Achieving Margin Maximization Exponentially Fast via Progressive Norm Rescaling

Mingze Wang [1] **Zeping Min** [1] **Lei Wu** [1 2]

Abstract

In this work, we investigate the margin-maximization bias exhibited by gradient-based algorithms in classifying linearly separable data. We present an in-depth analysis of the specific properties of the velocity field associated with (normalized) gradients, focusing on their role in margin maximization. Inspired by this analysis, we propose a novel algorithm called Progressive Rescaling Gradient Descent (PRGD) and show that PRGD can maximize the margin at an *exponential rate*. This stands in stark contrast to all existing algorithms, which maximize the margin at a slow *polynomial rate*. Specifically, we identify mild conditions on data distribution under which existing algorithms such as gradient descent (GD) and normalized gradient descent (NGD) *provably fail* in maximizing the margin efficiently. To validate our theoretical findings, we present both synthetic and real-world experiments. Notably, PRGD also shows promise in enhancing the generalization performance when applied to linearly non-separable datasets and deep neural networks.

1. Introduction

In modern machine learning, models are often over-parameterized in the sense that they can easily interpolate all training data, giving rise to a loss landscape with many global minima. Although all these minima yield zero training loss, their generalization ability can vary significantly. Intriguingly, it is often observed that Stochastic Gradient Descent (SGD) and its variants consistently converge to solutions with favorable generalization properties even without needing any explicit regularization (Neyshabur et al., 2014; Zhang et al., 2017). This phenomenon implies that the "implicit bias" of SGD plays a crucial role in ensuring the efficacy of deep learning; therefore, revealing the underlying mechanism is of paramount importance.

Soudry et al. (2018) investigated implicit bias of GD for classifying linearly separable data with linear models. They showed that GD trained with exponentially-tailed loss functions can implicitly maximize the ℓ_2-margin during its convergence process, ultimately locating a max-margin solution. This discovery offers valuable insights into the superior generalization performance often observed with GD, as larger margins are generally associated with better generalization (Boser et al., 1992; Bartlett et al., 2017). However, the rate at which GD maximizes the margin has been shown to be merely $\mathcal{O}(1/\log t)$ (Soudry et al., 2018). This naturally leads to the question: can we design a better gradient-based algorithm to accelerate the margin maximization. In the pursuit of this, Nacson et al. (2019b); Ji & Telgarsky (2021) has demonstrated that employing GD with aggressively loss-scaled step sizes can achieve polynomial rates in margin maximization. Notably, Ji & Telgarsky (2021) specifically established that the rate of NGD is $\mathcal{O}(1/t)$. Building on this, Ji et al. (2021) further introduced a momentum-based gradient method by applying Nesterov acceleration to the dual formulation of this problem, which achieves a remarkable margin-maximization rate of $\mathcal{O}(\log t/t^2)$ and Wang et al. (2022b) further improved it to $\mathcal{O}(1/t^2)$, currently standing as the state-of-the-art algorithm for this problem.

Our Contributions. In this paper, we begin by introducing a toy dataset to elucidate the causes of inefficiency in GD/NGD and to clarify the underlying intuition for accelerating margin maximization. Subsequently, we demonstrate that these insights are applicable to a broader range of scenarios.

- We reveal that the rate of directional convergence and margin maximization is governed by the centripetal velocity–the component orthogonal to the max-margin direction. We show that under mild conditions on data distribution, NGD and GD will inevitably be trapped in a region where the centripetal velocity

[1]School of Mathematical Sciences, Peking University, Beijing, China [2]Center for Machine Learning Research, Peking University, Beijing, China. Correspondence to: Mingze Wang <mingzewang@stu.pku.edu.cn>, Lei Wu <leiwu@math.pku.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

is diminished, thereby explaining the inefficiency of GD/NGD. Specifically, we establish that the aforementioned margin-maximization rates: $\mathcal{O}(1/\log t)$ for GD and $\mathcal{O}(1/t)$ for NGD also serve as *lower bounds*.

- Based on the above observations, we propose to speed up the margin maximization by maintaining a non-degenerate centripetal velocity. We show that there exists a favorable region, where the centripetal velocity is uniformly lower-bounded and moreover, we can reposition parameters into this region via a simple *norm rescaling*. Leveraging these properties, we introduce an algorithm called Progressive Rescaling Gradient Descent (PRGD). Notably, we prove that PRGD can achieve both directional convergence and margin maximization at an *exponential* rate $\mathcal{O}(e^{-\Omega(t)})$. This stands in stark contrast to all existing algorithms, which maximize the margin at a slow polynomial rate.
- Lastly, we validate our theoretical findings through both synthetic and real-world experiments. In particular, when applying PRGD to linearly non-separable datasets and homogenized deep neural networks—beyond the scope of our theory—we still observe consistent test performance improvements.

In Table 1, we summarize our main theoretical results and compare them with existing ones.

2. Related Work

Unraveling the implicit bias of optimization algorithms has become a fundamental problem in theoretical deep learning and has garnered extensive attention recently.

Margin Maximization in Gradient-based Algorithms. The tendency of GD to favor max-margin solutions when trained with exponentially-tailed loss functions was first identified in the seminal work by Soudry et al. (2018). Beyond aforementioned studies, Nacson et al. (2019c) explored this bias for SGD and Gunasekar et al. (2018a); Wang et al. (2022a); Sun et al. (2022); Wang et al. (2023) extended the analysis to various other optimization algorithms. Notably, Ji & Telgarsky (2019b) considered the situation where dataset are linearly non-separable, providing insights into the robustness of these findings in more complex settings. More recently, Ji et al. (2020) investigated the impact of the tail behavior of loss functions and Wu et al. (2023) analyzed the impact of edge of stability, a phenomenon previously observed by Wu et al. (2018); Jastrzebski et al. (2020); Cohen et al. (2021).

Additionally, the margin-maximization analysis has also been extended to nonlinear models. This includes studies by Ji & Telgarsky (2019a); Gunasekar et al. (2018b) on deep linear networks, as well as research by Chizat & Bach (2020) on wide two-layer ReLU networks. Notably, Nacson et al. (2019a); Lyu & Li (2019); Ji & Telgarsky (2020) demonstrated that for general homogeneous networks, Gradient Flow (GF) and GD converge to solutions corresponding the KKT point of the max-margin problem. Kunin et al. (2023) has recently extended this analysis to quasi-homogeneous networks. Moreover, for two-layer (leaky-)ReLU neural networks, Lyu et al. (2021); Vardi et al. (2022); Wang & Ma (2023) studied whether the convergent KKT point of GF is a global optimum of the max-margin problem.

Other Implicit Biases. There are many other attempts to explain the implicit bias of deep learning algorithms (Vardi, 2023). Among them, the most popular one is the *flat minima hypothesis*: SGD favors flat minima (Keskar et al., 2016) and flat minima generalize well (Hochreiter & Schmidhuber, 1997; Jiang et al., 2019). Recent studies (Wu et al., 2018; Ma & Ying, 2021; Wu et al., 2022) pro-

Table 1: Comparison of the directional convergence and margin maximization rates of different algorithms under Assumption 3.1, 5.4, and $\gamma^\star \boldsymbol{w}^\star \neq \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \boldsymbol{x}_i y_i$. In this table, $\boldsymbol{w}^\star$ and $\gamma^\star$ denote the ℓ_2 max-margin solution and the corresponding margin, respectively and $\boldsymbol{w}(t)$ denotes the solution at the t-th step.

Algorithm	Error of Direction $\mathrm{e}(t) = \|\hat{\boldsymbol{w}}(t) - \boldsymbol{w}^\star\|$
GD	$\mathrm{e}(t) = \mathcal{O}(1/\log t)$ (Soudry et al., 2018); $\mathrm{e}(t) = \Theta\left(1/\log t\right)$ (**Thm** 6.4)
NGD	$\mathrm{e}(t) = \mathcal{O}(1/t)$ (Ji & Telgarsky, 2021); $\mathrm{e}(t_k) = \Theta(1/t_k)$ (**Thm** 6.4)
PRGD (Ours)	$\mathrm{e}(t) = e^{-\Omega(t)}$ (**Thm** 6.1)

Algorithm	Error of Margin $\mathrm{r}(t) = \gamma^\star - \gamma(\boldsymbol{w}(t))$
GD	$\mathrm{r}(t) = \mathcal{O}(1/\log t)$ (Soudry et al., 2018); $\mathrm{r}(t) = \Omega\left(1/\log^2 t\right)$ (**Thm** 6.4)
NGD	$\mathrm{r}(t) = \mathcal{O}(1/t)$ (Ji & Telgarsky, 2021); $\mathrm{r}(t_k) = \Omega(1/t_k^2)$ (**Thm** 6.4)
Nesterov Acceleration	$\mathrm{r}(t) = \tilde{\mathcal{O}}(1/t^2)$ (Ji et al., 2021; Wang et al., 2022b)
PRGD (Ours)	$\mathrm{r}(t) = e^{-\Omega(t)}$ (**Thm** 6.1)

vided explanations for why SGD tends to select flat minima from a dynamical stability perspective. Moreover, Blanc et al. (2020); Li et al. (2022a); Lyu et al. (2022); Ma et al. (2022) offered in-depth characterizations of the dynamical process of SGD in reducing the sharpness near the global minima manifold. Additionally, beyond empirical observations, recent studies (Ma & Ying, 2021; Mulayoff et al., 2021; Gatmiry et al., 2023; Wu & Su, 2023) provided theoretical evidence for the superior generalization performance of flat minima. Besides, Woodworth et al. (2020); Pesme et al. (2021); Nacson et al. (2022); Pesme & Flammarion (2023); Even et al. (2023) investigated the implicit bias on linear diagonal networks, such as how the initialization scale and the step size affect the selection bias of GF, GD, and SGD in different regimes. Additionally, various studies explored how other training components impact implicit bias, such as normalization (Wu et al., 2020; Li et al., 2020; Lyu et al., 2022; Dai et al., 2023), re-parametrization (Li et al., 2022b), weight decay (Andriushchenko et al., 2023), cyclic learning rate (Wang & Wu, 2023), and sharpness-aware minimization (Wen et al., 2023a;b; Long & Bartlett, 2023).

3. Preliminaries

Notation. We use bold letters for vectors and lowercase letters for scalars, e.g. $\boldsymbol{x} = (x_1, \cdots, x_d)^\top \in \mathbb{R}^d$. We use $\langle \cdot, \cdot \rangle$ for the standard Euclidean inner product between two vectors, and $\|\cdot\|$ for the ℓ_2 norm of a vector or the spectral norm of a matrix. Let $\hat{\boldsymbol{w}} = \boldsymbol{w}/\|\boldsymbol{w}\|$ the normalized vector. We use standard big-O notations $\mathcal{O}, \Omega, \Theta$ to hide absolute positive constants, and use $\tilde{\mathcal{O}}, \tilde{\Omega}, \tilde{\Theta}$ to further hide logarithmic constants. For any positive integer n, let $[n] = \{1, \cdots, n\}$.

Problem Setup. We consider the problem of binary classification with a linear decision function $\boldsymbol{x} \mapsto \langle \boldsymbol{w}, \boldsymbol{x} \rangle$. Let $\mathcal{S} = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$ for any $i \in [n]$ be the training set. Without loss of generality, we assume $\|\boldsymbol{x}_i\| \leq 1, \forall i \in [n]$. Throughout this paper, we assume $\mathcal{S}$ to be linearly separable:

Assumption 3.1 (linear separability)**.** There exists a $\boldsymbol{w} \in \mathbb{S}^{d-1}$ such that $\min\limits_{i \in [n]} y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle > 0$.

Under this assumption, the solutions that classify all training data correctly may not be unique. Among them, **the max-margin solution** is often favorable due to its superior generalization ability as suggested by the theory of support vector machine (Vapnik, 1999). For any $\boldsymbol{w} \in \mathbb{R}^d$, the normalized ℓ_2-margin is defined by $\gamma(\boldsymbol{w}) := \min_{i \in [n]} y_i \langle \hat{\boldsymbol{w}}, \boldsymbol{x}_i \rangle$. The max-margin solution and the corresponding max margin are defined by

$$\boldsymbol{w}^\star := \arg\max_{\boldsymbol{w} \in \mathbb{S}^{d-1}} \gamma(\boldsymbol{w}), \quad \gamma^\star := \gamma(\boldsymbol{w}^\star). \tag{1}$$

For the margin function, we have the following important properties:

- **Homogeneity.** $\gamma(c\boldsymbol{w}) = \gamma(\boldsymbol{w})$ for any $c > 0$ and $\boldsymbol{w} \in \mathbb{R}^d$.
- **Directional Convergence.** Under Assumption 3.1, $\gamma^\star - \gamma(\boldsymbol{w}) \leq \|\hat{\boldsymbol{w}} - \boldsymbol{w}^\star\|$ (Lemma A.1).

Therefore, instead of directly inspecting the margin maximization, we can focus on the analysis of directional convergence, which is often must easier.

In this paper, we are interested in algorithms that minimize the following objective

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^n \ell\left(y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle\right). \tag{2}$$

where $\ell : \mathbb{R} \mapsto \mathbb{R}_{\geq 0}$ is a loss function. We assume $\ell(z) = e^{-z}$ for simplicity and the extension to general loss functions with exponential-decay tails such as logistic loss $\ell(z) = \log(1 + e^{-z})$ are straightforward (Soudry et al., 2018; Nacson et al., 2019b).

Consider to solve the optimization problem (2) with GD:

$$\textbf{GD:} \quad \boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \eta \nabla \mathcal{L}(\boldsymbol{w}(t)). \tag{3}$$

Soudry et al. (2018) showed under Assumption 3.1, GD (3) with $\eta \leq 1$ converges in direction to the max-margin solution $\boldsymbol{w}^\star$ despite the non-uniqueness of solutions. However, this occurs at a slow logarithmic rate $\gamma^\star - \gamma(\boldsymbol{w}(t)) = \mathcal{O}(1/\log t)$. To accelerate the convergence, Nacson et al. (2019b); Ji & Telgarsky (2021) proposed the following Normalized Gradient Descent (NGD):

$$\textbf{NGD:} \quad \boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \eta \frac{\nabla \mathcal{L}(\boldsymbol{w}(t))}{\mathcal{L}(\boldsymbol{w}(t))}, \tag{4}$$

and Ji & Telgarsky (2021) show that for NGD with $\eta \leq 1$, the margin maximization is much faster: $\gamma^\star - \gamma(\boldsymbol{w}(t)) = \mathcal{O}(1/t)$.

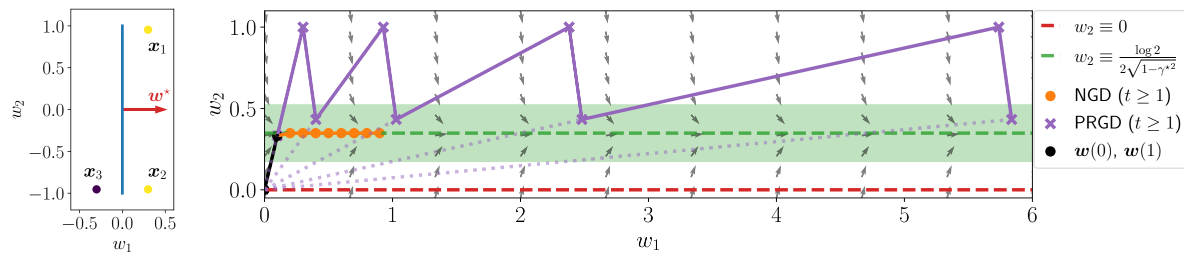
4. Motivations and the Algorithm

In this section, we propose a toy problem to showcase why NGD is slow in maximizing the margin and why our proposed algorithm can accelerate it significantly.

Dataset 1. $\mathcal{S} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), (\boldsymbol{x}_3, y_3)\}$ where $\boldsymbol{x}_1 = (\gamma^\star, \sqrt{1 - \gamma^{\star 2}})^\top$, $y_1 = 1$, $\boldsymbol{x}_2 = (\gamma^\star, -\sqrt{1 - \gamma^{\star 2}})^\top$, $y_2 = 1$, $\boldsymbol{x}_3 = (-\gamma^\star, -\sqrt{1 - \gamma^{\star 2}})^\top$, $y_3 = -1$, and $\gamma^\star > 0$.

For this particular dataset, the max-margin solution is $\boldsymbol{w}^\star = \boldsymbol{e}_1 = (1, 0)^\top$ and $\gamma^\star$ represents the associate margin. A visualization of this dataset can be found in Figure 1a.

The Vector Field. To gain an intuitive understanding of why NGD dynamics is slow in margin maximization for



(a) Dataset 1. (b) The vector field and the trajectories of NGD and PRGD for Dataset 1.

Figure 1: (a) A visualization of Dataset 1 where $\boldsymbol{w}^\star$ is the max-margin solution. (b) The vector field and the trajectories of NGD and PRGD for Dataset 1. The gray arrows plot the vector field $-\nabla\mathcal{L}(\cdot)/\left\|\nabla\mathcal{L}(\cdot)\right\|$; the red dashed line corresponds to the max-margin solution $\boldsymbol{w}^\star$; the green zone $\mathbb{A}$ is an "attractor" of NGD dynamics. We plot the trajectories of PPGD and NGD for 8 iterations starting from the same initial point $\boldsymbol{w}(1)$ (black), where $\boldsymbol{w}(1)$ is trained by NGD starting from $\boldsymbol{w}(0) = \mathbf{0}$ (black).

Dataset 1, we visualize the direction of normalized gradient in Figure 1b (the gray arrows). We can see that the **centripetal velocity**, i.e., the component orthogonal to $\boldsymbol{w}^\star$, becomes tiny in the green zone $\mathbb{A}$ and diminishes to zero at the green line. Consequently, NGD (the orange curve) always enters and is attracted in $\mathbb{A}$ (and remain close the green line). A rigorous analysis in Appendix A shows that the "attractor" of NGD (i.e., the green zone) is given by $\mathbb{A} := \left\{\boldsymbol{w} : \log 2/4 \leq w_2\sqrt{1-\gamma^{\star 2}} \leq 3\log 2/4\right\}$.

The Inefficiency of NGD. We can see that while NGD keeps trapped in the attractor $\mathbb{A}$, $\boldsymbol{w}(t)$ moves towards infinity and accordingly, $\hat{\boldsymbol{w}}(t) \to \boldsymbol{w}^\star$ as $t \to \infty$. However, due to the gap between $\mathbb{A}$ and the max-margin direction $\boldsymbol{w}^\star$, we have $|w_2(t)| = \Theta(1)$ for all $t \in \mathbb{N}$. Consequently, the directional convergence of NGD is cursed to have a rate at most

$$\begin{aligned}\left\|\hat{\boldsymbol{w}}(t) - \boldsymbol{w}^\star\right\| =& \sqrt{\left(\frac{w_1(t)}{\left\|\boldsymbol{w}(t)\right\|} - 1\right)^2 + \left(\frac{w_2(t)}{\left\|\boldsymbol{w}(t)\right\|}\right)^2} \\ =& \Theta\left(\frac{|w_2(t)|}{\left\|\boldsymbol{w}(t)\right\|}\right) = \Theta\left(\frac{1}{t}\right),\end{aligned}$$

where we use the fact that the norm $\left\|\boldsymbol{w}(t)\right\|$ grows at $\Theta(t)$ rate (Lemma C.5).

Acceleration via Amplifying the Centripetal Velocity. In Figure 1b, we can see that the centripetal velocity is helpful for converging towards $\boldsymbol{w}^\star$ when $\boldsymbol{w}$ is away from $\boldsymbol{w}^\star$. The inefficiency of NGD stems from the fact that NGD is trapped in the green zone where the centripetal velocity is tiny. Therefore, to accelerate the directional convergence, we can stretch $\boldsymbol{w}(t)$ outside the the green zone via *rescaling*: $\boldsymbol{w}(t) \to c\boldsymbol{w}(t)$ for some $c > 1$. This rescaling does not change the margin (due to the homogeneity) but reposition $\boldsymbol{w}(t)$ into a region where the centripetal velocity is lower bounded, thereby enabling a faster directional convergence when employing NGD steps there.

Based the above intuition, we propose the *Progressive Rescaling Gradient Descent* (PRGD) given in Alg. 1. The additional projection step in Alg. 1 is proposed to stablize training by avoiding the rapid explosion of parameters' norm in each cycle. It is shown to be useful in experiments but does not affect our theoretical results (Theorem 6.1).

Algorithm 1 Progressive Rescaling Gradient Descent (PRGD)

Input: Dataset $\mathcal{S}$; Initialization $\boldsymbol{w}(0)$; Progressive Time $\{T_k\}_{k=0}^K$; Progressive Radius $\{R_k\}_{k=0}^K$;

for $k = 0, 1, 2, \cdots, K$ **do**

  $\boldsymbol{w}(T_k + 1) = R_k \frac{\boldsymbol{w}(T_k)}{\left\|\boldsymbol{w}(T_k)\right\|}$; ▷ progressive rescaling step

  for $T_k + 1 \leq t \leq T_{k+1} - 1$ **do**

    $\boldsymbol{v}(t+1) = \boldsymbol{w}(t) - \eta\frac{\nabla\mathcal{L}(\boldsymbol{w}(t))}{\mathcal{L}(\boldsymbol{w}(t))}$; ▷ normalized gradient descent step

    $\boldsymbol{w}(t+1) = \mathrm{Proj}_{\mathbb{B}(\mathbf{0}, R_k)}\left(\boldsymbol{v}(t+1)\right)$; ▷ projection step

Output: $\boldsymbol{w}(T_K + 1)$.

The Efficiency of PRGD. In Figure 1b, we plot the trajectory of PRGD (the purple curve) with hyperparameter $T_{k+1} - T_k = 2$. This means that in each cycle, PRGD executes one step of norm rescaling, followed by one step of projected NGD. It is evident from the figure that PRGD converges towards $\boldsymbol{w}^*$ in direction much faster. This acceleration can be attributed to the following mechanism: the rescaling step allows PRGD to move out of the attractor $\mathbb{A}$ (where the centripetal velocity is tiny) and to undertake NGD in the line $w_2 = 1$, where *the centripetal velocity has a uniformly positive lower bound.* Utilizing this fact, we can show that the norm increases exponentially fast via simple geometric calculation. Consequently, the directional convergence is exponentially fast

$$\left\|\hat{\boldsymbol{w}}(2k-1) - \boldsymbol{w}^\star\right\| = \left\|\hat{\boldsymbol{w}}(2k) - \boldsymbol{w}^\star\right\|$$

$$=\Theta\left(\frac{|w_2(2k)|}{\|\boldsymbol{w}(2k)\|}\right)=\Theta\left(\frac{1}{\|\boldsymbol{w}(2k)\|}\right)=e^{-\Omega(k)}.$$

The following proposition formalizes the above intuitive analysis of NGD and PRGD for Dataset 1, whose proof is deferred to Appendix A.

