactive, consistent with Lemma 15. The blue region shows primal variables that remain positive over training, whereas the red region corresponds to dual variables that are sufficiently negative and remain unchanged. As training proceeds, $\boldsymbol{w}_{\oplus}$ fits all positively labeled examples and $\boldsymbol{w}_{\ominus}$ fits all negatively labeled examples. The experiment uses $n = 10$, $d = 2000$, features $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and labels satisfying $|y| \sim \mathcal{U}(0.1, 1)$ with $\mathrm{sign}(y)$ uniformly distributed over $\{\pm 1\}$.



(*a*) Two ReLU model gradient descent dynamic ($t = 0$).

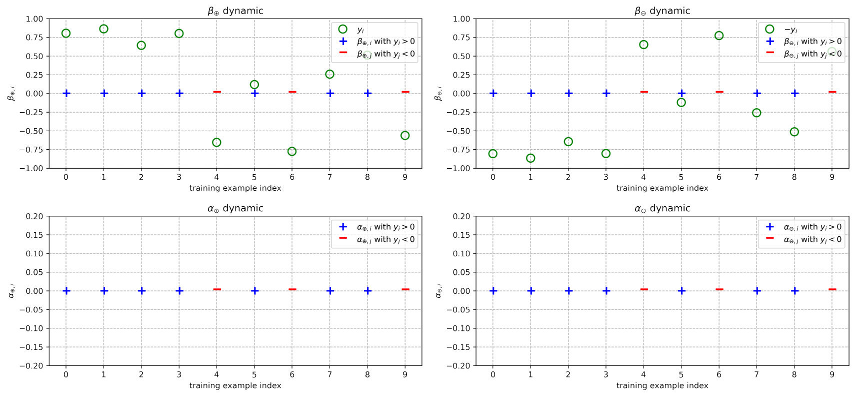


(*b*) Two ReLU model gradient descent dynamic ($t = 1$).

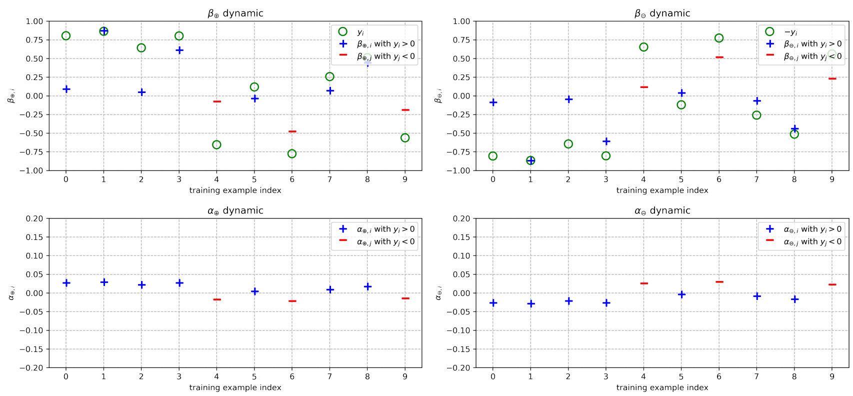


(*c*) Two ReLU model gradient descent dynamic ($t = 40$).

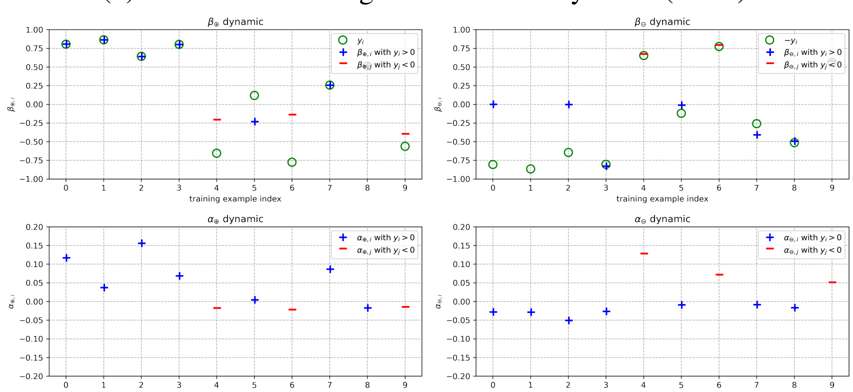
Figure 5: Simulation with random initialization in the high-dimensional regime, which violates our initialization assumption in Theorem 7. Under random initialization, the sufficient conditions of Lemma 15 are violated at the first gradient step. As a result, positively labeled examples do not all remain in the active (blue) regime (e.g. example no. 5), nor do negatively labeled examples consistently enter the inactive (red) regime (e.g. example no. 7). *Consequently, during training, this model fails to converge to a global minimum.* The experiment uses $n = 10$, $d = 2000$, features $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and labels satisfying $|y| \sim \mathcal{U}(0.1, 1)$ with $\operatorname{sign}(y)$ uniformly distributed over $\{\pm 1\}$.



(a) Two ReLU model gradient descent dynamic ($t = 0$).