Proposition 4.1. *Consider Dataset 1. Then NGD (4) can only maximize the margin polynomially fast, while PRGD (Alg. 1) can maximize the margin exponentially fast. Specifically,*

- ***(NGD).*** *Let $\boldsymbol{w}(t)$ be NGD (4) solution at time t with $\eta=1$ starting from $\boldsymbol{w}(0)=\mathbf{0}$. Then both the margin maximization and directional convergence are at (tight) polynomial rates: $\|\hat{\boldsymbol{w}}(t)-\boldsymbol{w}^\star\|=\Theta(1/t),\gamma^\star-\gamma(\boldsymbol{w}(t))=\Theta(1/t)$.*
- ***(PRGD).*** *Let $\boldsymbol{w}(1)$ be NGD (4) solution at time 1 with $\eta=1$ starting from $\boldsymbol{w}(0)=\mathbf{0}$, and let $\boldsymbol{w}(t)$ be PRGD solution (Alg. 1) at time t with $\eta=1$ starting from $\boldsymbol{w}(1)$. Then there exists a set of hyperparameters $\{R_k\}_k$ and $\{T_k\}_k$ such that $R_k=e^{\Theta(k)}$ and $T_k=\Theta(k)$, and both the margin maximization and directional convergence are at (tight) exponential rate: $\|\hat{\boldsymbol{w}}(t)-\boldsymbol{w}^\star\|=e^{-\Theta(t)},\gamma^\star-\gamma(\boldsymbol{w}(t))=e^{-\Theta(t)}$.*

5. Centripetal Velocity Analysis

In the above analysis, the key property enabling the acceleration for Dataset 1 is the existence a region where the centripetal velocity is uniformly lower bounded and we can stretch $\boldsymbol{w}(t)$ to this region by simple norm rescaling. In this section, we demonstrate that this property holds generally.

We needs the following decomposition of parameter for our fine-grained analysis of directional dynamics. Note that the same decomposition has been employed in Ji & Telgarsky (2021); Wu et al. (2023).

Definition 5.1. Let $\mathcal{P}(\boldsymbol{w}):=\langle\boldsymbol{w},\boldsymbol{w}^\star\rangle\boldsymbol{w}^\star$ and $\mathcal{P}_\perp(\boldsymbol{w}):=\boldsymbol{w}-\langle\boldsymbol{w},\boldsymbol{w}^\star\rangle\boldsymbol{w}^\star$. It is worth noting that for any $\boldsymbol{w}\in\mathbb{R}^d$, we have the following decomposition

$$\boldsymbol{w}=\mathcal{P}(\boldsymbol{w})+\mathcal{P}_\perp(\boldsymbol{w}).$$

We can now formally propose the definition of the "centripetal velocity" as follows and a visual illustration of the definition is provided in Figure 2.

Definition 5.2 (Centripetal Velocity)**.** The normalized gradient at $\boldsymbol{w}\in\mathbb{R}^d$ is $\nabla\mathcal{L}(\boldsymbol{w})/\mathcal{L}(\boldsymbol{w})$ and we define the centripetal velocity $\varphi(\boldsymbol{w})$ at $\boldsymbol{w}$ by

$$\varphi(\boldsymbol{w}):=\left\langle-\frac{\nabla\mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})},-\frac{\mathcal{P}_\perp(\boldsymbol{w})}{\|\mathcal{P}_\perp(\boldsymbol{w})\|}\right\rangle.$$

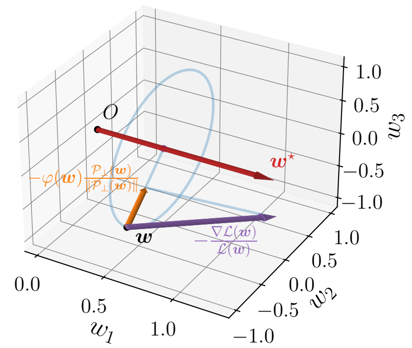


Figure 2: A visual illustration of Definition 5.2 in $\mathbb{R}^3$. The red arrow corresponds to the max-margin direction $\boldsymbol{w}^\star$. At $\boldsymbol{w}\in\mathbb{R}^3$, the purple arrow signifies the normalized negative gradient; the orange arrow depicts the projection of $-\nabla\mathcal{L}(\boldsymbol{w})/\mathcal{L}(\boldsymbol{w})$ along the centripetal direction $-\mathcal{P}_\perp(\boldsymbol{w})/\|\mathcal{P}_\perp(\boldsymbol{w})\|$, reflecting the *centripetal velocity* $\varphi(\boldsymbol{w})$.

In addition, our subsequent analysis crucially relies on the following geometry:

Definition 5.3 (Semi-infinite Hollow Cylinder)**.** We use

$$\begin{aligned}\mathbb{C}(D_1,D_2;H):=\big\{&\boldsymbol{w}\in\operatorname{span}\{\boldsymbol{x}_i:i\in[n]\}:\\&D_1\le\|\mathcal{P}_\perp(\boldsymbol{w})\|\le D_2;\langle\boldsymbol{w},\boldsymbol{w}^\star\rangle\ge H\big\}\end{aligned}\tag{5}$$

to denote a cylinder which starts from the height H and extends infinitely along the direction $\boldsymbol{w}^\star$.

Our subsequent analysis will concentrate on this semi-infinite hollow cylinder as PRGD ensures the iterations will be confined in the region. Additionally, it is crucial to note that our attention is restricted to the smaller subspace $\operatorname{span}\{\boldsymbol{x}_i:i\in[n]\}$, rather than the entire space $\mathbb{R}^d$. This is justified by fact that the trajectories of GD, NGD, and PRGD, when initialized from $\mathbf{0}$, will remain staying in this subspace indefinitely.

5.1. The Existence of a Favorable Semi-infinite Hollow Cylinder

In this subsection, we undertake a theoretical examination of the centripetal velocity, as defined in Definition 5.2, on the semi-infinite hollow cylinder described in Definition 5.3. Our investigation aims to address the following query:

Does a "favorable" semi-infinite hollow cylinder exist where the centripetal velocity consistently maintains a positive lower bound?

Assumption 5.4 (Non-degenerate data)**.** Let $\mathcal{I}$ be the index set of the support vectors. Assume there exist $\alpha_i>0$ ($i\in\mathcal{I}$) such that $\boldsymbol{w}^\star=\sum_{i\in\mathcal{I}}\alpha_i\boldsymbol{x}_iy_i$ and $\operatorname{span}\{\boldsymbol{x}_i:i\in\mathcal{I}\}=\operatorname{span}\{\boldsymbol{x}_i:i\in[n]\}$.

Under this assumption, we can establish the existence of a favorable semi-infinite cylindrical surface as follows. The proof is deferred to Appendix B.

Theorem 5.5 (Centripetal Velocity Analysis, Main result)**.** *Under Assumption 3.1 and 5.4, there exists a semi-infinite hollow cylinder $\mathbb{C}(D, 2D; H)$ and a positive constant $\mu > 0$ such that*

$$\inf_{\boldsymbol{w}\in\mathbb{C}(D,2D;H)} \varphi(\boldsymbol{w}) \geq \mu.$$

Assumption 5.4 has been widely used in prior analysis of the margin-maximization bias of gradient-based algorithms. The first part, requiring strictly positive dual variables to ensure directional convergence, is rather weak and holds for almost all linearly separable data (Soudry et al., 2018; Nacson et al., 2019b; Ji & Telgarsky, 2021; Wang et al., 2022a). The second part requires the support vectors to span the entire dataset. This condition has been adopted in refined analysis of the residual in Theorem 4 of (Soudry et al., 2018), the analysis of SGD dynamics (Nacson et al., 2019c), and the edge of stability (Wu et al., 2023). We emphasize that the second part is only a technical condition as experimental results for real-world datasets in Section 7.1 demonstrate that PRGD performs effectively even in cases where this condition is not met. We leave the relaxation of this condition for future work.

6. Convergence Analysis

6.1. Exponentially Fast Margin Maximization of PRGD

Theorem 5.5 ensures the existence of a favorable semi-infinite hollow cylinder. The following theorem shows that PRGD (Alg. 1) can leverage the favorability to achieve an exponential rate in directional convergence and margin maximization.

Theorem 6.1 (PRGD, Main Result)**.** *Suppose that Assumption 3.1 and 5.4 hold. Let $\mathbf{w}(t)$ be solution generated by the following* ***two-phase*** *algorithms starting from $\mathbf{w}(0) = \mathbf{0}$:*

- *Warm-up Phase: Run GD (3) or NGD (4) with $\eta \leq 1$ for T_{w} steps starting from $\mathbf{w}(0)$;*
- *Acceleration Phase: Run PRGD (Alg. 1) with some $\eta, \{R_k\}_k, \{T_k\}_k$ starting from $\mathbf{w}(T_{\mathrm{w}})$.*

Then there exist a set of hyperparameters $\eta = \Theta(1)$, $R_k = e^{\Theta(k)}$ and $T_k = \Theta(k)$, ensuring both directional convergence and margin maximization occur at ***exponential*** *rates:*

$$\|\hat{\boldsymbol{w}}(t) - \boldsymbol{w}^\star\| = e^{-\Omega(t)}; \quad \gamma^\star - \gamma(\boldsymbol{w}(t)) = e^{-\Omega(t)}.$$

This theorem demonstrates that PRGD can achieve both directional convergence and margin maximization exponentially fast. In stark contrast, all existing algorithms maximize the margin at notably slower rates, including $\mathcal{O}(1/\log t)$ for GD (Soudry et al., 2018), $\mathcal{O}(1/t)$ for NGD (Ji & Telgarsky, 2021), and $\tilde{\mathcal{O}}(1/t^2)$ for Dual Acceleration (Ji et al., 2021; Wang et al., 2022b).

The complete proof of Theorem 6.1 is deferred to Appendix C.1 and here, we provide a proof sketch to illustrate the intuition behind:

- The initial warm-up phase utilizes GD to secure a preliminary directional convergence, albeit at a slower rate, such that the condition $\|\hat{\boldsymbol{w}}(T_{\mathrm{w}}) - \boldsymbol{w}^\star\| < \min\{D/2H, 1/2\}$ is satisfied. This condition is crucial as it allows for the subsequent stretching of $\boldsymbol{w}(T_{\mathrm{w}})$ to the favorable semi-infinite hollow cylinder $\mathbb{C}(D, 2D; H)$ (in Theorem 5.5) through a straightforward norm rescaling. Without this condition, rescaling cannot reposition $\boldsymbol{w}$ into the favorable region.
- Following the warm-up phase, rescaling is employed to position $\boldsymbol{w}(T_{\mathrm{w}})$ into the favorable semi-infinite hollow cylinder $\mathbb{C}(D, 2D; H)$ by choosing $R_1 = \frac{D}{\|\mathcal{P}_\perp(\boldsymbol{w}(T_{\mathrm{w}}))\|}$. NGD steps taken thereafter yield a significant directional convergence, however, NGD steps also drive solutions to leave away from $\mathbb{C}(D, 2D; H)$. To overcome this issue, by setting a sequence of progressively increasing radii $\{R_k\}$, we can reposition the parameter back to $\mathbb{C}(D, 2D; H)$ again, as is evident illustrated in Figure 1b. Lastly, through a simple geometric calculation, we can demonstrate that such directional convergence is exponentially fast.

Remark 6.2. We clarify that Theorem 6.1 establishes the exponentially fast margin maximization of PRGD under a particular family of hyperparameters. For a broader range of hyperparameter choices, we delve into experimetnal explorations in Section 7.

Remark 6.3. In Proposition 4.1, we have provided a tightly exponentially fast rate on Dataset 1, which satisfies Assumption 3.1 and 5.4. Hence, the tightness of Theorem 6.1 is ensured.

6.2. Inefficiency of GD and NGD

In order to theoretically justify RPGD's superiority over GD and NGD, we need lower bounds of directional convergence and margin maximization rates for GD and NGD. However, as mentioned above, previous studies have only established the upper bounds of GD and NGD as $\mathcal{O}(1/\log t)$ and $\mathcal{O}(1/t)$, respectively. In this section, we further identify mild assumptions, under which we show that those rates also serve as the lower bounds for GD and NGD.

Theorem 6.4 (GD and NGD, Main Results)**.** *Suppose Assumption 3.1 and 5.4 hold. Additionally, we assume $\gamma^\star \boldsymbol{w}^\star \neq \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \boldsymbol{x}_i y_i$. Then,*

- *For GD (3) with $\eta \leq \eta_0$ starting from $\boldsymbol{w}(0) = \mathbf{0}$ (where η_0 is a constant), it holds that*

$$\|\hat{\boldsymbol{w}}(t) - \boldsymbol{w}^\star\| = \Theta(1/\log t);$$
$$\gamma^\star - \gamma(\boldsymbol{w}(t)) = \Omega\left(1/\log^2 t\right).$$

- *For NGD (4) with $\eta \leq \eta_0$ starting from $\boldsymbol{w}(0) = \mathbf{0}$ (where η_0 is a constant), there exists a subsequence $\boldsymbol{w}(t_k)$ ($t_k \to \infty$) such that*

$$\|\hat{\boldsymbol{w}}(t_k) - \boldsymbol{w}^\star\| = \Theta(1/t_k);$$
$$\gamma^\star - \gamma(\boldsymbol{w}(t_k)) = \Omega\left(1/t_k^2\right).$$

To the best of our knowledge, Theorem 6.4 provides the **first** lower bounds of both directional convergence and margin maximization for GD and NGD. We anticipate that this result can help build intuitions for future analysis of the implicit bias and convergence of gradient-based algorithms.

As presented in Table 1, under the same conditions–Assumption 3.1, 5.4, and $\gamma^\star \boldsymbol{w}^\star \neq \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \boldsymbol{x}_i y_i$, PRGD can achieve directional convergence *exponentially fast* with the rate $e^{-\Omega(t)}$. In contrast, Theorem 6.4 ensures that GD exhibits a *tight* bound with exponentially slow rate $\Theta(1/\log t)$, as well as NGD maintains a tight bound of polynomial speed $\Theta(1/t_k)$. Moreover, for margin maximization, Theorem 6.4 also provides *nearly tight* lower bounds for GD and NGD.

The detailed proof of Theorem 6.4 is provided in Appendix C.2. While the proof of the lower bounds involves a more intricate convex optimization analysis compared to Proposition 4.1 (especially for NGD due to its aggressive step size), the fundamental insights shared by both proofs are remarkably similar. Specifically, for NGD, our conditions ensures the existence of a nearly "attractor" region, such that (i) there exists a sequence of NGD $\boldsymbol{w}(t_k)$ ($t_k \to \infty$) falls within this region; (ii) the condition $\gamma^\star \boldsymbol{w}^\star \neq \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \boldsymbol{x}_i y_i$ ensures a $\Omega(1)$ distance between the attractor region and the max-margin direction. Since the norm grows as $\|\boldsymbol{w}(t_k)\| = \Theta(t_k)$, NGD is cursed to have only $\Omega(1/\|\boldsymbol{w}(t_k)\|) = \Omega(1/t_k)$ directional convergence rate.

7. Numerical Experiments

7.1. Linearly Separable Datasets

Experiments on Synthetic Dataset. We start our experimental validations with two synthetic linearly separable datasets. For synthetic datasets, the value of $\gamma^\star$ is explicit, and as such, we can explicitly compute the margin gap. To ensure a fair comparison, we maintain the same step size $\eta = 1$ for all GD, NGD, and PRGD.

While our theoretical analysis (Theorem 6.1) is confined to a specific set of hyper-parameters, it is worth noting that Theorem 5.5, a crucial property used in the proof of Theorem 6.1, holds over a relatively broad region. This flexibility enables a simpler selection of hyperparameters. Following the guidelines provided in Theorem 6.1, we employ PRGD(exp) with hyperparameters:

$$T_{k+1} - T_k \equiv 5,\ R_k = R_0 \times 1.2^k.$$

To illustrate the impact of the progressive radius, we also examine PRGD(poly) configured with

$$T_{k+1} - T_k \equiv 5,\ R_k = R_0 \times k^{1.2},$$

where the progressive radius increases polynomially. For more experimental details, refer to Appendix E.1. Some of the experimental results are provided in Figure 3, and the complete results are referred to Appendix E.1. Consistent with Theorem 6.1, PRGD(exp) indeed maximizes the margin (super-)exponentially fast, and surprisingly, PRGD(poly) also performs relatively well for this task. In contrast, NGD and GD reduce the margin gaps much more slowly, which substantiates Theorem 6.4.

Experiments on Real-World Datasets. In this case, we extend our experiments to real-world datasets. Specifically, we employ the `digit` datasets from `Sklearn`, which are image classification tasks with $d = 64$, $n = 300$. In this real-world setting, we lack prior knowledge of the exact $\gamma^\star$. Instead, we approximate $\gamma^\star$ by employing $\gamma(\boldsymbol{w}(t))$

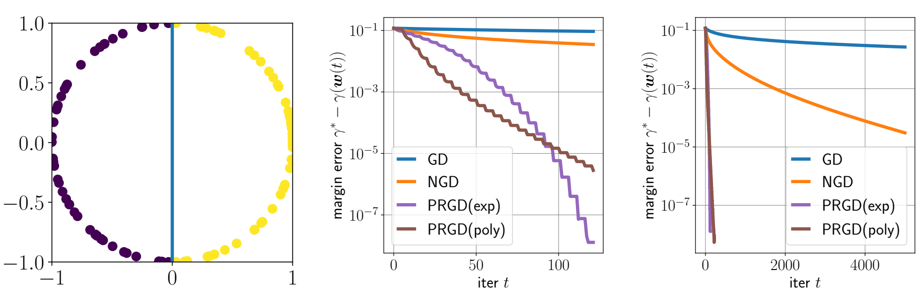


Figure 3: Comparison of margin maximization rates of different algorithms on a synthetic dataset. (left) A visualization of the 2d synthetic dataset. The yellow points represent the data with label 1, while the purple points corresponds to the data with label 1; (middle)(right) The comparison of margin maximization rates of different algorithms on this dataset at small and large time scales, respectively.

obtained by a sufficiently trained NGD. Additionally, it is worth noting that these datasets do not satisfy the second part of Assumption 5.4. For instance, in the `digit-01` dataset, $\mathrm{rank}\{\boldsymbol{x}_i : i \in \mathcal{I}\} = 2 < \mathrm{rank}\{\boldsymbol{x}_i : i \in [n]\} = 51$, where $\mathrm{rank}\{\boldsymbol{x}_i : i \in \mathcal{I}\}$ is calculated by approximating $\boldsymbol{w}^\star$ using $\hat{\boldsymbol{w}}(t)$, obtained through training NGD sufficiently.

In real experiments, we test both PRGD(exp) and PRGD(poly) and consistently observe that the latter performs much better. Therefore, in this experiment, we employ a modified variant of PRGD with smaller progressive radii:

$$R_k = R_0 \cdot k^{\alpha}, \quad T_{k+1} - T_k = T_0 \cdot k^{\beta}, \tag{6}$$

where α, β are hyperparameters to be tuned.

The results with well-tuned hyperparameters $\alpha = \beta = 0.6$ are presented in Figure 4. It is evident that, in these real-world datasets, PRGD consistently beats GD and NGD in terms of margin maximization rates.

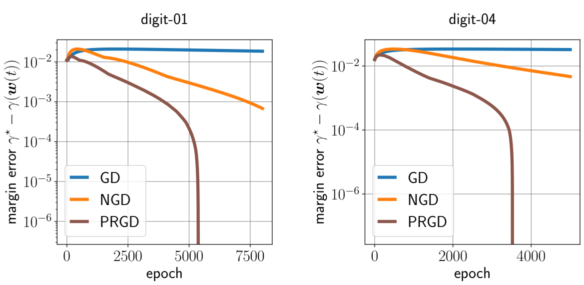
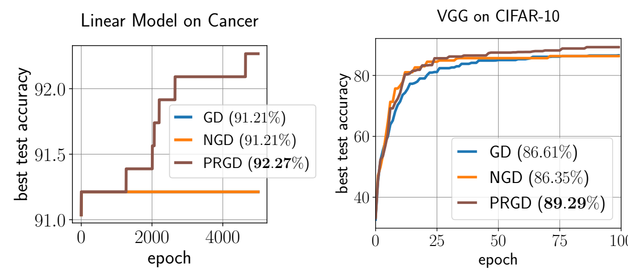


Figure 4: Comparison of margin maximization rates of different algorithms on `digit` (real-word) datasets. (Left) the results on `digit-01` dataset; (Right) the results on `digit-04` dataset.

7.2. Linearly Non-separable Datasets and Deep Neural Networks

In this subsection, we further explore the potential practical utilities of PRGD for datasets that are not linearly separable. 1) In the **first** experiment, we still consider linear models but for classifying a linearly non-separable dataset, `Cancer` in `Sklearn`. 2) For the **second** experiment, we examine the performance of PRGD for deep neural networks. Inspired by Lyu & Li (2019); Ji & Telgarsky (2020), the max-margin bias also exists for homogenized neural networks. Thus, we follow Lyu & Li (2019) and examine our algorithm for homogenized VGG-16 network (Simonyan & Zisserman, 2015) on the full CIFAR-10 dataset (Krizhevsky & Hinton, 2009), without employing any explicit regularization. Additionally, in this setting, we employ mini-batch stochastic gradient instead of the full gradient for these algorithms, and we also fine-tune the learning rates of GD, NGD, and PRGD. Both NGD and PRGD share the same learning rate scheduling strategy as described in Lyu & Li (2019). For both experiments, we follow the same strategy as described in (6) to tune the progressive hyperparameters of PRGD. For more experimental details, please refer to Appendix E.2.

The experimental results are presented in Fig 5a and Fig 5b, respectively. One can see that our PRGD algorithm archives better generalization performance and outperforms GD and NGD for both tasks.



(a) Linear Model on `Cancel` (b) VGG on Cifar-10

Figure 5: Comparison of the generalization performance of GD, NGD, and PRGD for non-linearly separable datasets and deep neural networks.

8. Concluding Remark

In this work, we investigate the mechanisms driving the convergence of gradient-based algorithms towards max-margin solutions. Specifically, we elucidate why GD and NGD can only achieve polynomially fast margin maximization by examining the properties of the velocity field linked to (normalized) gradients. This analysis inspires the design of a novel algorithm called PRGD that significantly accelerates the process of margin maximization. To substantiate our theoretical claims, we offer both synthetic and real-world experimental results, thereby underscoring the potential practical utility of the proposed PRGD algorithm. Looking ahead, an intriguing avenue for future research is the application of progressive norm rescaling techniques to state-of-the-art real-world models. In addition, it would be worthwhile to explore how PRGD can cooperate with other explicit regularization techniques, such as batch normalization, dropout, and sharpness-aware minimization (Foret et al., 2021), to further improve the generalization performance.

Acknowledgements

Mingze Wang is supported in part by the National Key Basic Research Program of China (No. 2015CB856000). Lei Wu is supported in part by a startup fund from Peking University. We thank Dr. Chao Ma, Zihao Wang, Zhonglin Xie for helpful discussions and anonymous reviewers for their valuable suggestions.

Impact Statement

This paper presents work whose goal is to advance the theoretical aspects of Machine Learning and, specifically, the margin maximization bias of gradient-based algorithms. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Andriushchenko, M., D'Angelo, F., Varre, A., and Flammarion, N. Why do we need weight decay in modern deep learning? *arXiv preprint arXiv:2310.04415*, 2023. 3

Bartlett, P., Foster, D. J., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 2017. 1

Blanc, G., Gupta, N., Valiant, G., and Valiant, P. Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process. In *Conference on learning theory*, pp. 483–513. PMLR, 2020. 3

Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, 1992. 1

Bubeck, S. et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015. 35

Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pp. 1305–1338. PMLR, 2020. 2

Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient descent on neural networks typically occurs at the edge of stability. *International Conference on Learning Representations*, 2021. 2

Dai, Y., Ahn, K., and Sra, S. The crucial role of normalization in sharpness-aware minimization. *Advances in Neural Information Processing Systems*, 2023. 3

Even, M., Pesme, S., Gunasekar, S., and Flammarion, N. (S)GD over diagonal linear networks: Implicit regularisation, large stepsizes and edge of stability. *arXiv preprint arXiv:2302.08982*, 2023. 3

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *International Conference on Learning Representations*, 2021. 8

Gatmiry, K., Li, Z., Chuang, C.-Y., Reddi, S., Ma, T., and Jegelka, S. The inductive bias of flatness regularization for deep matrix factorization. *arXiv preprint arXiv:2306.13239*, 2023. 3

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018a. 2

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, 31, 2018b. 2

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural computation*, 9(1):1–42, 1997. 2

Jastrzebski, S., Szymczak, M., Fort, S., Arpit, D., Tabor, J., Cho, K., and Geras, K. The break-even point on optimization trajectories of deep neural networks. In *International Conference on Learning Representations*, 2020. 2

Ji, Z. and Telgarsky, M. Gradient descent aligns the layers of deep linear networks. *International Conference on Learning Representations*, 2019a. 2

Ji, Z. and Telgarsky, M. Risk and parameter convergence of logistic regression. *Conference on Learning Theory*, 2019b. 2

Ji, Z. and Telgarsky, M. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020. 2, 8

Ji, Z. and Telgarsky, M. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pp. 772–804. PMLR, 2021. 1, 2, 3, 5, 6, 35

Ji, Z., Dudík, M., Schapire, R. E., and Telgarsky, M. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pp. 2109–2136. PMLR, 2020. 2, 16, 35

Ji, Z., Srebro, N., and Telgarsky, M. Fast margin maximization via dual acceleration. In *International Conference on Machine Learning*, pp. 4860–4869. PMLR, 2021. 1, 2, 6

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2019. 2

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2016. 2

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images, 2009. URL https://www.cs.toronto.edu/~kriz/cifar.html. 8

Kunin, D., Yamamura, A., Ma, C., and Ganguli, S. The asymmetric maximum margin bias of quasi-homogeneous neural networks. *International Conference on Learning Representations*, 2023. 2

Li, Z., Lyu, K., and Arora, S. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*, 33:14544–14555, 2020. 3

Li, Z., Wang, T., and Arora, S. What happens after SGD reaches zero loss?–a mathematical framework. *International Conference on Learning Representations*, 2022a. 3

Li, Z., Wang, T., Lee, J. D., and Arora, S. Implicit bias of gradient descent on reparametrized models: On equivalence to mirror descent. *Advances in Neural Information Processing Systems*, 35:34626–34640, 2022b. 3

Long, P. M. and Bartlett, P. L. Sharpness-aware minimization and the edge of stability. *arXiv preprint arXiv:2309.12488*, 2023. 3

Lyu, K. and Li, J. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019. 2, 8, 37

Lyu, K., Li, Z., Wang, R., and Arora, S. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34, 2021. 2

Lyu, K., Li, Z., and Arora, S. Understanding the generalization benefit of normalization layers: Sharpness reduction. *Advances in Neural Information Processing Systems*, 35: 34689–34708, 2022. 3

Ma, C. and Ying, L. On linear stability of SGD and input-smoothness of neural networks. *Advances in Neural Information Processing Systems*, 34:16805–16817, 2021. 2, 3

Ma, C., Kunin, D., Wu, L., and Ying, L. Beyond the quadratic approximation: The multiscale structure of neural network loss landscapes. *Journal of Machine Learning*, 1(3):247–267, 2022. 3

Mulayoff, R., Michaeli, T., and Soudry, D. The implicit bias of minima stability: A view from function space. *Advances in Neural Information Processing Systems*, 34: 17749–17761, 2021. 3

Nacson, M. S., Gunasekar, S., Lee, J., Srebro, N., and Soudry, D. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. In *International Conference on Machine Learning*, pp. 4683–4692. PMLR, 2019a. 2

Nacson, M. S., Lee, J., Gunasekar, S., Savarese, P. H. P., Srebro, N., and Soudry, D. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3420–3428. PMLR, 2019b. 1, 3, 6

Nacson, M. S., Srebro, N., and Soudry, D. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3051–3059. PMLR, 2019c. 2, 6

Nacson, M. S., Ravichandran, K., Srebro, N., and Soudry, D. Implicit bias of the step size in linear diagonal neural networks. In *International Conference on Machine Learning*, pp. 16270–16295. PMLR, 2022. 3

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014. 1

Pesme, S. and Flammarion, N. Saddle-to-saddle dynamics in diagonal linear networks. *Advances in Neural Information Processing Systems*, 2023. 3

Pesme, S., Pillaud-Vivien, L., and Flammarion, N. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021. 3

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015. 8

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018. 1, 2, 3, 6, 16, 35

Sun, H., Ahn, K., Thrampoulidis, C., and Azizan, N. Mirror descent maximizes generalized margin and can be implemented efficiently. *Advances in Neural Information Processing Systems*, 2022. 2

Vapnik, V. *The nature of statistical learning theory*. Springer science & business media, 1999. 3

Vardi, G. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6):86–93, 2023. 2

Vardi, G., Shamir, O., and Srebro, N. On margin maximization in linear and ReLU networks. *Advances in Neural Information Processing Systems*, 35:37024–37036, 2022. 2

Wang, B., Meng, Q., Zhang, H., Sun, R., Chen, W., and Ma, Z.-M. Momentum doesn't change the implicit bias. *Advances in Neural Information Processing Systems*, 2022a. 2, 6

Wang, G., Hanashiro, R., Guha, E., and Abernethy, J. On accelerated perceptrons and beyond. *arXiv preprint arXiv:2210.09371*, 2022b. 1, 2, 6

Wang, G., Hu, Z., Muthukumar, V., and Abernethy, J. Faster margin maximization rates for generic optimization methods. *Advances in Neural Information Processing Systems*, 2023. 2

Wang, M. and Ma, C. Understanding multi-phase optimization dynamics and rich nonlinear behaviors of ReLU networks. *Advances in Neural Information Processing Systems*, 2023. 2

Wang, M. and Wu, L. The noise geometry of stochastic gradient descent: A quantitative and analytical characterization. *arXiv preprint arXiv:2310.00692*, 2023. 3

Wen, K., Ma, T., and Li, Z. How sharpness-aware minimization minimizes sharpness? In *The Eleventh International Conference on Learning Representations*, 2023a. 3

Wen, K., Ma, T., and Li, Z. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. *Advances in Neural Information Processing Systems*, 2023b. 3

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020. 3

Wu, J., Braverman, V., and Lee, J. D. Implicit bias of gradient descent for logistic regression at the edge of stability. *Advances in Neural Information Processing Systems*, 2023. 2, 5, 6

Wu, L. and Su, W. J. The implicit regularization of dynamical stability in stochastic gradient descent. In *The 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37656–37684. PMLR, 2023. 3

Wu, L., Ma, C., and E, W. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31:8279–8288, 2018. 2

Wu, L., Wang, M., and Su, W. J. The alignment property of SGD noise and how it helps select flat minima: A stability analysis. *Advances in Neural Information Processing Systems*, 35:4680–4693, 2022. 2

Wu, X., Dobriban, E., Ren, T., Wu, S., Li, Z., Gunasekar, S., Ward, R., and Liu, Q. Implicit regularization and convergence for weight normalization. *Advances in Neural Information Processing Systems*, 33:2835–2847, 2020. 3

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. 1

Appendix

A. Proofs in Section 4

A.1. Proof of Proposition 4.1

Proof of Proposition 4.1.
Step I. Regularized path analysis.

For simplicity, we denote $\boldsymbol{z}_1 = \boldsymbol{x}_1 y_1$ and $\boldsymbol{z}_2 = \boldsymbol{x}_2 y_2$.

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{3}\left(2e^{-\boldsymbol{w}^\top \boldsymbol{z}_1} + e^{-\boldsymbol{w}^\top \boldsymbol{z}_2}\right) = \frac{1}{3}e^{-w_1\gamma}\left(2e^{-w_2\sqrt{1-\gamma^2}} + e^{w_2\sqrt{1-\gamma^2}}\right).$$

$$\nabla\mathcal{L}(\boldsymbol{w}) = \begin{pmatrix} -\frac{1}{3}e^{-w_1\gamma}\gamma\left(2e^{-w_2\sqrt{1-\gamma^2}} + e^{w_2\sqrt{1-\gamma^2}}\right) \\ \frac{1}{3}e^{-w_1\gamma}\sqrt{1-\gamma^2}\left(-2e^{-w_2\sqrt{1-\gamma^2}} + e^{w_2\sqrt{1-\gamma^2}}\right) \end{pmatrix}.$$

For any fixed $R > 0$, we will calculate the regularized solution in the ball $\|\boldsymbol{w}\|_2 \leq R$.

From the expression of $\nabla\mathcal{L}(\boldsymbol{w})$, we know $\nabla\mathcal{L}(\boldsymbol{w}) \neq \mathbf{0}$ for any $\boldsymbol{w} \in \mathbb{R}^d$. Hence, it must holds $\left\|\boldsymbol{w}^*_{\text{reg}}(R)\right\|_2 = R$. Moreover, we can determine the signal of $w^*_{\text{reg},1}(R)$ and $w^*_{\text{reg},2}(R)$. From the symmetry of the ℓ_2 ball, we know $w^*_{\text{reg},1}(R) < 0$ and $w^*_{\text{reg},2}(R) > 0$. This is because: if $w^*_{\text{reg},1}(R) > 0$, then $\mathcal{L}(-w^*_{\text{reg},1}(R), w^*_{\text{reg},2}(R)) < \mathcal{L}(w^*_{\text{reg},1}(R), w^*_{\text{reg},2}(R))$, which is contradict to the optimum of $\boldsymbol{w}^*_{\text{reg}}(R)$.

Then from the optimum and differentiability, we have

$$\frac{\left\langle \boldsymbol{w}^*_{\text{reg}}(R), -\nabla\mathcal{L}\left(\boldsymbol{w}^*_{\text{reg}}(R)\right)\right\rangle}{R\left\|\nabla\mathcal{L}\left(\boldsymbol{w}^*_{\text{reg}}(R)\right)\right\|_2} = 1,$$

which means

$$\boldsymbol{w}_{\mathrm{reg}}^*(R) \;/\!/\; \nabla\mathcal{L}\left(\boldsymbol{w}_{\mathrm{reg}}^*(R)\right), \quad \left\langle \boldsymbol{w}_{\mathrm{reg}}^*(R), \nabla\mathcal{L}\left(\boldsymbol{w}_{\mathrm{reg}}^*(R)\right)\right\rangle < 0.$$

For simplicity, we use the notation $w_1(R) := w_{\mathrm{reg},1}^*(R)$, $w_2(R) := w_{\mathrm{reg},2}^*(R)$ in the proof below.

By a straightforward calculation and taking the square, we have

$$\frac{(1-\gamma^2)\left(e^{2w_2(B)\sqrt{1-\gamma^2}} + 4e^{-2w_2(B)\sqrt{1-\gamma^2}} - 4\right)}{\gamma^2\left(e^{2w_2(R)\sqrt{1-\gamma^2}} + 4e^{-2w_2\sqrt{1-\gamma^2}} + 4\right)} = \frac{w_2^2(B)}{w_1^2(R)} = \frac{w_2^2(R)}{R^2 - w_2^2(R)},$$

which is equivalent to

$$\frac{R^2}{w_2^2(R)} = \frac{1}{1-\gamma^2} + \frac{8\gamma^2}{(1-\gamma^2)\left(e^{2w_2(R)\sqrt{1-\gamma^2}} + 4e^{-2w_2(R)\sqrt{1-\gamma^2}} - 4\right)}. \tag{7}$$

With the help of Lemma A.2, we know

$$\lim_{R\to\infty}\left\langle \boldsymbol{w}^*, \frac{\boldsymbol{w}_{\mathrm{reg}}^*(R)}{R}\right\rangle = \lim_{R\to\infty}\frac{w_1(R)}{\sqrt{w_1^2(R) + w_2^2(R)}} = 1,$$

which means $\lim\limits_{R\to\infty}\frac{w_2^2(R)}{R^2} = 0$. Then taking $R\to\infty$ in (7), we have

$$\lim_{R\to\infty}\left(e^{2w_2(R)\sqrt{1-\gamma^2}} + 4e^{-2w_2(R)\sqrt{1-\gamma^2}}\right) = 4.$$

A straight-forward calculation gives us

$$\lim_{R\to\infty} w_2(R) = \frac{\log 2}{2\sqrt{1-\gamma^2}}.$$

Step II. Proof for NGD.

Following the proof, we have

$$-\frac{\nabla\mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})} = \begin{pmatrix}\gamma \\ \sqrt{1-\gamma^2}\left(2 - e^{2w_2\sqrt{1-\gamma^2}}\right) / \left(2 + e^{2w_2\sqrt{1-\gamma^2}}\right)\end{pmatrix}.$$

For NGD, it holds that:

$$\begin{aligned} w_1(t+1) &= w_1(t) + \gamma, \\ w_2(t+1) &= w_2(t) + \sqrt{1-\gamma^2}\left(2 - e^{2w_2(t)\sqrt{1-\gamma^2}}\right) / \left(2 + e^{2w_2(t)\sqrt{1-\gamma^2}}\right). \end{aligned}$$

It is worth noticing that the dynamics of $w_1(t)$ and $w_2(t)$ are decoupled. For $w_1(t)$, it is easy to verify that $w_1(t) = \gamma t$, $\forall t \geq 1$. As for $w_2(t)$, we will estimate the uniform upper and lower bounds.

For simplicity, we denote $x(t) := 2w_2(t)\sqrt{1-\gamma^2} - \log 2$. From the dynamics of $w_2(t)$, the dynamics of $x(t)$ are

$$x(t+1) = x(t) + 2(1-\gamma^2)\frac{1 - e^{x(t)}}{1 + e^{x(t)}} = x(t) + 2(1-\gamma^2)\left(\frac{2}{1 + e^{x(t)}} - 1\right).$$

Then we will prove that $|x(t)| \leq \frac{1}{2}\log 2$ holds for $t \geq 1$ by induction.

From $x(0) = -\log 2$, we have $x(1) = -\log 2 + \frac{2(1-\gamma^2)}{3} \in \left[-\frac{1}{2}\log 2, \frac{1}{2}\log 2\right]$.

Assume that $x(t) \in \left[-\frac{1}{2}\log 2, \frac{1}{2}\log 2\right]$ holds for any $t \leq k$, and we denote $h(x) := x + 2(1-\gamma^2)\left(\frac{2}{1+e^x} - 1\right)$. Then with the help of Lemma D.2, the following estimate holds for $t = k+1$:

$$x(k+1) = h\left(x(k)\right) \leq h\left(\frac{1}{2}\log 2\right) = \frac{1}{2}\log 2 + 2(1-\gamma^2)\frac{1-\sqrt{2}}{1+\sqrt{2}} < \frac{1}{2}\log 2;$$
$$x(k+1) = h\left(x(k)\right) \geq h\left(-\frac{1}{2}\log 2\right) = -\frac{1}{2}\log 2 + 2(1-\gamma^2)\frac{\sqrt{2}-1}{\sqrt{2}+1} > -\frac{1}{2}\log 2.$$

By induction, we have proved that $x(t) \in \left[-\frac{1}{2}\log 2, \frac{1}{2}\log 2\right]$ holds for any $t \geq 1$. This implies that $w_2(t) \in \left[\frac{\log 2}{4\sqrt{1-\gamma^2}}, \frac{3\log 2}{4\sqrt{1-\gamma^2}}\right]$ holds for any $t \geq 1$. Hence,

$$\frac{\log 2}{4t\gamma\sqrt{1-\gamma^2}} \leq \frac{w_2(t)}{w_1(t)} \leq \frac{3\log 2}{4t\gamma\sqrt{1-\gamma^2}},\ \forall t \geq 1.$$

From the definition of directional convergence, we have:

$$\left\|\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|} - \boldsymbol{w}^\star\right\| = \sqrt{2\left(1 - \left\langle\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|}, \boldsymbol{e}_1\right\rangle\right)} = \sqrt{2\left(1 - \frac{w_1(t)}{\sqrt{w_1^2(t) + w_2^2(t)}}\right)}$$
$$= \sqrt{2\left(1 - \frac{1}{\sqrt{\frac{w_2^2(t)}{w_1^2(t)} + 1}}\right)} = \Theta\left(\left|\frac{w_2(t)}{w_1(t)}\right|\right) = \Theta\left(\frac{1}{t}\right).$$

From the definition of margin, we have:

$$\gamma(\boldsymbol{w}(t)) - \gamma^\star = \min_{i\in[2]}\left\langle\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|}, \boldsymbol{z}_i\right\rangle - \gamma^\star = \left\langle\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|}, \boldsymbol{z}_2\right\rangle - \gamma^\star$$
$$= \frac{w_1(t)\gamma^\star - w_2(t)\sqrt{1-\gamma^{\star 2}}}{\sqrt{w_1^2(t) + w_2^2(t)}} - \gamma^\star = \frac{-\frac{w_2(t)}{w_1(t)}\sqrt{1-\gamma^{\star 2}} + \gamma^\star}{\sqrt{\frac{w_2^2(t)}{w_1^2(t)} + 1}} - \gamma^\star$$
$$= -\frac{\frac{w_2(t)}{w_1(t)}\sqrt{1-\gamma^{\star 2}}}{\sqrt{\frac{w_2^2(t)}{w_1^2(t)} + 1}} + \gamma^\star\left(\left(\frac{w_2^2(t)}{w_1^2(t)} + 1\right)^{-1/2} - 1\right) = -\Theta\left(\frac{w_2(t)}{w_1(t)}\right) - \Theta\left(\frac{w_2^2(t)}{w_1^2(t)}\right)$$
$$= -\Theta\left(\frac{w_2(t)}{w_1(t)}\right) = -\Theta\left(\frac{1}{t}\right).$$

Step II. Proof for PRGD.

For PRGD, to maximize margin exponentially fast, we only need to select $R_k = e^{\Theta(k)}$ and $T_k = \Theta(k)$. Notice that the choices of R_k and T_k are not unique. For simplicity, we use the following choice to make our proof clear.

- Phase I. We run NGD with $\eta = 1$ for one step.
- Phase II. We run PRGD with $\eta = 1$ for $t \geq 1$. Specifically, we select T_k and R_k such that:

$$T_0 = 1; \quad T_{k+1} = T_k + 2,\ \forall k \geq 0; \quad R_k = \frac{\|\boldsymbol{w}(T_k)\|}{w_2(T_k)},\ \forall k \geq 0.$$

Recalling our proof for NGD, at the end of Phase I, it holds that $w_2(1) \in \left[\frac{\log 2}{4\sqrt{1-\gamma^{\star 2}}}, \frac{3\log 2}{4\sqrt{1-\gamma^{\star 2}}}\right]$ and $w_1(1) = \gamma^\star$. Then we analyze Phase II. For simplicity, we denote an abosulte constant $q = 1 + \sqrt{1-\gamma^{\star 2}}\left(2 - e^{2\sqrt{1-\gamma^{\star 2}}}\right) / \left(2 + e^{2\sqrt{1-\gamma^{\star 2}}}\right) \in (0, 1)$.

- (S1). $w_2(2k+2)=1$ and $\|\boldsymbol{w}(2k+2)\|=R_k$ hold for any $k\geq 0$;
- (S2). $\frac{w_2(2k+2)}{w_1(2k+2)}=\frac{w_2(2k+1)}{w_2(2k+1)}$ holds for any $k\geq 0$;
- (S3). $w_1(2k+2)=\left(\frac{1}{q}\right)^k\left(w_1(2)+\frac{\gamma^\star}{1-q}\right)-\frac{\gamma^\star}{1-q}=e^{\Theta(k)}$.
- (S4). $R_k=e^{\Theta(k)}$.
- (S5). $\frac{w_2(t)}{w_1(t)}=e^{-\Theta(t)}$.

According to the update rule of Algorithm 1, (S1)(S2) hold directly.

Then we will prove (S3). Recalling the update rule, for $2k+3$, it holds that

$$\begin{aligned}&w_2(2k+3)=w_2(2k+2)+\sqrt{1-\gamma^{\star 2}}\left(2-e^{2w_2(2k+2)\sqrt{1-\gamma^{\star 2}}}\right)/\left(2+e^{2w_2(2k+2)\sqrt{1-\gamma^{\star 2}}}\right)\\=&1+\sqrt{1-\gamma^{\star 2}}\left(2-e^{2\sqrt{1-\gamma^{\star 2}}}\right)/\left(2+e^{2\sqrt{1-\gamma^{\star 2}}}\right):=q;\end{aligned}$$

$$w_1(2k+3)=w_1(2k+2)+\gamma^\star.$$

Recalling (S2), we have:

$$w_1(2k+4)=w_2(2k+4)\frac{w_1(2k+3)}{w_2(2k+3)}=1\cdot\frac{w_1(2k+2)+\gamma^\star}{q}=\frac{1}{q}\left(w_1(2k+2)+\gamma^\star\right),$$

where $1+\sqrt{1-\gamma^{\star 2}}\left(2-e^{2\sqrt{1-\gamma^{\star 2}}}\right)/\left(2+e^{2\sqrt{1-\gamma^{\star 2}}}\right)\in(0,1)$.

Consequently, a simple calculation can imply (S3):

$$w_1(2k+2)=\left(\frac{1}{q}\right)^k\left(w_1(2)+\frac{\gamma^\star}{1-q}\right)-\frac{\gamma^\star}{1-q}=e^{\Theta(k)},\quad\forall k\geq 0.$$

Then using (S1) and (S3), (S4) holds:

$$R_k=\|\boldsymbol{w}(2k+2)\|=\sqrt{1+w_1^2(2k+2)}=e^{\Theta(k)}.$$

By (S1)(S2) and (S3), we have:

$$\frac{w_2(2k+1)}{w_1(2k+1)}=\frac{w_2(2k+2)}{w_1(2k+2)}=\frac{1}{w_1(2k+2)}=e^{-\Theta(2k+2)},$$

which implies (S4).