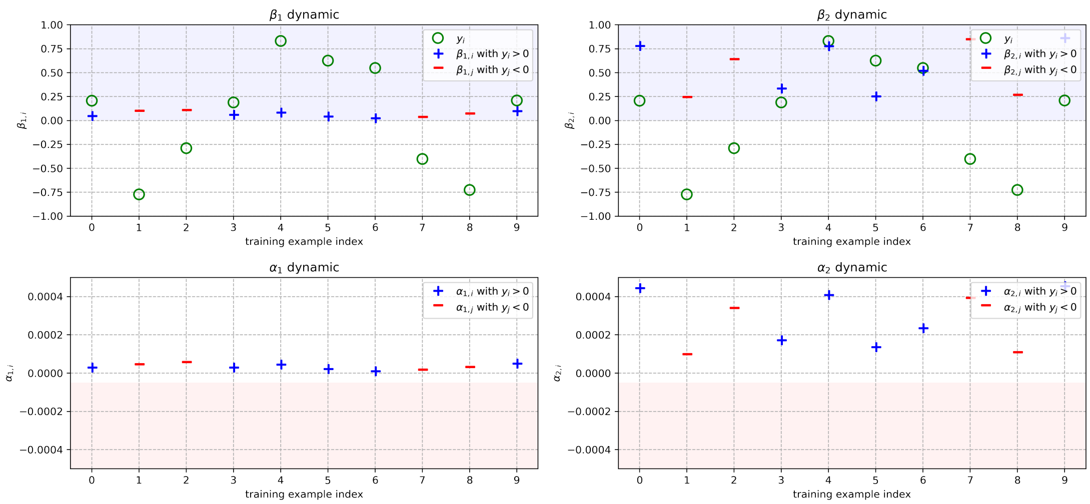
(b) Two ReLU model gradient descent dynamic ($t = 1$).



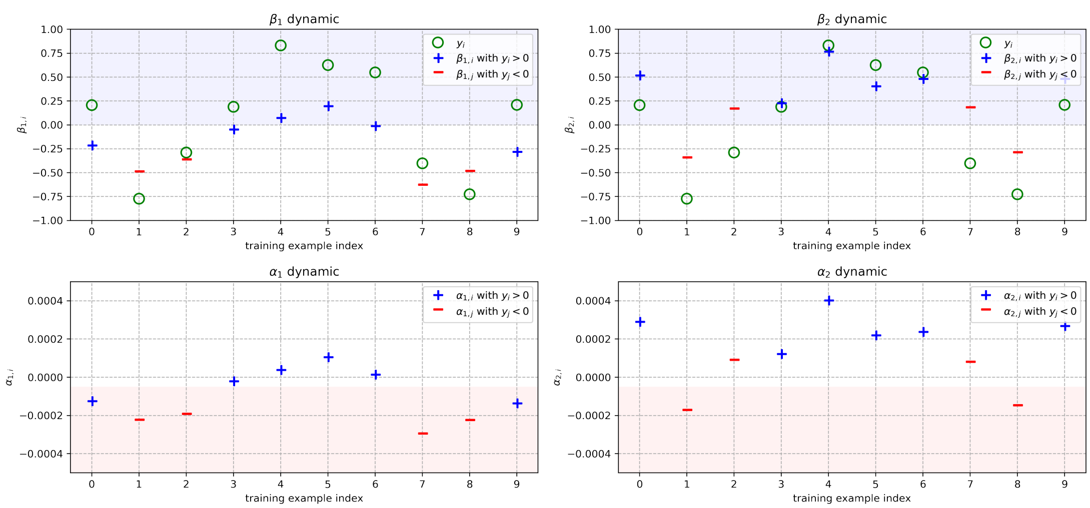
(c) Two ReLU model gradient descent dynamic ($t = 40$).

Figure 6: Simulation with all-positive initialization outside the high-dimensional regime. When the data dimension is not sufficiently large, the feature vectors are no longer approximately orthogonal. As a result, the clear separation into active (blue) and inactive (red) regimes observed in Figures 4 and 5 disappears. Consequently, the gradient dynamics become highly coupled across examples and are no longer analytically tractable using our high-dimensional arguments. The experiment uses $n = 10$, $d = 15$, features $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and labels satisfying $|y| \sim \mathcal{U}(0.1, 1)$ with $\mathrm{sign}(y)$ uniformly distributed over $\{\pm 1\}$.

H.3. Gradient Descent Dynamics of Multiple ReLU Models



(a) Multiple ReLU model gradient descent dynamic ($t = 0$).



(b) Multiple ReLU model gradient descent dynamic ($t = 1$).

Figure 7: Failure of stable activation patterns in multiple ReLU models. We illustrate the training dynamics of a multiple ReLU model when multiple neurons share the same sign. In this setting, the sufficient conditions of Lemma 19 are violated, and positive primal variables do not necessarily remain in the active (blue) regime throughout training (e.g. training example no. 0). As a result, the activation pattern becomes unstable, and the resulting primal dynamics are no longer tractable. The experiment uses $n = 10$, $d = 2000$, $m = 4$, with neuron signs $s_1 = s_2 = 1$ and $s_3 = s_4 = -1$, features $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and labels satisfying $|y| \sim \mathcal{U}(0.1, 1)$ with $\operatorname{sign}(y)$ uniformly distributed over $\{\pm 1\}$.