From the definition of directional convergence, we have:

$$\begin{aligned}&\left\|\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|}-\boldsymbol{w}^\star\right\|=\sqrt{2\left(1-\left\langle\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|},\boldsymbol{e}_1\right\rangle\right)}=\sqrt{2\left(1-\frac{w_1(t)}{\sqrt{w_1^2(t)+w_2^2(t)}}\right)}\\=&\sqrt{2\left(1-\frac{1}{\sqrt{\frac{w_2^2(t)}{w_1^2(t)}+1}}\right)}=\Theta\left(\frac{w_2(t)}{w_1(t)}\right)=e^{-\Theta(t)}.\end{aligned}$$

From the definition of margin, we have:

$$\gamma(\boldsymbol{w}(t))-\gamma^\star=\min_{i\in[2]}\left\langle\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|},\boldsymbol{z}_i\right\rangle-\gamma^\star=\left\langle\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|},\boldsymbol{z}_2\right\rangle-\gamma^\star$$

$$
\begin{aligned}
=&\frac{w_1(t)\gamma^\star - w_2(t)\sqrt{1-\gamma^{\star 2}}}{\sqrt{w_1^2(t)+w_2^2(t)}} - \gamma^\star = \frac{-\frac{w_2(t)}{w_1(t)}\sqrt{1-\gamma^{\star 2}}+\gamma^\star}{\sqrt{\frac{w_2^2(t)}{w_1^2(t)}+1}} - \gamma^\star \\
=&-\frac{\frac{w_2(t)}{w_1(t)}\sqrt{1-\gamma^{\star 2}}}{\sqrt{\frac{w_2^2(t)}{w_1^2(t)}+1}} + \gamma^\star\left(\left(\frac{w_2^2(t)}{w_1^2(t)}+1\right)^{-1/2}-1\right) = -\Theta\left(\frac{w_2(t)}{w_1(t)}\right) - \Theta\left(\frac{w_2^2(t)}{w_1^2(t)}\right) \\
=&-\Theta\left(\frac{w_2(t)}{w_1(t)}\right) = -e^{-\Theta(t)}.
\end{aligned}
$$

□

A.2. Useful Lemmas

Lemma A.1 (Margin error and Directional error)**.** *Under Assumption 3.1, for any $\boldsymbol{w}\in\mathbb{R}^d$, it holds that $\gamma^\star - \gamma(\boldsymbol{w}) \leq \left\|\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}-\boldsymbol{w}^\star\right\|$.*

Proof of Lemma A.1.
Let $\boldsymbol{w}\in\mathbb{R}^d$ and denote $i_0\in\arg\min_{i\in[n]} y_i\left\langle\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|},\boldsymbol{x}_i\right\rangle$. Then we have:

$$
\begin{aligned}
&\gamma^\star-\gamma(\boldsymbol{w}) = \min_i y_i\left\langle\boldsymbol{w}^\star,\boldsymbol{x}_i\right\rangle - \min_i y_i\left\langle\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|},\boldsymbol{x}_i\right\rangle \\
=&\min_i y_i\left\langle\boldsymbol{w}^\star,\boldsymbol{x}_i\right\rangle - y_{i_0}\left\langle\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|},\boldsymbol{x}_{i_0}\right\rangle \\
\leq& y_{i_0}\left\langle\boldsymbol{w}^\star,\boldsymbol{x}_{i_0}\right\rangle - y_{i_0}\left\langle\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|},\boldsymbol{x}_{i_0}\right\rangle = y_{i_0}\left\langle\boldsymbol{w}^\star-\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|},\boldsymbol{x}_{i_0}\right\rangle \\
\leq&\left\|\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}-\boldsymbol{w}^\star\right\|.
\end{aligned}
$$

Lemma A.2 (Integration of (Soudry et al., 2018; Ji et al., 2020))**.** *For problem (2), Gradient Flow convergences to the ℓ_2 max-margin direction $\boldsymbol{w}^*$, hence the regularization path also convergences to the ℓ_2 max-margin solution: $\lim_{B\to\infty}\frac{\boldsymbol{w}^*_{\rm reg}(B)}{B}=\boldsymbol{w}^*$.*

□

B. Proofs in Section 5

Proof of Theorem 5.5.
Without loss of generality, we can assume $\mathrm{span}\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}=\mathbb{R}^d$. If $\mathrm{span}\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}\neq\mathbb{R}^d$, we only need to change the proof in the subspace $\mathrm{span}\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$.

Recall the definition of $\mathbb{C}(D_1,D_2;H)$:

$$
\mathbb{C}(D_1,D_2;H):=\left\{\boldsymbol{w}\in\mathrm{span}\{\boldsymbol{x}_i:i\in[n]\}:D_1\leq\|\mathcal{P}_\perp(\boldsymbol{w})\|\leq D_2;\langle\boldsymbol{w},\boldsymbol{w}^\star\rangle\geq H\right\},
$$

we further define $\mathbb{C}(D;H)$ as:

$$
\mathbb{C}(D;H):=\left\{\boldsymbol{w}\in\mathrm{span}\{\boldsymbol{x}_i:i\in[n]\}:\|\mathcal{P}_\perp(\boldsymbol{w})\|=D;\langle\boldsymbol{w},\boldsymbol{w}^\star\rangle\geq H\right\}.
$$

It holds that

$$
\mathbb{C}(D;H)=\left\{h\boldsymbol{w}^\star+D\boldsymbol{v}:h\geq H,\boldsymbol{v}\perp\boldsymbol{w}^\star,\|\boldsymbol{v}\|=1\right\},
$$

and the following relationship is easy to verified:

$$\mathbb{C}(D_1, D_2; H) = \bigcup_{D_1 \le D \le D_2} \mathbb{C}(D; H).$$

In the following steps, we first prove the lower bound for $\mathbb{C}(D; H)$, and then prove for $\mathbb{C}(D_1, D_2; H)$.

Step I. Strip out the important ingredients.

For any $\boldsymbol{w} \in \mathbb{C}(D; H)$, we have:

$$\begin{aligned}
&\left\langle \frac{\nabla \mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}, \frac{\mathcal{P}_\perp(\boldsymbol{w})}{\|\mathcal{P}_\perp(\boldsymbol{w})\|} \right\rangle = \left\langle \frac{\nabla \mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}, \boldsymbol{v} \right\rangle \\
=& \left\langle \frac{\frac{1}{n}\sum_{i=1}^n (-y_i \boldsymbol{x}_i) \exp(-\langle \boldsymbol{w}, y_i \boldsymbol{x}_i \rangle)}{\frac{1}{n}\sum_{i=1}^n \exp(-y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)}, \boldsymbol{v} \right\rangle \\
=& \frac{\sum_{i=1}^n \langle \boldsymbol{v}, -y_i \boldsymbol{x}_i \rangle \exp\left(-h \langle \boldsymbol{w}^\star, y_i \boldsymbol{x}_i \rangle\right) \exp\left(-D \langle \boldsymbol{v}, y_i \boldsymbol{x}_i \rangle\right)}{\sum_{i=1}^n \exp\left(-h \langle \boldsymbol{w}^\star, y_i \boldsymbol{x}_i \rangle\right) \exp\left(-D \langle \boldsymbol{v}, y_i \boldsymbol{x}_i \rangle\right)}.
\end{aligned} \tag{8}$$

For the numerator of (8), it holds that

$$\begin{aligned}
&\sum_{i=1}^n \langle \boldsymbol{v}, -y_i \boldsymbol{x}_i \rangle \exp\left(-h \langle \boldsymbol{w}^\star, y_i \boldsymbol{x}_i \rangle\right) \exp\left(-D \langle \boldsymbol{v}, y_i \boldsymbol{x}_i \rangle\right) \\
=& \sum_{i \in \mathcal{I}} \langle \boldsymbol{v}, -y_i \boldsymbol{x}_i \rangle \exp\left(-h \langle \boldsymbol{w}^\star, y_i \boldsymbol{x}_i \rangle\right) \exp\left(-D \langle \boldsymbol{v}, y_i \boldsymbol{x}_i \rangle\right) \\
&+ \sum_{i \notin \mathcal{I}} \langle \boldsymbol{v}, -y_i \boldsymbol{x}_i \rangle \exp\left(-h \langle \boldsymbol{w}^\star, y_i \boldsymbol{x}_i \rangle\right) \exp\left(-D \langle \boldsymbol{v}, y_i \boldsymbol{x}_i \rangle\right) \\
=& \exp\left(-h\gamma^\star\right) \sum_{i \in \mathcal{I}} \langle \boldsymbol{v}, -y_i \boldsymbol{x}_i \rangle \exp\left(-D \langle \boldsymbol{v}, y_i \boldsymbol{x}_i \rangle\right) \\
&+ \sum_{i \notin \mathcal{I}} \langle \boldsymbol{v}, -y_i \boldsymbol{x}_i \rangle \exp\left(-h \langle \boldsymbol{w}^\star, y_i \boldsymbol{x}_i \rangle\right) \exp\left(-D \langle \boldsymbol{v}, y_i \boldsymbol{x}_i \rangle\right) \\
\ge& \exp\left(-h\gamma^\star\right) \sum_{i \in \mathcal{I}} \langle \boldsymbol{v}, -y_i \boldsymbol{x}_i \rangle \exp\left(-D \langle \boldsymbol{v}, y_i \boldsymbol{x}_i \rangle\right) - \sum_{i \notin \mathcal{I}} \exp\left(-h \langle \boldsymbol{w}^\star, y_i \boldsymbol{x}_i \rangle\right) \exp\left(D\right) \\
\ge& \exp\left(-h\gamma^\star\right) \sum_{i \in \mathcal{I}} \langle \boldsymbol{v}, -y_i \boldsymbol{x}_i \rangle \exp\left(-D \langle \boldsymbol{v}, y_i \boldsymbol{x}_i \rangle\right) - (n - |\mathcal{I}|) \exp\left(-h\gamma^\star_{\text{sub}}\right) \exp\left(D\right);
\end{aligned}$$

For the denominator of (8), it holds that:

$$\begin{aligned}
&\sum_{i=1}^n \exp\left(-h \langle \boldsymbol{w}^\star, y_i \boldsymbol{x}_i \rangle\right) \exp\left(-D \langle \boldsymbol{v}, y_i \boldsymbol{x}_i \rangle\right) \\
\le& \sum_{i=1}^n \exp\left(-h \langle \boldsymbol{w}^\star, y_i \boldsymbol{x}_i \rangle\right) \exp\left(D\right) \\
=& \left(\sum_{i \in \mathcal{I}} \exp\left(-h \langle \boldsymbol{w}^\star, y_i \boldsymbol{x}_i \rangle\right) + \sum_{i \notin \mathcal{I}} \exp\left(-h \langle \boldsymbol{w}^\star, y_i \boldsymbol{x}_i \rangle\right) \right) \exp\left(D\right) \\
\le& \left(|\mathcal{I}| \exp\left(-h\gamma^\star\right) + (n - |\mathcal{I}|) \exp\left(-h\gamma^\star_{\text{sub}}\right) \right) \exp(D).
\end{aligned}$$

Combining these two estimates, we obtain:

$$\begin{aligned}
&\left\langle \frac{\nabla \mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}, \frac{\mathcal{P}_\perp(\boldsymbol{w})}{\|\mathcal{P}_\perp(\boldsymbol{w})\|} \right\rangle \\
\geq & \frac{\exp\left(-h\gamma^\star\right) \sum_{i\in\mathcal{I}} \left\langle \boldsymbol{v}, -y_i\boldsymbol{x}_i \right\rangle \exp\left(-D\left\langle \boldsymbol{v}, y_i\boldsymbol{x}_i\right\rangle\right) - (n-|\mathcal{I}|)\exp\left(-h\gamma^\star_{\text{sub}}\right)\exp\left(D\right)}{\left(|\mathcal{I}|\exp\left(-h\gamma^\star\right) + (n-|\mathcal{I}|)\exp\left(-h\gamma^\star_{\text{sub}}\right)\right)\exp(D)} \\
= & \frac{\frac{1}{|\mathcal{I}|\exp(D)} \sum_{i\in\mathcal{I}} \left\langle \boldsymbol{v}, -y_i\boldsymbol{x}_i \right\rangle \exp\left(-D\left\langle \boldsymbol{v}, y_i\boldsymbol{x}_i\right\rangle\right) - \frac{n-|\mathcal{I}|}{|\mathcal{I}|}\exp\left(-h(\gamma^\star_{\text{sub}}-\gamma^\star)\right)}{1 + \frac{n-|\mathcal{I}|}{|\mathcal{I}|}\exp\left(-h(\gamma^\star_{\text{sub}}-\gamma^\star)\right)}.
\end{aligned} \tag{9}$$

Notice that the term $\frac{n-|\mathcal{I}|}{|\mathcal{I}|}\exp\left(-h(\gamma^\star_{\text{sub}}-\gamma^\star)\right)$ converges to 0 when h goes to $+\infty$. Thus, we only need to derive the uniform lower bound for the term

$$\sum_{i\in\mathcal{I}} \left\langle \boldsymbol{v}, -y_i\boldsymbol{x}_i \right\rangle \exp\left(-D\left\langle \boldsymbol{v}, y_i\boldsymbol{x}_i\right\rangle\right)$$

for any $\boldsymbol{v} \in \{\boldsymbol{v} : \boldsymbol{v} \perp \boldsymbol{w}^\star, \|\boldsymbol{v}\| = 1\}$.

Step II. Uniform Lower bound of $\sum_{i\in\mathcal{I}} \left\langle \boldsymbol{v}, -y_i\boldsymbol{x}_i \right\rangle \exp\left(D\left\langle \boldsymbol{v}, -y_i\boldsymbol{x}_i\right\rangle\right)$ for $\{\boldsymbol{v} : \boldsymbol{v} \perp \boldsymbol{w}^\star, \|\boldsymbol{v}\| = 1\}$.

For simplicity, we denote $\boldsymbol{u}_i := \mathcal{P}_\perp(-y_i\boldsymbol{x}_i)$ for $i \in [n]$. It is worth noticing that $\langle \boldsymbol{v}, -y_i\boldsymbol{x}_i\rangle = \langle \boldsymbol{v}, \boldsymbol{u}_i\rangle$ for any $\boldsymbol{v} \in \{\boldsymbol{v} : \boldsymbol{v} \perp \boldsymbol{w}^\star, \|\boldsymbol{v}\| = 1\}$. Therefore,

$$\sum_{i\in\mathcal{I}} \left\langle \boldsymbol{v}, -y_i\boldsymbol{x}_i \right\rangle \exp\left(D\left\langle \boldsymbol{v}, -y_i\boldsymbol{x}_i\right\rangle\right) = \sum_{i\in\mathcal{I}} \left\langle \boldsymbol{v}, \boldsymbol{u}_i \right\rangle \exp\left(D\left\langle \boldsymbol{v}, \boldsymbol{u}_i\right\rangle\right).$$

First, recalling Assumption 5.4, there exist $\alpha_i > 0$ $(i \in \mathcal{I})$ such that $\boldsymbol{w}^\star = \sum_{i\in\mathcal{I}} \alpha_i y_i \boldsymbol{x}_i$, where $\sum_{i\in\mathcal{I}} \alpha_i = 1$. This implies: $\mathbf{0} = \sum_{i\in\mathcal{I}} \alpha_i \boldsymbol{u}_i$. Thus, we define $\boldsymbol{k}_i := \alpha_i \boldsymbol{u}_i$, then

$$\sum_{i\in\mathcal{I}} \boldsymbol{k}_i = \mathbf{0}.$$

Recalling Assumption 5.4, it holds $\text{span}\{\boldsymbol{x}_i : i \in \mathcal{I}\} = \text{span}\{\boldsymbol{x}_i : i \in [n]\}$, which implies that $\text{span}\{\boldsymbol{u}_i : i \in \mathcal{I}\} = \{\boldsymbol{v} : \boldsymbol{v} \perp \boldsymbol{w}^\star\}$. Therefore, there exists an absolute constant $\lambda_{\min} > 0$ such that

$$\sum_{i\in\mathcal{I}} \langle \boldsymbol{v}, \boldsymbol{k}_i\rangle^2 = \boldsymbol{v}^\top \left(\sum_{i\in\mathcal{I}} \boldsymbol{k}_i\boldsymbol{k}_i^\top\right)\boldsymbol{v} \geq \lambda_{\min}, \quad \forall \boldsymbol{v} \in \{\boldsymbol{v} \perp \boldsymbol{w}^\star, \|\boldsymbol{v}\| = 1\}.$$

By a rough estimate, it holds that:

$$\begin{aligned}
&\sum_{i\in\mathcal{I}} |\langle \boldsymbol{v}, \boldsymbol{k}_i\rangle| = \left(\left(\sum_{i\in\mathcal{I}} |\langle \boldsymbol{v}, \boldsymbol{k}_i\rangle|\right)^2\right)^{1/2} \\
\geq & \left(\sum_{i\in\mathcal{I}} |\langle \boldsymbol{v}, \boldsymbol{k}_i\rangle|^2\right)^{1/2} \geq \sqrt{\lambda_{\min}}, \quad \forall \boldsymbol{v} \in \{\boldsymbol{v} \perp \boldsymbol{w}^\star, \|\boldsymbol{v}\| = 1\}.
\end{aligned}$$

Notice that $\sum_{i\in\mathcal{I}} \langle \boldsymbol{v}, \boldsymbol{k}_i\rangle = \left\langle \boldsymbol{v}, \sum_{i\in\mathcal{I}} \boldsymbol{k}_i\right\rangle = 0$, it holds that

$$\sum_{i\in\mathcal{I}, \langle \boldsymbol{v}, \boldsymbol{k}_i\rangle > 0} \langle \boldsymbol{v}, \boldsymbol{k}_i\rangle = -\sum_{i\in\mathcal{I}, \langle \boldsymbol{v}, \boldsymbol{k}_i\rangle < 0} \langle \boldsymbol{v}, \boldsymbol{k}_i\rangle,$$

which means:

$$\sum_{i\in\mathcal{I},\langle \boldsymbol{v},\boldsymbol{k}_i\rangle>0}\langle \boldsymbol{v},\boldsymbol{k}_i\rangle=\frac{1}{2}\sum_{i\in\mathcal{I}}|\langle \boldsymbol{v},\boldsymbol{k}_i\rangle|.$$

Consequently, we can do the following estimate for the largest $\langle \boldsymbol{v},\boldsymbol{k}_i\rangle$:

$$\begin{aligned}&\max_{i\in\mathcal{I}}\langle \boldsymbol{v},\boldsymbol{k}_i\rangle\geq\frac{1}{|\{i:i\in\mathcal{I},\langle \boldsymbol{v},\boldsymbol{k}_i\rangle>0\}|}\sum_{i\in\mathcal{I},\langle \boldsymbol{v},\boldsymbol{k}_i\rangle>0}\langle \boldsymbol{v},\boldsymbol{k}_i\rangle\\ \geq&\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I},\langle \boldsymbol{v},\boldsymbol{k}_i\rangle>0}\langle \boldsymbol{v},\boldsymbol{k}_i\rangle=\frac{1}{2|\mathcal{I}|}\sum_{i\in\mathcal{I}}|\langle \boldsymbol{v},\boldsymbol{k}_i\rangle|\geq\frac{\sqrt{\lambda_{\min}}}{2|\mathcal{I}|},\quad\forall\boldsymbol{v}\in\{\boldsymbol{v}\perp\boldsymbol{w}^\star,\|\boldsymbol{v}\|=1\}.\end{aligned}$$

Hence, we obtain the uniform lower bound by the following splitting:

$$\begin{aligned}&\sum_{i\in\mathcal{I}}\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\exp\left(D\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\right)\\ =&\sum_{i\in\mathcal{I}:\langle \boldsymbol{v},\boldsymbol{u}_i\rangle>0}\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\exp\left(D\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\right)+\sum_{i\in\mathcal{I}:\langle \boldsymbol{v},\boldsymbol{u}_i\rangle<0}\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\exp\left(D\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\right)\\ \geq&\max_{i\in\mathcal{I}}\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\exp\left(D\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\right)+\sum_{i\in\mathcal{I}:\langle \boldsymbol{v},\boldsymbol{u}_i\rangle<0}\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\exp\left(D\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\right)\\ \geq&\max_{i\in\mathcal{I}}\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\exp\left(D\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\right)+\sum_{i\in\mathcal{I}:\langle \boldsymbol{v},\boldsymbol{u}_i\rangle<0}\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\exp\left(0\right)\\ \geq&\max_{i\in\mathcal{I}}\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\exp\left(D\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\right)-\sum_{i\in\mathcal{I}:\langle \boldsymbol{v},\boldsymbol{u}_i\rangle<0}\|\boldsymbol{u}_i\|\\ \geq&\max_{i\in\mathcal{I}}\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\exp\left(D\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\right)-|\mathcal{I}|\\ \geq&\frac{\sqrt{\lambda_{\min}}}{2|\mathcal{I}|}\exp\left(D\frac{\sqrt{\lambda_{\min}}}{2|\mathcal{I}|}\right)-|\mathcal{I}|,\quad\forall\boldsymbol{v}\in\{\boldsymbol{v}\perp\boldsymbol{w}^\star,\|\boldsymbol{v}\|=1\}.\end{aligned}$$

Step III. The final bound.

First, we select

$$D_0=\log\left(\frac{4|\mathcal{I}|^2}{\sqrt{\lambda_{\min}}}\right),\quad H_0=\frac{1}{\gamma^\star_{\rm sub}-\gamma^\star}\log\left(\max\left\{\frac{2(n-|\mathcal{I}|)}{|\mathcal{I}|},2\right\}\right).$$

For any $D\geq D_0$ and $h\geq H_0$, we have: it holds that:

$$\begin{aligned}&\inf_{\boldsymbol{v}\in\{\boldsymbol{v}\perp\boldsymbol{w}^\star,\|\boldsymbol{v}\|=1\}}\sum_{i\in\mathcal{I}}\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\exp\left(D\langle \boldsymbol{v},\boldsymbol{u}_i\rangle\right)\\ \geq&\frac{\sqrt{\lambda_{\min}}}{2|\mathcal{I}|}\exp\left(D\frac{\sqrt{\lambda_{\min}}}{2|\mathcal{I}|}\right)-|\mathcal{I}|\geq\frac{\sqrt{\lambda_{\min}}}{4|\mathcal{I}|}\exp\left(D\frac{\sqrt{\lambda_{\min}}}{2|\mathcal{I}|}\right);\end{aligned}$$

$$\frac{n-|\mathcal{I}|}{|\mathcal{I}|}\exp\left(-h(\gamma^\star_{\rm sub}-\gamma^\star)\right)\leq\frac{1}{2}.$$

Therefore,

$$(9)\geq\frac{\frac{1}{|\mathcal{I}|\exp(D)}\frac{\sqrt{\lambda_{\min}}}{4|\mathcal{I}|}\exp\left(D\frac{\sqrt{\lambda_{\min}}}{2|\mathcal{I}|}\right)-\frac{n-|\mathcal{I}|}{|\mathcal{I}|}\exp\left(-h(\gamma^\star_{\rm sub}-\gamma^\star)\right)}{1+\frac{1}{2}}$$

$$\geq \frac{\frac{\sqrt{\lambda_{\min}}}{4|\mathcal{I}|^2}\exp(-D) - \frac{n-|\mathcal{I}|}{|\mathcal{I}|}\exp\left(-h(\gamma_{\text{sub}}^\star - \gamma^\star)\right)}{2}.$$

Now we select

$$D_1 = D_0, \quad D_2 = 2D_0, \quad H = \max\left\{H_0, \frac{1}{\gamma_{\text{sub}}^\star - \gamma^\star}\left(2D_0 + \log\left(\frac{4|\mathcal{I}|(n-|\mathcal{I}|)}{\sqrt{\lambda_{\min}}}\right)\right)\right\}.$$

Thus, for any $D \in [D_1, D_2]$ and $\boldsymbol{w} \in \mathbb{C}(D; H)$, it holds that

$$\begin{aligned}
&\left\langle \frac{\nabla\mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}, \frac{\mathcal{P}_\perp(\boldsymbol{w})}{\|\mathcal{P}_\perp(\boldsymbol{w})\|} \right\rangle \\
\geq &\frac{\frac{\sqrt{\lambda_{\min}}}{4|\mathcal{I}|^2}\exp(-D) - \frac{n-|\mathcal{I}|}{|\mathcal{I}|}\exp\left(-h(\gamma_{\text{sub}}^\star - \gamma^\star)\right)}{2} \\
\geq &\frac{\sqrt{\lambda_{\min}}\exp(-D)}{16|\mathcal{I}|^2} \geq \frac{\sqrt{\lambda_{\min}}\exp(-2D_0)}{16|\mathcal{I}|^2}.
\end{aligned}$$

Lastly, we obtain our result:

$$\begin{aligned}
&\inf_{\boldsymbol{w}\in\mathbb{C}(D_1,D_2;H)} \left\langle \frac{\nabla\mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}, \frac{\mathcal{P}_\perp(\boldsymbol{w})}{\|\mathcal{P}_\perp(\boldsymbol{w})\|} \right\rangle \\
= &\inf_{D\in[D_1,D_2]} \inf_{\boldsymbol{w}\in\mathbb{C}(D;H)} \left\langle \frac{\nabla\mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}, \frac{\mathcal{P}_\perp(\boldsymbol{w})}{\|\mathcal{P}_\perp(\boldsymbol{w})\|} \right\rangle \geq \frac{\sqrt{\lambda_{\min}}\exp(-D)}{16|\mathcal{I}|^2} \\
\geq &\frac{\sqrt{\lambda_{\min}}\exp(-2D_0)}{16|\mathcal{I}|^2} > 0.
\end{aligned}$$

□

C. Proofs in Section 6

C.1. Proof of Theorem 6.1

Proof of Theorem 6.1.
According Theorem 5.5, there exist constants $H, D, \mu > 0$ such that

$$\left\langle \frac{\nabla\mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}, \frac{\mathcal{P}_\perp(\boldsymbol{w})}{\|\mathcal{P}_\perp(\boldsymbol{w})\|} \right\rangle \geq \mu \text{ holds for any } \boldsymbol{w} \in \mathbb{C}(D, 2D; H),$$

where $\mathbb{C}(D, 2D; H) := \{\boldsymbol{w} \in \text{span}\{\boldsymbol{x}_i : i \in [n]\} : D \leq \|\mathcal{P}_\perp(\boldsymbol{w})\| \leq 2D; \langle \boldsymbol{w}, \boldsymbol{w}^\star\rangle \geq H\}$.

Following the proof of Theorem 5.5, we further define $\mathbb{C}(D; H)$ as:

$$\mathbb{C}(D; H) := \{\boldsymbol{w} \in \text{span}\{\boldsymbol{x}_i : i \in [n]\} : \|\mathcal{P}_\perp(\boldsymbol{w})\| = D; \langle \boldsymbol{w}, \boldsymbol{w}^\star\rangle \geq H\}.$$

It holds that $\mathbb{C}(D; H) = \{h\boldsymbol{w}^\star + D\boldsymbol{v} : h \geq H, \boldsymbol{v} \perp \boldsymbol{w}^\star, \|\boldsymbol{v}\| = 1\}$ and $\mathbb{C}(D; H) \subset \mathbb{C}(D, 2D; H)$.

Analysis of Phase I.

Phase I is a warm-up phase. We will prove that at the end of this phase, trained $\boldsymbol{w}$ can be scaled onto $\mathbb{C}(D; H)$. First, we choose the error

$$\epsilon = \min\left\{\frac{D}{2H}, \frac{1}{2}\right\}.$$

Notice that Theorem C.5 ensures that under Assumption 3.1 and 5.4, the directional convergence of GD and NGD with $\eta \leq 1$ holds, and the rates are $\mathcal{O}(1/\log t)$ and $\mathcal{O}(1/t)$,respectively.

Therefore, there exists T^ϵ such that $\left\|\frac{\boldsymbol{w}(T^\epsilon)}{\|\boldsymbol{w}(T^\epsilon)\|}-\boldsymbol{w}^\star\right\|<\epsilon$, which implies the inner satisfies:

$$\left\langle\frac{\boldsymbol{w}(T^\epsilon)}{\|\boldsymbol{w}(T^\epsilon)\|},\boldsymbol{w}^\star\right\rangle=\frac{1}{2}\left(2-\left\|\frac{\boldsymbol{w}(T^\epsilon)}{\|\boldsymbol{w}(T^\epsilon)\|}-\boldsymbol{w}^\star\right\|^2\right)>1-\frac{\epsilon^2}{2}.$$

Therefore, at T^ϵ, it holds that:

$$\begin{aligned}
&\frac{\|\mathcal{P}_\perp(\boldsymbol{w}(T^\epsilon))\|}{\langle\boldsymbol{w}(T^\epsilon),\boldsymbol{w}^\star\rangle}=\frac{\|\boldsymbol{w}(T^\epsilon)-\mathcal{P}(\boldsymbol{w}(T^\epsilon))\|}{\langle\boldsymbol{w}(T^\epsilon),\boldsymbol{w}^\star\rangle}=\left\|\frac{\boldsymbol{w}(T^\epsilon)}{\langle\boldsymbol{w}(T^\epsilon),\boldsymbol{w}^\star\rangle}-\boldsymbol{w}^\star\right\|\\
=&\left\|\frac{\boldsymbol{w}(T^\epsilon)}{\langle\boldsymbol{w}(T^\epsilon)),\boldsymbol{w}^\star\rangle}-\boldsymbol{w}^\star\right\|=\left\|\frac{\boldsymbol{w}(T^\epsilon)}{\langle\boldsymbol{w}(T^\epsilon)),\boldsymbol{w}^\star\rangle}-\frac{\boldsymbol{w}(T^\epsilon)}{\|\boldsymbol{w}(T^\epsilon)\|}+\frac{\boldsymbol{w}(T^\epsilon)}{\|\boldsymbol{w}(T^\epsilon)\|}-\boldsymbol{w}^\star\right\|\\
\leq&\left\|\frac{\boldsymbol{w}(T^\epsilon)}{\langle\boldsymbol{w}(T^\epsilon)),\boldsymbol{w}^\star\rangle}-\frac{\boldsymbol{w}(T^\epsilon)}{\|\boldsymbol{w}(T^\epsilon)\|}\right\|+\left\|\frac{\boldsymbol{w}(T^\epsilon)}{\|\boldsymbol{w}(T^\epsilon)\|}-\boldsymbol{w}^\star\right\|<\left|\frac{\langle\boldsymbol{w}(T^\epsilon)),\boldsymbol{w}^\star\rangle-\|\boldsymbol{w}(T^\epsilon)\|}{\langle\boldsymbol{w}(T^\epsilon)),\boldsymbol{w}^\star\rangle}\right|+\epsilon\\
=&\left|1-\frac{1}{\left\langle\frac{\boldsymbol{w}(T^\epsilon)}{\|\boldsymbol{w}(T^\epsilon)\|},\boldsymbol{w}^\star\right\rangle}\right|+\epsilon<\frac{1}{1-\frac{\epsilon^2}{2}}-1+\epsilon=\frac{\frac{\epsilon^2}{2}}{1-\frac{\epsilon^2}{2}}+\epsilon<\frac{4\epsilon^2}{7}+\epsilon\\
\leq&\left(\frac{2}{7}+1\right)\epsilon\leq 2\epsilon\leq\min\left\{\frac{D}{H},1\right\}.
\end{aligned}$$

We choose $T_{\rm w}=T^\epsilon=\Theta(1)$, and we obtain $\boldsymbol{w}(T_{\rm w})$ at the end of Phase I.

Analysis of Phase II.

For simplicity, due to $T_{\rm w}$ is an constant, we replace the time t to $t-T_{\rm w}$ in the proof of Phase II. This means that Phase II starts from $t=0$ with the initialization $\boldsymbol{w}(0)\leftarrow\boldsymbol{w}(T_{\rm w})$.

In this proof, we choose

$$\eta=\mu D,\ T_k=2k,\ R_k=\frac{D\,\|\boldsymbol{w}(T_k)\|}{\|\mathcal{P}_\perp(\boldsymbol{w}(T_k))\|},\ \forall k\geq 0.$$

Recalling Algorithm 1, the update rule is:

$$\begin{aligned}
&\cdots;\\
\boldsymbol{w}(2k+1)&=R_k\frac{\boldsymbol{w}(2k)}{\|\boldsymbol{w}(2k)\|};\\
\boldsymbol{v}(2k+2)&=\boldsymbol{w}(2k+1)-\eta\frac{\nabla\mathcal{L}(\boldsymbol{w}(2k+1))}{\mathcal{L}(\boldsymbol{w}(2k+1))};\\
\boldsymbol{w}(2k+2)&=\mathrm{Proj}_{\mathbb{B}(0,\|\boldsymbol{w}(2k+1)\|)}\left(\boldsymbol{v}(2k+2)\right);\\
\boldsymbol{w}(2k+3)&=R_{k+1}\frac{\boldsymbol{w}(2k+2)}{\|\boldsymbol{w}(2k+2)\|};\\
&\cdots
\end{aligned}$$

In general, we aim to prove the following statements:

(S1). $\boldsymbol{w}(2k+1)\in\mathbb{C}(D;H),\ \forall k\geq 0.$

(S2). $\langle\boldsymbol{w}(2k+1),\boldsymbol{w}^\star\rangle\geq\frac{1}{\left(\sqrt{1-2\mu}\right)^k}\left(\langle\boldsymbol{w}(1),\boldsymbol{w}^\star\rangle+\frac{\gamma^\star}{1-\sqrt{1-2\mu}}\right)-\frac{\gamma^\star}{1-\sqrt{1-2\mu}},\ \forall k\geq 0;$

$$\langle\boldsymbol{w}(2k+1),\boldsymbol{w}^\star\rangle\leq\frac{1}{\left(\sqrt{1-\mu^2}\right)^k}\left(\langle\boldsymbol{w}(1),\boldsymbol{w}^\star\rangle+\frac{1}{1-\sqrt{1-\mu^2}}\right)-\frac{1}{1-\sqrt{1-\mu^2}},\ \forall k\geq 0.$$

(S3). $D\sqrt{1-2\mu} \leq \|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\| \leq D\sqrt{1-\mu^2},\ \forall k \geq 0.$

(S4). $\frac{\langle \boldsymbol{w}(2k+2), \boldsymbol{w}^\star \rangle}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+2))\|} = \frac{\langle \boldsymbol{w}(2k+3), \boldsymbol{w}^\star \rangle}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+3))\|} = \frac{\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle}{D} e^{\Theta(k)}.$

(S5). $R_{k+1} = \langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle e^{\Theta(k)}.$

(S6). $\left\| \frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|} - \boldsymbol{w}^\star \right\| = \frac{D}{\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle} e^{-\Theta(t)}.$

(S7). $\gamma^\star - \gamma(\boldsymbol{w}(t)) = \frac{D}{\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle} e^{-\Theta(t)}.$

Step I. Proof of (S1)(S2).

In this step, we will prove (S1)(S2) by induction.

Step I (i). We prove (S1)(S2) for $k=0$. Recalling our analysis of Phase I, it holds that

$$\frac{\|\mathcal{P}_\perp(\boldsymbol{w}(0))\|}{\langle \boldsymbol{w}(0)), \boldsymbol{w}^\star \rangle} \leq \min\left\{\frac{D}{H}, 1\right\}.$$

Thus, if we choose $R_0 = \frac{D\|\boldsymbol{w}(0)\|}{\|\mathcal{P}_\perp(\boldsymbol{w}(0))\|}$ in Algorithm 1, then $\boldsymbol{w}(1) = \frac{D}{\|\mathcal{P}_\perp(\boldsymbol{w}(0))\|} \cdot \boldsymbol{w}(0)$ and $\boldsymbol{w}(1)$ satisfies:

$$\|\mathcal{P}_\perp(\boldsymbol{w}(1))\| = \left\|\mathcal{P}_\perp\left(\frac{D}{\|\mathcal{P}_\perp(\boldsymbol{w}(0))\|}\boldsymbol{w}(0)\right)\right\| = \left\|\frac{D\mathcal{P}_\perp(\boldsymbol{w}(0))}{\|\mathcal{P}_\perp(\boldsymbol{w}(0))\|}\right\| = D;$$

$$\begin{aligned}&\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle = \left\langle \frac{D}{\|\mathcal{P}_\perp(\boldsymbol{w}(0))\|}\boldsymbol{w}(0), \boldsymbol{w}^\star \right\rangle = D\left\langle \frac{\boldsymbol{w}(0)}{\|\mathcal{P}_\perp(\boldsymbol{w}(0))\|}, \boldsymbol{w}^\star \right\rangle \\ =&D\frac{\langle \boldsymbol{w}(0), \boldsymbol{w}^\star \rangle}{\|\mathcal{P}_\perp(\boldsymbol{w}(0))\|} \geq \frac{D}{\min\left\{\frac{D}{H}, 1\right\}} = \max\left\{H, D\right\}.\end{aligned}$$

which means that (S1) $\boldsymbol{w}(1) \in \mathbb{C}(D; H)$ holds for $k=0$. As for (S2), it is trivial for $k=0$.

Step I (ii). Assume (S1)(S2) hold for any $0 \leq k' \leq k$. Then we will prove for $k' = k+1$.

First, it is easy to bound the difference between $\langle \boldsymbol{v}(2k+2), \boldsymbol{w}^\star \rangle$ and $\langle \boldsymbol{w}(2k+1), \boldsymbol{w}^\star \rangle$:

$$\begin{aligned}&\langle \boldsymbol{v}(2k+2), \boldsymbol{w}^\star \rangle - \langle \boldsymbol{w}(2k+1), \boldsymbol{w}^\star \rangle \\ =&\eta\left\langle -\frac{\nabla\mathcal{L}(\boldsymbol{w}(2k+1))}{\mathcal{L}(\boldsymbol{w}(2k+1))}, \boldsymbol{w}^\star \right\rangle \overset{\text{Lemma C.1}}{\in} [\eta\gamma^\star, \eta] = [\mu\gamma^\star D, \mu D].\end{aligned} \tag{10}$$

Secondly, notice the following fact about $\boldsymbol{w}(2k+3)$:

$$\begin{aligned}&\boldsymbol{w}(2k+3) = R_{k+1}\frac{\boldsymbol{w}(2k+2)}{\|\boldsymbol{w}(2k+2)\|} = \frac{D\,\|\boldsymbol{w}(2k+2)\|}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+2))\|}\frac{\boldsymbol{w}(2k+2)}{\|\boldsymbol{w}(2k+2)\|} \\ =&\frac{D}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+2))\|}\boldsymbol{w}(2k+2) = \frac{D}{\|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|}\boldsymbol{v}(2k+2).\end{aligned} \tag{11}$$

With the help of the estimates above and the induction, now we can give the following two-sided bounds for $\langle \boldsymbol{w}(2k+3), \boldsymbol{w}^\star \rangle$.

- Upper bound for $\langle \boldsymbol{w}(2k+3), \boldsymbol{w}^\star \rangle$:

$$
\begin{aligned}
\langle \boldsymbol{w}(2k+3), \boldsymbol{w}^\star \rangle \overset{(11)}{=} & D \frac{\langle \boldsymbol{v}(2k+2), \boldsymbol{w}^\star \rangle}{\|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|} \\
\overset{(10)}{\leq} & D \frac{\langle \boldsymbol{w}(2k+1), \boldsymbol{w}^\star \rangle + 1}{\sqrt{1-\mu^2} D} = \frac{\langle \boldsymbol{w}(2k+1), \boldsymbol{w}^\star \rangle + 1}{\sqrt{1-\mu^2}} \\
\overset{\text{induction}}{\leq} & \frac{1}{\sqrt{1-\mu^2}} \left(\frac{1}{\left(\sqrt{1-\mu^2}\right)^k} \left(\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle + \frac{1}{1-\sqrt{1-\mu^2}} \right) - \frac{1}{1-\sqrt{1-\mu^2}} + 1 \right) \\
= & \frac{1}{\left(\sqrt{1-\mu^2}\right)^{k+1}} \left(\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle + \frac{1}{1-\sqrt{1-\mu^2}} \right) - \frac{1}{1-\sqrt{1-\mu^2}}.
\end{aligned}
\tag{12}
$$

- Lower bound for $\langle \boldsymbol{w}(2k+3), \boldsymbol{w}^\star \rangle$:

$$
\begin{aligned}
\langle \boldsymbol{w}(2k+3), \boldsymbol{w}^\star \rangle \overset{(11)}{=} & D \frac{\langle \boldsymbol{v}(2k+2), \boldsymbol{w}^\star \rangle}{\|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|} \\
\overset{(10)}{\geq} & D \frac{\langle \boldsymbol{w}(2k+1), \boldsymbol{w}^\star \rangle + \gamma^\star}{\sqrt{1-2\mu} D} = \frac{\langle \boldsymbol{w}(2k+1), \boldsymbol{w}^\star \rangle + \gamma^\star}{\sqrt{1-2\mu}} \\
\overset{\text{induction}}{\geq} & \frac{1}{\sqrt{1-2\mu}} \left(\frac{1}{\left(\sqrt{1-2\mu}\right)^k} \left(\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle + \frac{\gamma^\star}{1-\sqrt{1-2\mu}} \right) - \frac{\gamma^\star}{1-\sqrt{1-2\mu}} + \gamma^\star \right) \\
= & \frac{1}{\left(\sqrt{1-2\mu}\right)^{k+1}} \left(\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle + \frac{\gamma^\star}{1-\sqrt{1-2\mu}} \right) - \frac{\gamma^\star}{1-\sqrt{1-2\mu}}.
\end{aligned}
\tag{13}
$$

Hence, from (12)(13), we have proved that (S2) holds for $k+1$.

Moreover, we have the following facts:

$$
\begin{aligned}
\|\mathcal{P}_\perp(\boldsymbol{w}(2k+3))\| \overset{(11)}{=} & \left\| \mathcal{P}_\perp \left(\frac{D}{\|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|} \boldsymbol{v}(2k+2) \right) \right\| \\
= & \left\| D \frac{\mathcal{P}_\perp(\boldsymbol{v}(2k+2))}{\|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|} \right\| = D,
\end{aligned}
$$

$$
\begin{aligned}
\langle \boldsymbol{w}(2k+3), \boldsymbol{w}^\star \rangle \overset{(13)}{\geq} & \frac{1}{\left(\sqrt{1-2\mu}\right)^{k+1}} \left(\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle + \frac{\gamma^\star}{1-\sqrt{1-2\mu}} \right) - \frac{\gamma^\star}{1-\sqrt{1-2\mu}} \\
\geq & \frac{\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle}{\left(\sqrt{1-2\mu}\right)^{k+1}} \geq \langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle \geq H;
\end{aligned}
$$

which means that (S1) holds for $k+1$, i.e., $\boldsymbol{w}(2k+3) \in \mathbb{C}(D; H)$.

Now we have proved (S1)(S2) for any $k \geq 0$ by induction.

Step II. Proof of (S3).

In this step, we will prove (S3) directly. For any $k \geq 0$, we can derive the following two-sides bounds:

- For the upper bound of $\|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|$, we have:

$$\begin{aligned}
&\|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|^2 = \left\|\mathcal{P}_\perp(\boldsymbol{w}(2k+1)) - \eta\mathcal{P}_\perp\left(\frac{\nabla\mathcal{L}(\boldsymbol{w}(2k+1))}{\mathcal{L}(\boldsymbol{w}(2k+1))}\right)\right\|^2 \\
=&\|\mathcal{P}_\perp(\boldsymbol{w}(2k+1))\|^2 + \eta^2\left\|\mathcal{P}_\perp\left(\frac{\nabla\mathcal{L}(\boldsymbol{w}(2k+1))}{\mathcal{L}(\boldsymbol{w}(2k+1))}\right)\right\|^2 \\
&\quad - 2\eta\left\langle \mathcal{P}_\perp(\boldsymbol{w}(2k+1)), \mathcal{P}_\perp\left(\frac{\nabla\mathcal{L}(\boldsymbol{w}(2k+1))}{\mathcal{L}(\boldsymbol{w}(2k+1))}\right)\right\rangle \\
=&D^2 + \eta^2\left\|\mathcal{P}_\perp\left(\frac{\nabla\mathcal{L}(\boldsymbol{w}(2k+1))}{\mathcal{L}(\boldsymbol{w}(2k+1))}\right)\right\|^2 - 2\eta D\left\langle \frac{\mathcal{P}_\perp(\boldsymbol{w}(2k+1))}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+1))\|}, \frac{\nabla\mathcal{L}(\boldsymbol{w}(2k+1))}{\mathcal{L}(\boldsymbol{w}(2k+1))}\right\rangle \\
\le&D^2 + \eta^2\left\|\frac{\nabla\mathcal{L}(\boldsymbol{w}(2k+1))}{\mathcal{L}(\boldsymbol{w}(2k+1))}\right\|^2 - 2\eta D\left\langle \frac{\mathcal{P}_\perp(\boldsymbol{w}(2k+1))}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+1))\|}, \frac{\nabla\mathcal{L}(\boldsymbol{w}(2k+1))}{\mathcal{L}(\boldsymbol{w}(2k+1))}\right\rangle \\
\overset{\text{Lemma C.1}}{\le}& D^2 + \eta^2 - 2\eta D\mu = D^2 + \mu^2 D^2 - 2\mu^2 D^2 = (1-\mu^2)D^2.
\end{aligned} \tag{14}$$

- For the lower bound of $\|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|$, we have:

$$\begin{aligned}
&\|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|^2 = \left\|\mathcal{P}_\perp(\boldsymbol{w}(2k+1)) - \eta\mathcal{P}_\perp\left(\frac{\nabla\mathcal{L}(\boldsymbol{w}(2k+1))}{\mathcal{L}(\boldsymbol{w}(2k+1))}\right)\right\|^2 \\
=&D^2 + \eta^2\left\|\mathcal{P}_\perp\left(\frac{\nabla\mathcal{L}(\boldsymbol{w}(2k+1))}{\mathcal{L}(\boldsymbol{w}(2k+1))}\right)\right\|^2 \\
&\quad - 2\eta D\left\langle \frac{\mathcal{P}_\perp(\boldsymbol{w}(2k+1))}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+1))\|}, \frac{\nabla\mathcal{L}(\boldsymbol{w}(2k+1))}{\mathcal{L}(\boldsymbol{w}(2k+1))}\right\rangle \\
\ge&D^2 - 2\eta D\left\langle \frac{\mathcal{P}_\perp(\boldsymbol{w}(2k+1))}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+1))\|}, \frac{\nabla\mathcal{L}(\boldsymbol{w}(2k+1))}{\mathcal{L}(\boldsymbol{w}(2k+1))}\right\rangle \\
\ge&D^2 - 2\eta D\left\|\frac{\nabla\mathcal{L}(\boldsymbol{w}(2k+1))}{\mathcal{L}(\boldsymbol{w}(2k+1))}\right\| \\
\overset{\text{Lemma C.1}}{\ge}& D^2 - 2\eta D = D^2 - 2\mu D^2 = (1-2\mu)D^2.
\end{aligned} \tag{15}$$

Hence, we have proved (S3).

Step III. Proof of (S4)(S5)(S6).

First, we derive two-sided bounds for $\frac{\langle \boldsymbol{w}(2k+2), \boldsymbol{w}^\star\rangle}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+2))\|}$. For any $k\ge 0$, we have:

- Upper bound of $\frac{\langle \boldsymbol{w}(2k+2), \boldsymbol{w}^\star\rangle}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+2))\|}$.

$$\begin{aligned}
&\frac{\langle \boldsymbol{w}(2k+2), \boldsymbol{w}^\star\rangle}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+2))\|} = \frac{\langle \boldsymbol{v}(2k+2), \boldsymbol{w}^\star\rangle}{\|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|} \\
\overset{(10)}{\le}& \frac{\langle \boldsymbol{w}(2k+1), \boldsymbol{w}^\star\rangle + \mu D}{\|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|} \\
\overset{\text{(S2)(S3)}}{\le}& \frac{\frac{1}{\left(\sqrt{1-\mu^2}\right)^k}\left(\langle \boldsymbol{w}(1), \boldsymbol{w}^\star\rangle + \frac{1}{1-\sqrt{1-\mu^2}}\right) - \frac{1}{1-\sqrt{1-\mu^2}} + \mu D}{D\sqrt{1-2\mu}} \\
\le& \frac{\langle \boldsymbol{w}(1), \boldsymbol{w}^\star\rangle}{D} e^{\Theta(k)}.
\end{aligned}$$

- Lower bound of $\frac{\langle \boldsymbol{w}(2k+2), \boldsymbol{w}^\star\rangle}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+2))\|}$.

$$
\begin{aligned}
\frac{\langle \boldsymbol{w}(2k+2), \boldsymbol{w}^\star \rangle}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+2))\|} &= \frac{\langle \boldsymbol{v}(2k+2), \boldsymbol{w}^\star \rangle}{\|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|} \\
&\overset{(10)}{\geq} \frac{\langle \boldsymbol{w}(2k+1), \boldsymbol{w}^\star \rangle + \mu\gamma^\star D}{\|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|} \\
&\overset{\text{(S2)(S3)}}{\geq} \frac{\frac{1}{\left(\sqrt{1-2\mu}\right)^k}\left(\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle + \frac{\gamma^\star}{1-\sqrt{1-2\mu}}\right) - \frac{\gamma^\star}{1-\sqrt{1-2\mu}} + \mu\gamma^\star D}{(1-\mu)D} \\
&\geq \frac{\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle}{D} e^{\Theta(k)}.
\end{aligned}
$$

Additionally, notice

$$
\frac{\langle \boldsymbol{w}(2k+2), \boldsymbol{w}^\star \rangle}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+2))\|} = \frac{\langle \boldsymbol{w}(2k+3), \boldsymbol{w}^\star \rangle}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+3))\|},
$$

we obtain (S4):

$$
\frac{\langle \boldsymbol{w}(2k+2), \boldsymbol{w}^\star \rangle}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+2))\|} = \frac{\langle \boldsymbol{w}(2k+3), \boldsymbol{w}^\star \rangle}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+3))\|} = \frac{\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle}{D} e^{\Theta(k)}.
$$

Furthermore, Combining (S4) and the following fact

$$
\begin{aligned}
R_{k+1} &= \frac{D\,\|\boldsymbol{w}(2k+2)\|}{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+2))\|} = \frac{D\,\|\boldsymbol{v}(2k+2)\|}{\|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|} \\
&= D\frac{\sqrt{\langle \boldsymbol{v}(2k+2), \boldsymbol{w}^\star \rangle^2 + \|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|^2}}{\|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|} = D\sqrt{\frac{\langle \boldsymbol{v}(2k+2), \boldsymbol{w}^\star \rangle^2}{\|\mathcal{P}_\perp(\boldsymbol{v}(2k+2))\|^2} + 1},
\end{aligned}
$$

we can obtain (S5):

$$
R_{k+1} = \langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle\, e^{\Theta(k)}.
$$

In the same way, we can prove

$$
\begin{aligned}
&\left\| \frac{\boldsymbol{w}(2k)}{\|\boldsymbol{w}(2k)\|} - \boldsymbol{w}^\star \right\| = \left\| \frac{\boldsymbol{w}(2k+1)}{\|\boldsymbol{w}(2k+1)\|} - \boldsymbol{w}^\star \right\| = 2\left(1 - \left\langle \frac{\boldsymbol{w}(2k+1)}{\|\boldsymbol{w}(2k+1)\|}, \boldsymbol{w}^\star \right\rangle\right) \\
=&2\left(1 - \frac{\langle \boldsymbol{w}(2k+1), \boldsymbol{w}^\star \rangle}{\|\boldsymbol{w}(2k+1)\|}\right) = 2\left(1 - \frac{\langle \boldsymbol{w}(2k+1), \boldsymbol{w}^\star \rangle}{\sqrt{\langle \boldsymbol{w}(2k+1), \boldsymbol{w}^\star \rangle^2 + \|\mathcal{P}_\perp(\boldsymbol{w}(2k+1))\|^2}}\right) \\
=&2\left(1 - \frac{1}{\sqrt{1 + \frac{\|\mathcal{P}_\perp(\boldsymbol{w}(2k+1))\|^2}{\langle \boldsymbol{w}(2k+1), \boldsymbol{w}^\star \rangle^2}}}\right) \overset{\text{(S4)}}{=} 2\left(1 - \frac{1}{\sqrt{1 + \frac{D^2}{\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle^2} e^{-\Theta(k)}}}\right) \\
=&\frac{D}{\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle} e^{-\Theta(k)},
\end{aligned}
$$

which means (S6):

$$
\left\| \frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|} - \boldsymbol{w}^\star \right\| = \frac{D}{\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle} e^{-\Theta(t)},
$$

Step III. Proof of (S7). Using Lemma A.1 and (S6), we obtain (S7).

Conclusions.

From our proof of Phase II, we have $\langle \boldsymbol{w}(1), \boldsymbol{w}^\star \rangle \geq \max\{H, D\}$. Taking this fact into (S6)(S7), we obtain our conclusions:

$$
\begin{aligned}
\left\| \frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|} - \boldsymbol{w}^\star \right\| &= e^{-\Omega(t)}; \\
\gamma^\star - \gamma(\boldsymbol{w}(t)) &= e^{-\Omega(t)}.
\end{aligned}
$$

□

C.2. Proof of Theorem 6.4

The proof of GD is relatively straightforward. In contrast, the proof for NGD is significantly more intricate, necessitating a more rigorous convex optimization analysis than Proposition 4.1.

For NGD, we still focus on the dynamics of $\mathcal{P}_\perp(\boldsymbol{w}(t))$, which is orthogonal to $\boldsymbol{w}^\star$. Actually, we can prove that there exists a subsequence $\mathcal{P}_\perp(\boldsymbol{w}(t_k))$ ($t_k \to \infty$), which satisfies $\|\mathcal{P}_\perp(\boldsymbol{w}(t_k))\| = \Theta(1)$. Since the norm grows at $\|\boldsymbol{w}(t_k)\| = \Theta(t_k)$ (Thm C.5), NGD must have only $\Theta(\|\mathcal{P}_\perp(\boldsymbol{w}(t_k))\| / \|\boldsymbol{w}(t_k)\|) = \Theta(1/t_k)$ directional convergence rate. Furthermore, the non-degenerated data assumption 5.4 can also provide a two-sided bound for the margin error (Lemma C.2), which ensures $\Omega(1/t_k^2)$ margin maximization rate. Our crucial point is that the $(d-1)$-dim dynamics of $\mathcal{P}_\perp(\boldsymbol{w}(t))$ near $\mathbf{0} \in \text{span}\{\boldsymbol{x}_i : i \in \mathcal{I}\}$ are close to in-exact gradient descent dynamics on another strongly convex loss $\mathcal{L}_\perp(\cdot)$ with unique minimizer $\boldsymbol{v}^\star \in \mathbb{R}^{d-1}$. Moreover, our condition $\gamma^\star \boldsymbol{w}^\star \neq \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \boldsymbol{x}_i y_i$ can ensure that $\boldsymbol{v}^\star \neq \mathbf{0}$. Therefore, there must exists a sequence $\mathcal{P}_\perp(\boldsymbol{w}(t_k))$ which can escapes from a sufficient small ball $\mathbb{B}(\mathbf{0}_{d-1}; \epsilon_0)$, which means $\|\mathcal{P}_\perp(\boldsymbol{w}(t_k))\| = \Theta(1)$.

Proof for GD.
The proof for GD is straightforward.

By Theorem C.7, $\lim_{t\to+\infty} \Big(\boldsymbol{w}(t) - \boldsymbol{w}^\star \log t\Big) = \tilde{\boldsymbol{w}}$, where $\tilde{\boldsymbol{w}}$ is the solution to the equations:

$$\eta \exp\left(-\left\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i y_i\right\rangle\right) = \alpha_i, i \in \mathcal{I}.$$

<u>Step I. $\mathcal{P}_\perp(\tilde{\boldsymbol{w}}) \neq \mathbf{0}$.</u>

If we assume $\mathcal{P}_\perp(\tilde{\boldsymbol{w}}) = \mathbf{0}$, then there exists $c > 0$ such that $\tilde{\boldsymbol{w}} = c\boldsymbol{w}^\star$.

Notice that for any $i \in \mathcal{I}$, $\langle \boldsymbol{w}^\star, \boldsymbol{x}_i y_i\rangle = \gamma^\star$. Therefore, there exists $c' > 0$ such that $\alpha_i = c'$ for any $i \in \mathcal{I}$. Recalling $\boldsymbol{w}^\star = \sum_{i\in\mathcal{I}} \alpha_i \boldsymbol{x}_i y_i$, we have $\boldsymbol{w}^\star = c' \sum_{i\in\mathcal{I}} \boldsymbol{x}_i y_i$, which implies $\boldsymbol{w}^\star = \frac{1}{|\mathcal{I}|\gamma^\star}\sum_{i\in\mathcal{I}} \boldsymbol{x}_i y_i$. This is contradict to our condition $\gamma^\star \boldsymbol{w}^\star \neq \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \boldsymbol{x}_i y_i$.

Hence, we have proved $\mathcal{P}_\perp(\tilde{\boldsymbol{w}}) \neq \mathbf{0}$.

<u>Step II. The lower bound.</u>

Recalling $\lim_{t\to+\infty} \Big(\boldsymbol{w}(t) - \boldsymbol{w}^\star \log t\Big) = \tilde{\boldsymbol{w}}$ and our results in Step I, there exists $T_0 > 0$ such that

$$\|\boldsymbol{w}(t) - \boldsymbol{w}^\star \log t - \tilde{\boldsymbol{w}}\| \leq \frac{\|\mathcal{P}_\perp(\tilde{\boldsymbol{w}})\|}{2}, \quad \forall t \geq T_0.$$

Using the fact $\|\mathcal{P}_\perp(\boldsymbol{w})\| \leq \|\boldsymbol{w}\|$, we have

$$\|\mathcal{P}_\perp(\boldsymbol{w}(t)) - \mathcal{P}_\perp(\tilde{\boldsymbol{w}})\| \leq \|\boldsymbol{w}(t) - \boldsymbol{w}^\star \log t - \tilde{\boldsymbol{w}}\| \leq \frac{\|\mathcal{P}_\perp(\tilde{\boldsymbol{w}})\|}{2}, \quad \forall t \geq T_0,$$

which implies

$$\frac{\|\mathcal{P}_\perp(\tilde{\boldsymbol{w}})\|}{2} \leq \|\mathcal{P}_\perp(\boldsymbol{w}(t))\| \leq \frac{3\,\|\mathcal{P}_\perp(\tilde{\boldsymbol{w}})\|}{2}, \quad \forall t \geq T_0.$$

Recalling Theorem C.5, it holds that $\|\boldsymbol{w}(t)\| = \Theta(\log t)$. Then, a direct calculation ensures that:

$$\begin{aligned}
&\left\|\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|} - \boldsymbol{w}^\star\right\|^2 = 2 - 2\frac{\langle \boldsymbol{w}(t), \boldsymbol{w}^\star\rangle}{\|\boldsymbol{w}(t)\|} \\
=&2 - 2\frac{\langle \boldsymbol{w}(t), \boldsymbol{w}^\star\rangle}{\sqrt{\langle \boldsymbol{w}(t), \boldsymbol{w}^\star\rangle^2 + \|\mathcal{P}_\perp(\boldsymbol{w}(t))\|^2}} \\
=&2 - \frac{2}{\sqrt{1 + \frac{\|\mathcal{P}_\perp(\boldsymbol{w}(t))\|^2}{\langle \boldsymbol{w}(t), \boldsymbol{w}^\star\rangle^2}}} = \Theta\left(\frac{\|\mathcal{P}_\perp(\boldsymbol{w}(t))\|^2}{\langle \boldsymbol{w}(t), \boldsymbol{w}^\star\rangle^2}\right)
\end{aligned}$$

$$=\Theta\left(\frac{\left\|\mathcal{P}_{\perp}(\boldsymbol{w}(t_k))\right\|^2}{\left\|\boldsymbol{w}(t)\right\|^2-\left\|\mathcal{P}_{\perp}(\boldsymbol{w}(t))\right\|^2}\right)=\Theta\left(\frac{1}{\log^2 t}\right),$$

which implies the tight bound for the directional convergence rate: $\left\|\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|}-\boldsymbol{w}^\star\right\|=\Theta\left(\frac{1}{\log t}\right)$. Moreover, with the help of Lemma C.2, we have the lower bound for the margin maximization rate: $\gamma^\star-\gamma(\boldsymbol{w}(t))=\Omega\left(\frac{1}{\log^2 t}\right)$.

□

Proof for NGD.
NGD is more difficult to analyze than GD due to the more aggressive step size, and we need more refined convex optimization analysis.

Without loss of generality, we can assume $\mathrm{span}\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}=\mathbb{R}^d$. This is because: GD, NGD, and PRGD can only evaluate in $\mathrm{span}\{\boldsymbol{x}_i:i\in[n]\}$, i.e. $\boldsymbol{w}(t)\in\mathrm{span}\{\boldsymbol{x}_i:i\in[n]\}$. If $\mathrm{span}\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}\neq\mathbb{R}^d$, we only need to change the proof in the subspace $\mathrm{span}\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$. For simplicity, we still denote $\boldsymbol{z}_i:=\boldsymbol{x}_iy_i$ ($i\in[n]$).

With the help of Theorem C.5, the upper bounds hold: $\gamma^\star-\gamma(\boldsymbol{w}(t))\leq\left\|\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|}-\boldsymbol{w}^\star\right\|=\mathcal{O}(1/t)$. So we only need to prove the lower bounds for $\gamma^\star-\gamma(\boldsymbol{w}(t))$ and $\left\|\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|}-\boldsymbol{w}^\star\right\|$.

Proof Outline.

We **aim to prove the following claim**:

there exists constant $T_0>0$ and $\epsilon_0>0$, such that:
for any $T>T_0$, there exists $t>T$ s.t. $\|\mathcal{P}_{\perp}(\boldsymbol{w}(t))\|>\epsilon_0$.

If we can prove this conclusion, then there exists a subsequence $\boldsymbol{w}(t_k)$ satisfying $t_{k+1}>t_k$, $t_k\to\infty$, and $\|\mathcal{P}_{\perp}(\boldsymbol{w}(t_k))\|>\epsilon_0$. Recalling Theorem C.5, $\|\boldsymbol{w}(t_k)\|=\Theta(t_k)$. Therefore, it must holds $\left\|\frac{\boldsymbol{w}(t_k)}{\|\boldsymbol{w}(t_k)\|}-\boldsymbol{w}^\star\right\|=\Omega(1/t_k)$.

Proof Preparation.

For simplicity, we denote the optimization problem orthogonal to $\boldsymbol{w}^\star$ as

$$\min_{\boldsymbol{v}}:\mathcal{L}_{\perp}(\boldsymbol{v})=\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\exp\left(-y_i\left\langle\boldsymbol{v},\mathcal{P}_{\perp}(\boldsymbol{x}_i)\right\rangle\right),\boldsymbol{v}\in\mathrm{span}\{\mathcal{P}_{\perp}(\boldsymbol{x}_i):i\in\mathcal{I}\}.$$

In this proof, we focus on the dynamics of $\mathcal{P}_{\perp}(\boldsymbol{w}(t))$, satisfying:

$$\mathcal{P}_{\perp}(\boldsymbol{w}(t+1))=\mathcal{P}_{\perp}(\boldsymbol{w}(t))-\eta\mathcal{P}_{\perp}\left(\frac{\nabla\mathcal{L}(\boldsymbol{w}(t))}{\mathcal{L}(\boldsymbol{w}(t))}\right)$$
$$=\mathcal{P}_{\perp}(\boldsymbol{w}(t))-\eta\mathcal{P}_{\perp}\left(\frac{\frac{1}{n}\sum_{i=1}^n e^{-\langle\boldsymbol{w}(t),\boldsymbol{z}_i\rangle}\boldsymbol{z}_i}{\frac{1}{n}\sum_{i=1}^n e^{-\langle\boldsymbol{w}(t),\boldsymbol{z}_i\rangle}}\right)$$
$$=\mathcal{P}_{\perp}(\boldsymbol{w}(t))-\eta\frac{\sum_{i=1}^n e^{-\langle\boldsymbol{w}(t),\boldsymbol{z}_i\rangle}\mathcal{P}_{\perp}(\boldsymbol{z}_i)}{\sum_{i=1}^n e^{-\langle\boldsymbol{w}(t),\boldsymbol{z}_i\rangle}}.$$

With the help of Theorem C.6, we know that

(L1) the minimizer $\boldsymbol{v}^*\in\mathrm{span}\{\mathcal{P}_{\perp}(\boldsymbol{x}_i):i\in\mathcal{I}\}$ (of $\mathcal{L}_{\perp}(\cdot)$) is unique.
(L2) there exists an absolute constant $C>0$ such that $\|\mathcal{P}_{\perp}(\boldsymbol{w}(t))-\boldsymbol{v}^\star\|\leq C$, $\forall t$;
(L3) there exists $\mu>0$ such that $\mathcal{L}_{\perp}(\cdot)$ is μ-strongly convex in $\{\boldsymbol{v}:\|\boldsymbol{v}\|<C+\|\boldsymbol{v}^\star\|\}$;

It is also easy to verify the L-smoothness:

(L4) there exists $L>0$ such that $\mathcal{L}_{\perp}(\cdot)$ is L-smooth in $\{\boldsymbol{v}:\|\boldsymbol{v}\|<C+\|\boldsymbol{v}^\star\|\}$.

Step I. The minimizer $\boldsymbol{v}^\star \neq 0$.

If $\boldsymbol{v}^\star = \mathbf{0}$, then $\nabla \mathcal{L}_\perp(\mathbf{0}) = \mathbf{0}$, which implies

$$\mathbf{0} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} e^0 \mathcal{P}(\boldsymbol{z}_i) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \mathcal{P}(\boldsymbol{z}_i).$$

Therefore,

$$\begin{aligned}\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \boldsymbol{z}_i =& \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \langle \boldsymbol{z}_i, \boldsymbol{w}^\star\rangle \boldsymbol{w}^\star + \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \mathcal{P}(\boldsymbol{z}_i) \\ =& \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \gamma^\star \boldsymbol{w}^\star = \gamma^\star \boldsymbol{w}^\star,\end{aligned}$$

which is contradict to $\gamma^\star \boldsymbol{w}^\star \neq \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \boldsymbol{z}_i$.

Step II. The gradient error near $\mathbf{0} \in \mathrm{span}\{\mathcal{P}_\perp(\boldsymbol{x}_i) : i \in \mathcal{I}\}$.

Notice that the update rule of $\mathcal{P}_\perp(\boldsymbol{w}(t))$ can be written as

$$\mathcal{P}_\perp(\boldsymbol{w}(t+1)) = \mathcal{P}_\perp(\boldsymbol{w}(t)) - \eta \mathcal{P}_\perp\left(\frac{\nabla \mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}\right) = \mathcal{P}_\perp(\boldsymbol{w}(t)) - \eta\Big(\nabla \mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}(t))) + \boldsymbol{\epsilon}(\boldsymbol{w}(t))\Big).$$

In this step, we will prove: there exists $\epsilon_0 > 0$ and $R_0 > 0$ such that the gradient error

$$\|\boldsymbol{\epsilon}(\boldsymbol{w})\| = \left\|\mathcal{P}_\perp\left(\frac{\nabla \mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}\right) - \nabla \mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}))\right\| \leq \frac{1}{2}\|\nabla \mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}))\|$$

holds for any $\boldsymbol{w}$ satisfying $\langle \boldsymbol{w}, \boldsymbol{w}^\star\rangle > R_0$ and $\|\mathcal{P}_\perp(\boldsymbol{w})\| < \epsilon_0$.

Step II (i). For some $\epsilon_1 > 0$, $\|\nabla \mathcal{L}_\perp(\boldsymbol{v}) - \nabla \mathcal{L}_\perp(\mathbf{0})\| < \frac{1}{8}\|\nabla \mathcal{L}_\perp(\mathbf{0})\|$ for any $\boldsymbol{v} \in \mathbb{B}(\mathbf{0}; \epsilon_1)$.

Notice that Step I ensures $\nabla \mathcal{L}_\perp(\mathbf{0}) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \mathcal{P}(\boldsymbol{z}_i) \neq \mathbf{0}$. We choose

$$\epsilon_1 = \min\left\{\frac{\|\nabla \mathcal{L}_\perp(\mathbf{0})\|}{8L}, C + \|\boldsymbol{v}^\star\|\right\}.$$

Then for any $\boldsymbol{v} \in \mathbb{B}(\mathbf{0}, \epsilon_1)$, then (iv) ($L$-smooth) ensures that:

$$\|\nabla \mathcal{L}_\perp(\boldsymbol{v}) - \nabla \mathcal{L}_\perp(\mathbf{0})\| \leq L\|\boldsymbol{v} - \mathbf{0}\| < L\epsilon_1 \leq \frac{1}{8}\|\nabla \mathcal{L}_\perp(\mathbf{0})\|.$$

Step II (ii). For some $\epsilon_2 > 0$ and $R_0 > 0$, $\left\|\mathcal{P}_\perp\left(\frac{\nabla \mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}\right) - \nabla \mathcal{L}_\perp(\mathbf{0})\right\| \leq \frac{1}{8}\|\nabla \mathcal{L}_\perp(\mathbf{0})\|$ holds for any $\boldsymbol{w}$ satisfying $\langle \boldsymbol{w}, \boldsymbol{w}^\star\rangle > R_0$ and $\|\mathcal{P}_\perp(\boldsymbol{w})\| < \epsilon_2$.

Due to $\|\nabla \mathcal{L}_\perp(\mathbf{0})\|/16 \neq 0$, using Lemma C.3, there exists $\epsilon_2 > 0$ and $R_2 > 0$ such that: for any $\boldsymbol{w}$ satisfying $\langle \boldsymbol{w}, \boldsymbol{w}^\star\rangle > R_0$ and $\|\mathcal{P}_\perp(\boldsymbol{w})\| < \epsilon_2$, it holds that:

$$\begin{aligned}&\left\|\mathcal{P}_\perp\left(\frac{\nabla \mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}\right) - \nabla \mathcal{L}_\perp(\mathbf{0})\right\| \\ =&\left\|\frac{\sum_{i=1}^n e^{-\langle \boldsymbol{w}, \boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)}{\sum_{j=1}^n e^{-\langle \boldsymbol{w}, \boldsymbol{z}_j\rangle}} - \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \mathcal{P}_\perp(\boldsymbol{z}_i)\right\| \\ \leq&\left\|\frac{\sum_{i=1}^n e^{-\langle \boldsymbol{w}, \boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)}{\sum_{j=1}^n e^{-\langle \boldsymbol{w}, \boldsymbol{z}_j\rangle}} - \frac{\sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w}, \boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)}{\sum_{j\in\mathcal{I}} e^{-\langle \boldsymbol{w}, \boldsymbol{z}_j\rangle}}\right\| \\ &+\left\|\frac{\sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w}, \boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)}{\sum_{j\in\mathcal{I}} e^{-\langle \boldsymbol{w}, \boldsymbol{z}_j\rangle}} - \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \mathcal{P}_\perp(\boldsymbol{z}_i)\right\|\end{aligned}$$

$$\overset{\text{Lemma C.3 (iii)(iv)}}{\leq} \frac{\|\nabla\mathcal{L}_\perp(\mathbf{0})\|}{16} + \frac{\|\nabla\mathcal{L}_\perp(\mathbf{0})\|}{16} = \frac{\|\nabla\mathcal{L}_\perp(\mathbf{0})\|}{8}.$$

Step II (iii). Based on Step II (i) and (ii), we can select

$$\epsilon_3 := \min\{\epsilon_1, \epsilon_2\}, \quad R_0 := R_0.$$

Then for any $\boldsymbol{w}$ satisfying $\langle \boldsymbol{w}, \boldsymbol{w}^\star \rangle > R_0$ and $\|\mathcal{P}_\perp(\boldsymbol{w})\| < \epsilon_3$, it holds that:

$$\begin{aligned}
\|\boldsymbol{\epsilon}(\boldsymbol{w})\| &= \left\|\mathcal{P}_\perp\left(\frac{\nabla\mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}\right) - \nabla\mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}))\right\| \\
&\leq \left\|\mathcal{P}_\perp\left(\frac{\nabla\mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}\right) - \nabla\mathcal{L}_\perp(\mathbf{0})\right\| + \|\nabla\mathcal{L}_\perp(\mathbf{0}) - \nabla\mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}))\| \\
&\overset{\text{Step II (i) and (ii)}}{\leq} \frac{\|\nabla\mathcal{L}_\perp(\mathbf{0})\|}{8} + \frac{\|\nabla\mathcal{L}_\perp(\mathbf{0})\|}{8} = \frac{\|\nabla\mathcal{L}_\perp(\mathbf{0})\|}{4} \overset{\text{Step II (i)}}{<} \frac{\|\nabla\mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}))\|}{2}.
\end{aligned}$$

Step III. The proof of the main claim.

From Theorem C.5, we know $\|\boldsymbol{w}(t)\| = \Theta(t)$ and $\left\|\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|} - \boldsymbol{w}^\star\right\| = \mathcal{O}(1/t)$. Therefore, there exists $T_0 > 0$ such that $\langle \boldsymbol{w}(t), \boldsymbol{w}^\star \rangle > R_0$ holds for any $t > T_0$ (where R_0 is defined in Step II). Additionally, we choose

$$\epsilon_0 := \min\left\{\epsilon_3, \frac{1}{2}\|\boldsymbol{v}^\star\|\right\},$$

where ϵ_3 is defined in Step II.

Consequently, in this step, we aim to prove: there exists constant $T_0 > 0$ and $\epsilon_0 > 0$, such that: for any $T > T_0$, there exists $t > T$ s.t. $\|\mathcal{P}_\perp(\boldsymbol{w}(t))\| > \epsilon_0$.

Given any $T > T_0$, now we assume that $\|\mathcal{P}_\perp(\boldsymbol{w}(t))\| < \epsilon_0$ holds for any $t > T$.

Recalling Theorem C.6, it ensures that $\mathcal{L}_\perp(\cdot)$ is μ-strongly convex in $\mathbb{B}(\boldsymbol{v}^\star; C) - \mathbb{B}(\boldsymbol{v}^\star; \delta)$ for some $\mu > 0$. Therefore,

$$\|\nabla\mathcal{L}_\perp(\boldsymbol{w})\| \geq \mu\|\boldsymbol{v} - \boldsymbol{v}^\star\| \geq \mu\delta, \quad \forall \boldsymbol{v} \in \mathbb{B}(\boldsymbol{v}^\star; C) - \mathbb{B}(\boldsymbol{v}^\star; \delta).$$

By our result in Step II, for any $t > T$, the gradient error holds that

$$\left\|\mathcal{P}_\perp\left(\frac{\nabla\mathcal{L}(\boldsymbol{w}(t))}{\mathcal{L}(\boldsymbol{w}(t))}\right) - \nabla\mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}(t)))\right\| \leq \frac{1}{2}\|\nabla\mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}(t)))\|.$$

Hence, by setting $\eta \leq 1/9L$, the loss descent has the following lower bound: for any $t > T$,

$$\begin{aligned}
\mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}(t))) - \mathcal{L}_\perp^\star &= \mathcal{L}_\perp\left(\mathcal{P}_\perp(\boldsymbol{w}(t-1)) - \eta\mathcal{P}_\perp\left(\frac{\nabla\mathcal{L}(\boldsymbol{w}(t-1))}{\mathcal{L}(\boldsymbol{w}(t-1))}\right)\right) - \mathcal{L}_\perp^\star \\
&\overset{\text{Lemma D.3}}{\leq} \mathcal{L}_\perp\left(\mathcal{P}_\perp(\boldsymbol{w}(t-1))\right) - \mathcal{L}_\perp^\star - \eta\left\langle \nabla\mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}(t-1))), \mathcal{P}_\perp\left(\frac{\nabla\mathcal{L}(\boldsymbol{w}(t-1))}{\mathcal{L}(\boldsymbol{w}(t-1))}\right)\right\rangle \\
&\quad + \frac{L}{2}\eta^2\left\|\mathcal{P}_\perp\left(\frac{\nabla\mathcal{L}(\boldsymbol{w}(t-1))}{\mathcal{L}(\boldsymbol{w}(t-1))}\right)\right\|^2 \\
&= \mathcal{L}_\perp\left(\mathcal{P}_\perp(\boldsymbol{w}(t-1))\right) - \mathcal{L}_\perp^\star - \eta\|\nabla\mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}(t-1)))\|^2 \\
&\quad - \eta\left\langle \nabla\mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}(t-1))), \mathcal{P}_\perp\left(\frac{\nabla\mathcal{L}(\boldsymbol{w}(t-1))}{\mathcal{L}(\boldsymbol{w}(t-1))}\right) - \nabla\mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}(t-1)))\right\rangle \\
&\quad + \frac{L}{2}\eta^2\left\|\mathcal{P}_\perp\left(\frac{\nabla\mathcal{L}(\boldsymbol{w}(t-1))}{\mathcal{L}(\boldsymbol{w}(t-1))}\right)\right\|^2 \\
&\leq \mathcal{L}_\perp\left(\mathcal{P}_\perp(\boldsymbol{w}(t-1))\right) - \mathcal{L}_\perp^\star
\end{aligned}$$

$$
\begin{aligned}
&- \eta \left(\|\nabla \mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}(t-1)))\|^2 - \frac{1}{2} \|\nabla \mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}(t-1)))\|^2 \right) \\
&+ \frac{L\eta^2}{2} \left(\frac{3}{2}\right)^2 \|\nabla \mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}(t-1)))\|^2 \\
\le & \mathcal{L}_\perp \left(\mathcal{P}_\perp(\boldsymbol{w}(t-1))\right) - \mathcal{L}_\perp^\star - \eta \left(\frac{1}{2} - \frac{9\eta L}{8}\right) \|\nabla \mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}(t-1)))\|^2 \\
\le & \mathcal{L}_\perp \left(\mathcal{P}_\perp(\boldsymbol{w}(t-1))\right) - \mathcal{L}_\perp^\star - \frac{3\eta}{8} \|\nabla \mathcal{L}_\perp(\mathcal{P}_\perp(\boldsymbol{w}(t-1)))\|^2 \\
\overset{\text{Lemma D.3}}{\le} & \mathcal{L}_\perp \left(\mathcal{P}_\perp(\boldsymbol{w}(t-1))\right) - \mathcal{L}_\perp^\star - \frac{3\eta}{8} \cdot 2\mu \left(\mathcal{L}_\perp \left(\mathcal{P}_\perp(\boldsymbol{w}(t-1))\right) - \mathcal{L}_\perp^\star\right) \\
\le & \left(1 - \frac{3\eta\mu}{4}\right) \left(\mathcal{L}_\perp \left(\mathcal{P}_\perp(\boldsymbol{w}(t-1))\right) - \mathcal{L}_\perp^\star\right) \\
\le & \cdots \\
\le & \left(1 - \frac{3\eta\mu}{4}\right)^{t-T} \left(\mathcal{L}_\perp \left(\mathcal{P}_\perp(\boldsymbol{w}(T_\epsilon))\right) - \mathcal{L}_\perp^\star\right).
\end{aligned}
$$

Hence, there exists time $t > T$ such that $\mathcal{L}_\perp \left(\mathcal{P}_\perp(\boldsymbol{w}(t))\right) - \mathcal{L}_\perp^\star < \frac{\mu\epsilon_0}{4}$.

On the other hand, the strong convexity and Lemma D.3 implies that

$$
\begin{aligned}
&\mathcal{L}_\perp \left(\mathcal{P}_\perp(\boldsymbol{w}(t))\right) - \mathcal{L}_\perp^\star \ge \frac{\mu}{2} \|\mathcal{P}_\perp(\boldsymbol{w}(t)) - \boldsymbol{v}^\star\| \ge \frac{\mu}{2} \left(\|\boldsymbol{v}^\star\| - \|\mathcal{P}_\perp(\boldsymbol{w}(t)\|\right) \\
> & \frac{\mu}{2}(\|\boldsymbol{v}^\star\| - \epsilon_0) \ge \frac{\mu\epsilon_0}{4}.
\end{aligned}
$$

Thus, we obtain the **contradiction**. Hence, our **main claim** holds.

Step IV. Final Lower bound.

From our result in Step III, there exists a subsequence $\boldsymbol{w}(t_k)$ satisfying $t_{k+1} > t_k$, $t_k \to \infty$, and $\|\mathcal{P}_\perp(\boldsymbol{w}(t_k))\| > \epsilon_0$. Recalling (L2), it holds that $\|\mathcal{P}_\perp(\boldsymbol{w}(t_k))\| \le C + \|\boldsymbol{v}^\star\|$. Therefore, $\|\mathcal{P}_\perp \boldsymbol{w}(t_k)\| = \Theta(1)$.

Recalling Theorem C.5, it holds that $\|\boldsymbol{w}(t_k)\| = \Theta(t_k)$. Then, a direct calculation ensures that:

$$
\begin{aligned}
&\left\| \frac{\boldsymbol{w}(t_k)}{\|\boldsymbol{w}(t_k)\|} - \boldsymbol{w}^\star \right\|^2 = 2 - 2\frac{\langle \boldsymbol{w}(t_k), \boldsymbol{w}^\star \rangle}{\|\boldsymbol{w}(t_k)\|} \\
= & 2 - 2\frac{\langle \boldsymbol{w}(t_k), \boldsymbol{w}^\star \rangle}{\sqrt{\langle \boldsymbol{w}(t_k), \boldsymbol{w}^\star \rangle^2 + \|\mathcal{P}_\perp(\boldsymbol{w}(t_k))\|^2}} \\
= & 2 - \frac{2}{\sqrt{1 + \frac{\|\mathcal{P}_\perp(\boldsymbol{w}(t_k))\|^2}{\langle \boldsymbol{w}(t_k), \boldsymbol{w}^\star \rangle^2}}} = \Theta\left(\frac{\|\mathcal{P}_\perp(\boldsymbol{w}(t_k))\|^2}{\langle \boldsymbol{w}(t_k), \boldsymbol{w}^\star \rangle^2}\right) \\
= & \Theta\left(\frac{\|\mathcal{P}_\perp(\boldsymbol{w}(t_k))\|^2}{\|\boldsymbol{w}(t_k)\|^2 - \|\mathcal{P}_\perp(\boldsymbol{w}(t_k))\|^2}\right) = \Theta\left(\frac{1}{\|\boldsymbol{w}(t_k)\|^2 - \|\mathcal{P}_\perp(\boldsymbol{w}(t_k))\|^2}\right) = \Theta\left(\frac{1}{t_k^2}\right),
\end{aligned}
$$

which implies the tight bound for the directional convergence rate: $\left\| \frac{\boldsymbol{w}(t_k)}{\|\boldsymbol{w}(t_k)\|} - \boldsymbol{w}^\star \right\| = \Theta\left(\frac{1}{t_k}\right)$. Moreover, with the help of Lemma C.2, we have the lower bound for the margin maximization rate: $\gamma^\star - \gamma(\boldsymbol{w}(t_k)) = \Omega\left(\frac{1}{t_k^2}\right)$.

Hence, we have proved Theorem 6.4 for NGD.

□

C.3. Useful Lemmas

Lemma C.1. *Under Assumption 3.1, it holds that*

$$\gamma^\star \le \left\langle -\frac{\nabla\mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}, \boldsymbol{w}^\star \right\rangle \le 1, \quad \gamma^\star \le \left\| \frac{\nabla\mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})} \right\| \le 1, \quad \forall \boldsymbol{w} \in \mathbb{R}^d.$$

Proof of Lemma C.1. For any $\boldsymbol{w} \in \mathbb{R}^d$, we have:

$$\left\langle -\frac{\nabla\mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}, \boldsymbol{w}^\star \right\rangle = \frac{\frac{1}{n}\sum_{i=1}^n e^{-y_i\langle \boldsymbol{w},\boldsymbol{x}_i\rangle} y_i \langle \boldsymbol{w}^\star, \boldsymbol{x}_i\rangle}{\frac{1}{n}\sum_{i=1}^n e^{-y_i\langle \boldsymbol{w},\boldsymbol{x}_i\rangle}} \ge \frac{\frac{1}{n}\sum_{i=1}^n e^{-y_i\langle \boldsymbol{w},\boldsymbol{x}_i\rangle}\gamma^\star}{\frac{1}{n}\sum_{i=1}^n e^{-y_i\langle \boldsymbol{w},\boldsymbol{x}_i\rangle}} = \gamma^*,$$

$$\left\langle -\frac{\nabla\mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}, \boldsymbol{w}^\star \right\rangle = \frac{\frac{1}{n}\sum_{i=1}^n e^{-y_i\langle \boldsymbol{w},\boldsymbol{x}_i\rangle} y_i \langle \boldsymbol{w}^\star, \boldsymbol{x}_i\rangle}{\frac{1}{n}\sum_{i=1}^n e^{-y_i\langle \boldsymbol{w},\boldsymbol{x}_i\rangle}} \le \frac{\frac{1}{n}\sum_{i=1}^n e^{-y_i\langle \boldsymbol{w},\boldsymbol{x}_i\rangle}}{\frac{1}{n}\sum_{i=1}^n e^{-y_i\langle \boldsymbol{w},\boldsymbol{x}_i\rangle}} = 1.$$

For the lower bound of $\|\nabla\mathcal{L}(\boldsymbol{w})/\mathcal{L}(\boldsymbol{w})\|$, it holds that

$$\left\| \frac{\nabla\mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})} \right\| \ge \left\langle -\frac{\nabla\mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})}, \boldsymbol{w}^\star \right\rangle \ge \gamma^\star.$$

For the upper bound of $\|\nabla\mathcal{L}(\boldsymbol{w})/\mathcal{L}(\boldsymbol{w})\|$, it holds that

$$\left\| \frac{\nabla\mathcal{L}(\boldsymbol{w})}{\mathcal{L}(\boldsymbol{w})} \right\| = \left\| -\frac{\frac{1}{n}\sum_{i=1}^n e^{-y_i\langle \boldsymbol{w},\boldsymbol{x}_i\rangle} y_i\boldsymbol{x}_i}{\frac{1}{n}\sum_{i=1}^n e^{-y_i\langle \boldsymbol{w},\boldsymbol{x}_i\rangle}} \right\| \le \frac{\frac{1}{n}\sum_{i=1}^n e^{-y_i\langle \boldsymbol{w},\boldsymbol{x}_i\rangle} \|y_i\boldsymbol{x}_i\|}{\frac{1}{n}\sum_{i=1}^n e^{-y_i\langle \boldsymbol{w},\boldsymbol{x}_i\rangle}} \le 1.$$

□

Lemma C.2 ((Two-sided) Margin error and Directional error)**.** *Under Assumption 3.1 and 5.4, if $\boldsymbol{w}$ satisfies $\left\|\boldsymbol{w}^\star - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right\| < (\gamma^\star_{\rm sub} - \gamma^\star)/2$ (where $\gamma^\star_{\rm sub} = \min_{i\notin\mathcal{I}} \langle \boldsymbol{w}^\star, \boldsymbol{z}_i\rangle$), then it holds that*

$$\frac{\gamma^\star}{2}\left\|\boldsymbol{w}^\star - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right\|^2 \le \gamma^\star - \gamma(\boldsymbol{w}) \le \left\|\boldsymbol{w}^\star - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right\|.$$

Proof of Lemma C.2.
This lemma is an improved version of Lemma A.1. The second "$\le$" is ensured by Lemma A.1, and we only need to prove the first "$\le$". For simplicity, we still denote $\boldsymbol{z}_i := \boldsymbol{z}_i, i \in [n]$.

Step I. $\gamma(\boldsymbol{w}) = \min_{i\in\mathcal{I}}\left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{z}_i\right\rangle$.

For any $i \in [n]$, we have

$$\left|\langle \boldsymbol{w}^\star, \boldsymbol{z}_i\rangle - \left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{z}_i\right\rangle\right| \le \left\|\boldsymbol{w}^\star - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right\| < \frac{\gamma^\star_{\rm sub} - \gamma^\star}{2},$$

which implies $\langle \boldsymbol{w}^\star, \boldsymbol{z}_i\rangle - \frac{\gamma^\star_{\rm sub} - \gamma^\star}{2} < \left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{z}_i\right\rangle < \langle \boldsymbol{w}^\star, \boldsymbol{z}_i\rangle + \frac{\gamma^\star_{\rm sub} - \gamma^\star}{2}$. Furthermore,

$$\left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{z}_i\right\rangle < \gamma^\star + \frac{\gamma^\star_{\rm sub} - \gamma^\star}{2} = \frac{\gamma^\star_{\rm sub} + \gamma^\star}{2}, \quad i \in \mathcal{I};$$

$$\left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{z}_i \right\rangle > \gamma^\star_{\text{sub}} - \frac{\gamma^\star_{\text{sub}} - \gamma^\star}{2} = \frac{\gamma^\star_{\text{sub}} + \gamma^\star}{2}, \quad i \notin \mathcal{I}.$$

Therefore, it holds that

$$\gamma(\boldsymbol{w}) = \min_{i\in[n]} \left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{z}_i \right\rangle = \min_{i\in\mathcal{I}} \left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{z}_i \right\rangle.$$

Step II. The lower bound for $\gamma^\star - \gamma(\boldsymbol{w})$.

From the result in Step I,

$$\begin{aligned}
\gamma^\star - \gamma(\boldsymbol{w}) &= \gamma^\star - \min_{i\in\mathcal{I}} \left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{z}_i \right\rangle = \max_{i\in\mathcal{I}} \left(\gamma^\star - \left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{z}_i \right\rangle \right) \\
&= \max_{i\in\mathcal{I}} \left(\langle \boldsymbol{w}^\star, \boldsymbol{z}_i \rangle - \left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{z}_i \right\rangle \right) = \max_{i\in\mathcal{I}} \left\langle \boldsymbol{w}^\star - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{z}_i \right\rangle.
\end{aligned}$$

Recalling Assumption 5.4, $\boldsymbol{w}^\star = \sum_{i\in\mathcal{I}} \alpha_i \boldsymbol{z}_i$, where $\alpha_i > 0$ and $1/\gamma^\star = \sum_{i\in\mathcal{I}} \alpha_i$. Thus, for every $\alpha_i > 0$, we have $\alpha_i(\gamma^\star - \gamma(\boldsymbol{w})) = \alpha_i \max_{i\in\mathcal{I}} \left\langle \boldsymbol{w}^\star - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{z}_i \right\rangle$, which ensures:

$$\begin{aligned}
&\left(\sum_{i\in\mathcal{I}} \alpha_i\right)(\gamma^\star - \gamma(\boldsymbol{w})) = \left(\sum_{i\in\mathcal{I}} \alpha_i\right) \max_{i\in\mathcal{I}} \left\langle \boldsymbol{w}^\star - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{z}_i \right\rangle \\
\geq& \sum_{i\in\mathcal{I}} \alpha_i \left\langle \boldsymbol{w}^\star - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{z}_i \right\rangle = \left\langle \boldsymbol{w}^\star - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \sum_{i\in\mathcal{I}} \alpha_i \boldsymbol{z}_i \right\rangle \\
=& \left\langle \boldsymbol{w}^\star - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{w}^\star \right\rangle = 1 - \left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{w}^\star \right\rangle = \frac{1}{2}\left(2 - 2\left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{w}^\star \right\rangle\right) \\
=& \frac{1}{2}\left\| \boldsymbol{w}^\star - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right\|^2.
\end{aligned}$$

Hence, we obtain:

$$\gamma^\star - \gamma(\boldsymbol{w}) \geq \frac{1}{2\sum_{i\in\mathcal{I}} \alpha_i} \left\| \boldsymbol{w}^\star - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right\|^2 = \frac{\gamma^\star}{2} \left\| \boldsymbol{w}^\star - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right\|^2.$$

□

Lemma C.3. *For any $\epsilon > 0$, there exists $\delta \in (0, 1)$ and $R > 0$, such that for any $\boldsymbol{w}$ satisfying $\|\mathcal{P}_\perp(\boldsymbol{w})\| < \delta$ and $\langle \boldsymbol{w}, \boldsymbol{w}^\star \rangle > R$, it holds that*

$$\begin{aligned}
(i).\ & \left| \sum_{i=1}^n e^{-\langle \boldsymbol{w}, \boldsymbol{z}_i \rangle} - \sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w}, \boldsymbol{z}_i \rangle} \right| < \epsilon \sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w}, \boldsymbol{z}_i \rangle}; \\
(ii).\ & \left\| \sum_{i=1}^n e^{-\langle \boldsymbol{w}, \boldsymbol{z}_i \rangle} \mathcal{P}_\perp(\boldsymbol{z}_i) - \sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w}, \boldsymbol{z}_i \rangle} \mathcal{P}_\perp(\boldsymbol{z}_i) \right\| < \epsilon \sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w}, \boldsymbol{z}_i \rangle}; \\
(iii).\ & \left\| \frac{\sum_{i=1}^n e^{-\langle \boldsymbol{w}, \boldsymbol{z}_i \rangle} \mathcal{P}_\perp(\boldsymbol{z}_i)}{\sum_{j=1}^n e^{-\langle \boldsymbol{w}, \boldsymbol{z}_j \rangle}} - \frac{\sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w}, \boldsymbol{z}_i \rangle} \mathcal{P}_\perp(\boldsymbol{z}_i)}{\sum_{j\in\mathcal{I}} e^{-\langle \boldsymbol{w}, \boldsymbol{z}_j \rangle}} \right\| < \epsilon; \\
(iv).\ & \left\| \frac{\sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w}, \boldsymbol{z}_i \rangle} \mathcal{P}_\perp(\boldsymbol{z}_i)}{\sum_{j\in\mathcal{I}} e^{-\langle \boldsymbol{w}, \boldsymbol{z}_j \rangle}} - \frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} \mathcal{P}_\perp(\boldsymbol{z}_i) \right\| < \epsilon.
\end{aligned}$$

Proof of Lemma C.3.

For simplicity, in this proof, we still denote $\boldsymbol{z}_i := \boldsymbol{x}_i y_i$ ($i \in [n]$) and $\gamma^\star_{\text{sub}} := \min\limits_{i\notin\mathcal{I}} \left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \boldsymbol{z}_i \right\rangle$. And we are given an $\epsilon > 0$. Without loss of generality, we can assume $\delta < 1$.

Proof of (i).

Notice the following two estimates:

$$\sum_{i\notin\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle} = \sum_{i\notin\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle\langle \boldsymbol{w}^\star,\boldsymbol{z}_i\rangle} e^{-\langle \mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_i\rangle}$$
$$\leq \sum_{i\notin\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle\gamma^\star_{\rm sub}} e^{-\langle \mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_i\rangle} \leq \sum_{i\notin\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle\gamma^\star_{\rm sub}} e^{\|\mathcal{P}_\perp(\boldsymbol{w})\|},$$

$$\sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle} = \sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle\langle \boldsymbol{w}^\star,\boldsymbol{z}_i\rangle} e^{-\langle \mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_i\rangle}$$
$$= \sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle\gamma^\star} e^{-\langle \mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_i\rangle} \geq \sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle\gamma^\star} e^{-\|\mathcal{P}_\perp(\boldsymbol{w})\|}.$$

Then for any $\delta > 0$, $R_{>}0$, and $\boldsymbol{w}$ satisfying $\|\mathcal{P}_\perp(\boldsymbol{w})\| < \delta$ and $\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle > R$, we have:

$$\frac{\sum_{i\notin\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}}{\sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}} \leq \frac{\sum_{i\notin\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle\gamma^\star_{\rm sub}} e^{\|\mathcal{P}_\perp(\boldsymbol{w})\|}}{\sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle\gamma^\star} e^{-\|\mathcal{P}_\perp(\boldsymbol{w})\|}}$$
$$\leq \frac{n-|\mathcal{I}|}{|\mathcal{I}|}\exp\left(-\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle(\gamma^\star_{\rm sub}-\gamma^\star) + 2\|\mathcal{P}_\perp(\boldsymbol{w})\|\right)$$
$$\leq \frac{n-|\mathcal{I}|}{|\mathcal{I}|}\exp\left(-\left(1-\frac{\delta^2}{2}\right)\|\boldsymbol{w}\|(\gamma^\star_{\rm sub}-\gamma^\star) + \sqrt{2}\delta\|\boldsymbol{w}\|\right)$$
$$= \frac{n-|\mathcal{I}|}{|\mathcal{I}|}\exp\left(-\|\boldsymbol{w}\|\left(\left(1-\frac{\delta^2}{2}\right)(\gamma^\star_{\rm sub}-\gamma^\star) + \sqrt{2}\delta\right)\right)$$
$$\leq \frac{n-|\mathcal{I}|}{|\mathcal{I}|}\exp\left(-\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle\left(\left(1-\frac{\delta^2}{2}\right)(\gamma^\star_{\rm sub}-\gamma^\star) + \sqrt{2}\delta\right)\right).$$

Due to $\gamma^\star_{\rm sub} - \gamma^\star > 0$, there exist constants $\delta_1 > 0$ and $R_1 > 0$ such that: for any $\boldsymbol{w}$ satisfying $\|\mathcal{P}_\perp(\boldsymbol{w})\| < \delta_1$ and $\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle > R_1$, it holds $\frac{\sum_{i\notin\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}}{\sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}} < \epsilon$, which means (i) holds.

Proof of (ii).

Based on the proof of (i), there exists $\delta_1 > 0$ and $R_1 > 0$ such that: for any $\boldsymbol{w}$ satisfying $\|\mathcal{P}_\perp(\boldsymbol{w})\| < \delta_1$ and $\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle > R_1$, it holds that $\frac{\sum_{i\notin\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}}{\sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}} < \epsilon$, which means

$$\left\|\sum_{i\in[n]} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(z_i) - \sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(z_i)\right\| = \left\|\sum_{i\notin\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(z_i)\right\|$$
$$\leq \sum_{i\notin\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle} < \epsilon \sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}.$$

Proof of (iii).

From the results of (i)(ii), for $\epsilon/2$, there exists $\delta_2 > 0$ and $R_2 > 0$ such that: for any $\boldsymbol{w}$ satisfying $\|\mathcal{P}_\perp(\boldsymbol{w})\| < \delta_2$ and $\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle > R_2$,

$$\left|\sum_{i=1}^n e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle} - \sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}\right| < \frac{\epsilon}{2}\sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle};$$
$$\left\|\sum_{i=1}^n e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i) - \sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)\right\| < \frac{\epsilon}{2}\sum_{i\in\mathcal{I}} e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle};$$

Therefore, we have

$$
\begin{aligned}
&\left\|\frac{\sum_{i\in[n]}e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)}{\sum_{j\in[n]}e^{-\langle \boldsymbol{w},\boldsymbol{z}_j\rangle}}-\frac{\sum_{i\in\mathcal{I}}e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)}{\sum_{j\in\mathcal{I}}e^{-\langle \boldsymbol{w},\boldsymbol{z}_j\rangle}}\right\|\\
\leq&\left\|\frac{\sum_{i\in[n]}e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)}{\sum_{j\in[n]}e^{-\langle \boldsymbol{w},\boldsymbol{z}_j\rangle}}-\frac{\sum_{i\in\mathcal{I}}e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)}{\sum_{j\in[n]}e^{-\langle \boldsymbol{w},\boldsymbol{z}_j\rangle}}\right\|\\
&\quad+\left\|\frac{\sum_{i\in\mathcal{I}}e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)}{\sum_{j\in[n]}e^{-\langle \boldsymbol{w},\boldsymbol{z}_j\rangle}}-\frac{\sum_{i\in\mathcal{I}}e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)}{\sum_{j\in\mathcal{I}}e^{-\langle \boldsymbol{w},\boldsymbol{z}_j\rangle}}\right\|\\
\leq&\frac{\epsilon}{2}\frac{\sum_{i\in\mathcal{I}}e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}}{\sum_{j\in[n]}e^{-\langle \boldsymbol{w},\boldsymbol{z}_j\rangle}}+\left\|\sum_{i\in\mathcal{I}}e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)\right\|\left|\frac{\sum_{j\in\mathcal{I}}e^{-\langle \boldsymbol{w},\boldsymbol{z}_j\rangle}-\sum_{j\in[n]}e^{-\langle \boldsymbol{w},\boldsymbol{z}_j\rangle}}{\left(\sum_{j\in\mathcal{I}}e^{-\langle \boldsymbol{w},\boldsymbol{z}_j\rangle}\right)\left(\sum_{j\in[n]}e^{-\langle \boldsymbol{w},\boldsymbol{z}_j\rangle}\right)}\right|\\
\leq&\frac{\epsilon}{2}\frac{\sum_{i\in\mathcal{I}}e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}}{\sum_{j\in[n]}e^{-\langle \boldsymbol{w},\boldsymbol{z}_j\rangle}}+\frac{\epsilon}{2}\frac{\left\|\sum_{i\in\mathcal{I}}e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)\right\|}{\sum_{j\in[n]}e^{-\langle \boldsymbol{w},\boldsymbol{z}_j\rangle}}\\
\leq&\frac{\epsilon}{2}+\frac{\epsilon}{2}=\epsilon/2.
\end{aligned}
$$

Proof of (iv).

There exists $\delta_3>0$ such that: for any $\boldsymbol{w}$ satisfying $\|\mathcal{P}_\perp(\boldsymbol{w})\|<\delta_3$,

$$
\left|e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_i\rangle}-1\right|\leq 2\left|\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_i\rangle\right|<2\left\|\mathcal{P}_\perp(\boldsymbol{w})\right\|<\epsilon/4.
$$

Then we have

$$
\left|\sum_{j\in\mathcal{I}}e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_i\rangle}-|\mathcal{I}|\right|\leq\sum_{j\in\mathcal{I}}\left|e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_i\rangle}-1\right|\leq\epsilon|\mathcal{I}|/4.
$$

Thus, we can derive:

$$
\begin{aligned}
&\left\|\frac{\sum_{i\in\mathcal{I}}e^{-\langle \boldsymbol{w},\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)}{\sum_{j\in\mathcal{I}}e^{-\langle \boldsymbol{w},\boldsymbol{z}_j\rangle}}-\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\mathcal{P}_\perp(\boldsymbol{z}_i)\right\|\\
=&\left\|\frac{\sum_{i\in\mathcal{I}}e^{-\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle\langle\boldsymbol{w}^\star,\boldsymbol{z}_i\rangle}e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)}{\sum_{j\in\mathcal{I}}e^{-\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle\langle\boldsymbol{w}^\star,\boldsymbol{z}_j\rangle}e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_j\rangle}}-\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\mathcal{P}_\perp(\boldsymbol{z}_i)\right\|\\
=&\left\|\frac{e^{-\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle\gamma^\star}\sum_{i\in\mathcal{I}}e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)}{e^{-\langle \boldsymbol{w},\boldsymbol{w}^\star\rangle\gamma^\star}\sum_{j\in\mathcal{I}}e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_j\rangle}}-\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\mathcal{P}_\perp(\boldsymbol{z}_i)\right\|\\
=&\left\|\frac{\sum_{i\in\mathcal{I}}e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_i\rangle}\mathcal{P}_\perp(\boldsymbol{z}_i)}{\sum_{j\in\mathcal{I}}e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_j\rangle}}-\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\mathcal{P}_\perp(\boldsymbol{z}_i)\right\|\\
=&\left\|\sum_{i\in\mathcal{I}}\left(\frac{e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_i\rangle}}{\sum_{j\in\mathcal{I}}e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_j\rangle}}-\frac{1}{|\mathcal{I}|}\right)\mathcal{P}_\perp(\boldsymbol{z}_i)\right\|\\
\leq&\sum_{j\in\mathcal{I}}\left|\frac{e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_i\rangle}}{\sum_{j\in\mathcal{I}}e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_j\rangle}}-\frac{1}{|\mathcal{I}|}\right|\\
\leq&\sum_{j\in\mathcal{I}}\left|\frac{e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_i\rangle}}{\sum_{j\in\mathcal{I}}e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_j\rangle}}-\frac{1}{\sum_{j\in\mathcal{I}}e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_j\rangle}}\right|+\sum_{j\in\mathcal{I}}\left|\frac{1}{\sum_{j\in\mathcal{I}}e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_j\rangle}}-\frac{1}{|\mathcal{I}|}\right|\\
\leq&\sum_{j\in\mathcal{I}}\frac{\epsilon}{4\sum_{j\in\mathcal{I}}e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_j\rangle}}+\sum_{j\in\mathcal{I}}\frac{\epsilon|\mathcal{I}|}{4\sum_{j\in\mathcal{I}}e^{-\langle\mathcal{P}_\perp(\boldsymbol{w}),\boldsymbol{z}_j\rangle}|\mathcal{I}|}
\end{aligned}
$$

$$\leq \sum_{j\in\mathcal{I}} \frac{\epsilon}{4(1-\epsilon)|\mathcal{I}|} + \sum_{j\in\mathcal{I}} \frac{\epsilon}{4(1-\epsilon)|\mathcal{I}|} = \frac{\epsilon}{2(1-\epsilon)} < \epsilon.$$

The final results.

We choose $\delta = \min\{\delta_2, \delta_3\}$ and $R = R_2$. From our proofs above, (i)$\sim$(iv) all hold for any $\boldsymbol{w}$ satisfying $\|\mathcal{P}_\perp(\boldsymbol{w})\| < \delta$ and $\langle \boldsymbol{w}, \boldsymbol{w}^\star\rangle > R$.

□

Lemma C.4 ((Ji et al., 2020)). *Under Assumption 3.1, let $\boldsymbol{w}(t)$ be trained by GD (3) with $\eta \leq 1/2$ starting from $\boldsymbol{w}(0) = \mathbf{0}$, then GD converges to the max-margin direction:*

$$\lim_{t\to+\infty} \frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|} \to \boldsymbol{w}^\star.$$

Lemma C.5 (Theorem 4.3, (Ji & Telgarsky, 2021)). *Under Assumption 3.1 and 5.4,*

(I) (GD). let $\boldsymbol{w}(t)$ be trained by GD (3) with $\eta \leq 1$ starting from $\boldsymbol{w}(0) = \mathbf{0}$. Then $\left\|\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|} - \boldsymbol{w}^\star\right\| = \mathcal{O}(1/\log t)$ and $\|\boldsymbol{w}(t)\| = \Theta(\log t)$.

(II) (NGD) let $\boldsymbol{w}(t)$ be trained by NGD (4) with $\eta \leq 1$ starting from $\boldsymbol{w}(0) = \mathbf{0}$. Then $\left\|\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|} - \boldsymbol{w}^\star\right\| = \mathcal{O}(1/t)$ and $\|\boldsymbol{w}(t)\| = \Theta(t)$.

Theorem C.6 (Theorem 4.4, (Ji & Telgarsky, 2021)). *Under the same conditions in Theorem C.5, let $\boldsymbol{w}(t)$ be trained by NGD with $\eta \leq 1$ starting from $\boldsymbol{w} = \mathbf{0}$. Then*

(i) $\mathcal{L}_\perp(\cdot)$ has a unique minimizer $\boldsymbol{v}^\star$ over $\operatorname{span}\{\mathcal{P}_\perp(\boldsymbol{x}_i) : i \in \mathcal{I}\}$;

(ii) $\mathcal{L}_\perp(\cdot)$ is strongly convex in any bounded set;

(iii) there exists an absolute constant $C > 0$ such that $\|\mathcal{P}_\perp(\boldsymbol{w}(t)) - \boldsymbol{v}^\star\| \leq C,\ \forall t$.

Theorem C.7 (Theorem 4, (Soudry et al., 2018)). *Under Assumption 3.1 and 5.4, let $\boldsymbol{w}(t)$ be trained by GD (3) with $\eta \leq 1$ starting from $\boldsymbol{w}(0) = \mathbf{0}$. If we denote $\boldsymbol{\rho}(t) = \boldsymbol{w}(t) - \boldsymbol{w}^\star \log t$, then*

$$\lim_{t\to+\infty} \boldsymbol{\rho}(t) = \tilde{\boldsymbol{w}},$$

where $\tilde{\boldsymbol{w}}$ is the solution to the equations: $\eta \exp\left(-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i y_i\rangle\right) = \alpha_i, i \in \mathcal{I}$.

D. Useful Inequalities

Lemma D.1. *(i) For any $x \geq 0$, $\sqrt{1+x} \leq 1 + \frac{x}{2}$; (ii) For any $0 \leq x \leq 1/3$, $\sqrt{1+x} \geq 1 + \frac{x}{3}$.*

Lemma D.2. *For a fixed $\gamma \in (0,1)$, consider the function $h(x) = x + 2(1-\gamma^2)\left(\frac{2}{1+e^x} - 1\right), x \in \mathbb{R}$. Then $h'(x) > 0$ holds for any $x \in \mathbb{R}$.*

Proof of Lemma D.2.

$$h'(x) = 1 - \frac{4(1-\gamma^2)e^x}{(1+e^x)^2} \geq 1 - \frac{4(1-\gamma^2)e^x}{(2e^{x/2})^2} = \gamma^2 > 0,\ \forall x \in \mathbb{R}.$$

□

Lemma D.3 ((Bubeck et al., 2015)). *Let the function $f(\cdot) : \mathcal{S} \to \mathbb{R}$ be both L-smooth and μ-strongly convex on some convex set $\mathcal{S} \subset \mathbb{R}^d$ (in terms of ℓ_2 norm). If the minimizer $\boldsymbol{w}^\star \in \operatorname{int}(\mathcal{S})$ with $f^\star = f(\boldsymbol{w}^\star)$, then for any $\boldsymbol{w}, \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{S}$, it holds that:*

$$f(\boldsymbol{w}_1) \leq f(\boldsymbol{w}_2) + \langle \nabla f(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2\rangle + \frac{L}{2}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|^2;$$

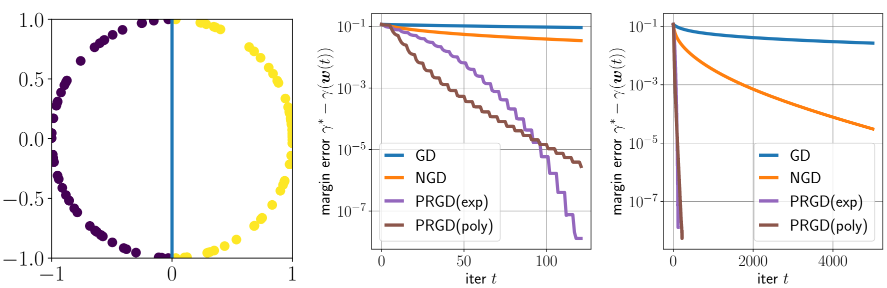
$$f(\boldsymbol{w}_1) \geq f(\boldsymbol{w}_2) + \langle \nabla f(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2 \rangle + \frac{\mu}{2} \left\| \boldsymbol{w}_1 - \boldsymbol{w}_2 \right\|^2 ;$$
$$f(\boldsymbol{w}) - f^\star \geq \frac{\mu}{2} \left\| \boldsymbol{w} - \boldsymbol{w}^\star \right\| ;$$
$$\left\| \nabla f(\boldsymbol{w}) \right\|^2 \geq 2\mu(f(\boldsymbol{w}) - f^\star).$$

E. Experimental Details

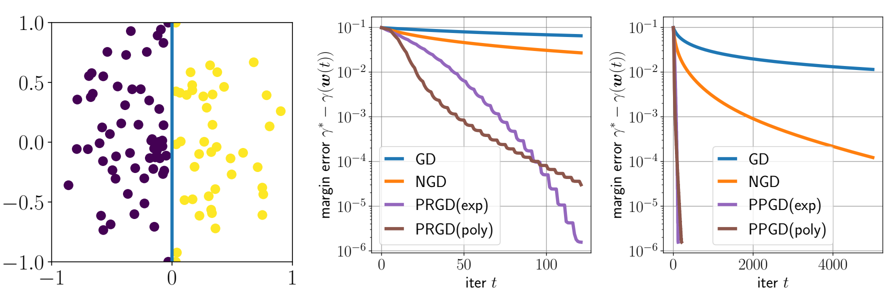
E.1. Experimental details on two synthetic datasets

- Dataset I. We set $\gamma^\star = \sin(\pi/100)$ and $n = 100$. Then we generate the dataset by setting $\boldsymbol{x}_1 = (\gamma^\star, \sqrt{1-\gamma^{\star 2}})$, $\boldsymbol{x}_2 = (-\gamma^\star, \sqrt{1-\gamma^{\star 2}})$, and generate $\boldsymbol{x}_i \sim \text{Unif}\left(\mathbb{S}^1 \cap \{\boldsymbol{x} : |x_1| \geq \gamma^\star\}\right)$ randomly for $i \geq 3$. As for the label, we set $y_i = \text{sgn}(x_1)$.
- Dataset II. We set $\gamma^\star = \sin(\pi/100)$ and $n = 100$. Then we generate the dataset by setting $\boldsymbol{x}_1 = (\gamma^\star, \sqrt{1-\gamma^{\star 2}})$, $\boldsymbol{x}_2 = (-\gamma^\star, \sqrt{1-\gamma^{\star 2}})$, and generate $\boldsymbol{x}_i \sim \text{Unif}\left(\mathbb{B}(0,1) \cap \{\boldsymbol{x} : |x_1| \geq \gamma^\star\}\right)$ randomly for $i \geq 3$. As for the label, we also set $y_i = \text{sgn}(x_1)$.
- PRGD. We follow the guidelines provided in Theorem 6.1. For the Warm-up phase, we use GD as the Warm-up Phase for 1000 iterations, and then turn it to PRGD. For the second Phase, we employ PRGD(exp) with hyperparameters $T_{k+1} - T_k \equiv 5$, $R_k = R_0 \times 1.2^k$. To illustrate the role of the progressive radius, we also examine PRGD(poly) configured with $T_{k+1} - T_k \equiv 5$, $R_k = R_0 \times k^{1.2}$, where the progressive radius increases polynomially.

The numerical results and comparison of different algorithms are shown in Figure 6 and Table 2.



(a) Dataset I



(b) Dataset II

Figure 6: (The detailed version of Figure 3) Comparison of margin Maximization rates of different algorithms on two synthetic datasets. (left) The visualization of two 2d synthetic dataset. The yellow points represent the data with label 1, while the purple points corresponds to the data with label 1; (middle)(right) The comparison of margin maximization rates of different algorithms on the corresponding dataset at small and large time scales, respective

Table 2: The number of iterations needed to achieve the same margin error on the sythetic datasets.

	GD	NGD	**PRGD(exp)**	PRGD(poly)
margin error `1e-6`, Dataset I	$+\infty$	12,508	**106**	142
margin error `1e-4`, Dataset II	$+\infty$	5,027	**94**	95

E.2. Experiments Details for VGG on CIFAR-10

Following Lyu & Li (2019), we examine our algorithm for the homogenized VGG-16. We explored the performance of the VGG-based neural network for image classification tasks on CIFAR-10.

Our experimental setup involved a modified VGG architecture implemented in PyTorch. The homogeneity requires that the bis term exists at most in the first layer (Lyu & Li, 2019). Specifically, the network's architecture comprised multiple convolutional layers without the bias terms, followed by ReLU activations and max pooling. The classifier section consisted of three fully connected layers with ReLU activations and dropout, excluding bias in linear transformations.

The network was trained using a batch size of 64, with the option to enable CUDA for GPU acceleration. Weight initialization was conducted using Kaiming normalization for convolutional layers and a uniform distribution for linear layers. The model was trained and evaluated using a custom DataLoader for both the training and test datasets. We used a base learning rate of 1×10^{-3}, a momentum of 0.9, and a weight decay of 5×10^{-4}. For the loss-based learning rate used in NGD and PRGD, we use the strategy in Lyu & Li (2019). Additionally, for $\|\boldsymbol{\theta}\|$ in the PRGD regime, we use the ℓ_2 norm of the parameters of all layers. And we configured PRGD with $T_k = 3000 \times \left(2 + \frac{k}{3000}\right)^3$, $R_k = \min\left(R_0 \times \left(2 + \frac{k}{3000}\right)^{0.2}, 1000\right)$.