

---

# Implicit Bias of Spectral Descent and Muon on Multiclass Separable Data

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 Different gradient-based methods for optimizing overparameterized models can all  
2 achieve zero training error yet converge to distinctly different solutions inducing  
3 different generalization properties. We provide the first complete characterization  
4 of implicit optimization bias for p-norm normalized steepest descent (NSD) and  
5 momentum steepest descent (NMD) algorithms in multi-class linear classification  
6 with cross-entropy loss. Our key theoretical contribution is proving that these algo-  
7 rithms converge to solutions maximizing the margin with respect to the classifier  
8 matrix’s p-norm, with established convergence rates. These results encompass  
9 important special cases including Spectral Descent and Muon, which we show  
10 converge to max-margin solutions with respect to the spectral norm. A key insight  
11 of our contribution is that the analysis of general entry-wise and Schatten p-norms  
12 can be reduced to the analysis of NSD/NMD with max-norm by exploiting a natural  
13 ordering property between all p-norms relative to the max-norm and its dual sum-  
14 norm. For the specific case of descent with respect to the max-norm, we further  
15 extend our analysis to include preconditioning, showing that Adam converges  
16 to the matrix’s max-norm solution. Our results demonstrate that the multi-class  
17 linear setting, which is inherently richer than the binary counterpart, provides  
18 the most transparent framework for studying implicit biases of matrix-parameter  
19 optimization algorithms.

## 20 1 Introduction

21 The ever-increasing training cost of large language models (LLMs) has demanded better optimizer  
22 designs with improved performance and efficiency [9, 1, 19]. The de facto standard optimizers for  
23 deep learning training are Adam and AdamW [28, 32]. However, these algorithms that employ  
24 diagonal preconditioners to independently adjust the learning rate of each coordinate, may fail to  
25 capture their inter-dependencies and fully leverage the geometry of the loss landscape [62]. This  
26 has spurred a series of research efforts on improving Adam or AdamW’s computational efficiency  
27 [16, 20, 42, 63], with LLM-training as the target application domain [27, 51, 38, 30].

28 A noticeable work by Jordan et al. [27] proposed the Muon optimizer, which was shown to have  
29 remarkable performances on NanoGPT benchmarks. More recently, it has been shown that Muon can  
30 be used for large-scale LLM training with the potential to replace AdamW as the standard choice [30].  
31 The key step in Muon is to orthogonalize the updates via the Newton-Schulz iteration [27, 5]. More  
32 precisely, the update (denoted as  $\Delta$ ) is (approximately) replaced by the product of its singular-vector  
33 matrices  $UV^T$  (where the (truncated) singular value decomposition (SVD) of  $\Delta$  is  $\Delta = U\Sigma V^T$ ).  
34 Even though the benefits of orthogonalization are not fully understood, Jordan et al. [27] pointed out  
35 that it could promote updates in directions of small magnitudes given the weight matrices are typically  
36 low-rank. Moreover, if the above SVD approximation is exact and gradient accumulations are turned

Table 1: Summary of margin convergence rates for NSD and NMD algorithms of different norm constraints for linear multiclass separable data with the CE loss. The (truncated) SVDs of the gradient and momentum are denoted as  $\nabla = U\Sigma V^T$  and  $M = \tilde{U}\tilde{\Sigma}\tilde{V}^T$  respectively.

Method	Norm Constraint	Update $\Delta$	Reference	Rate <sup>2</sup>
NGD	Unit $\ \cdot\ _2$ -ball	$\frac{\nabla}{\ \nabla\ _2}$	Hazan et al. [21]	-
NMD-GD		$\frac{M}{\ M\ _2}$	Cutkosky and Mehta [12]	-
SignGD	Unit $\ \cdot\ _{\max}$ -ball	$\text{sign}(\nabla)$	Bernstein et al. [6]	-
Signum		$\text{sign}(M)$	Bernstein et al. [6]	-
<i>Spectral-GD</i>	Unit $   \cdot   _{\infty}$ -ball	$UV^T$	Bernstein and Newhouse [5]	$O(\frac{\log t+n}{t^{1/2}})$
<i>Muon</i> <sup>1</sup>		$\tilde{U}\tilde{V}^T$	Jordan et al. [27]	$O(\frac{d \log t+dn}{t^{1/2}})$

<sup>1</sup> We consider EMA-style momentum of the form (5).

<sup>2</sup> NGD and SignGD rates are the same as Spectral-GD; Signum and NMD-GD rates are the same as Muon.

off, then Muon becomes spectral descent [11, 5], which is the (normalized) steepest descent w.r.t the spectral norm [5]. As noted by Bernstein and Newhouse [5], spectral descent is also Shampoo (which won the AlgoPerf competition [43, 13]) without accumulations in preconditioners. Thus, Muon can be viewed as (approximate) Shampoo when both optimizers are without accumulations. In essence, we observe that one important ingredient of Muon or Shampoo (without accumulations) is the spectral-descent step of the following:

$$W^\dagger = W - \eta UV^T \quad \text{where} \quad \nabla \mathcal{L}(W) = U\Sigma V^T.$$

Theoretical investigations of spectral descent or Muon mainly focus on characterizing the convergence rates of the algorithm (e.g., the rate of decrease of the gradient norm in the non-convex setting [2, 29, 38]). However, modern machine learning models are overparameterized, leading to multiple weight configurations that achieve identical training loss but exhibit markedly different generalization properties [61, 4]. The key insight is that gradient-based methods inherently prefer “simple” solutions according to optimizer-specific notions of simplicity. Understanding this implicit bias/regularization requires analyzing not just loss convergence, but the geometric trajectory of parameter updates throughout training. To this end, our work aims to address the fundamental question:

*What is the **implicit bias** of **spectral descent** (and its momentum variants) in linear multiclass classification with separable data and cross-entropy loss?*

The multiclass setting where the parameter is a **matrix**, is a natural place to study the class of spectral-descent algorithms, and provides an inherently richer setting. Our work captures this richness by establishing convergence with respect to not only entry-wise matrix norms, but also matrix Schatten norms. Hence, while the focus is on spectral descent and Muon, the analysis establishes implicit bias rates for a wide family of algorithms (Table 1), and we state the results in the most general form from the perspective of steepest descent with (unit) norm-ball constraints. Our contributions are:

1. For multiclass separable data trained with the cross-entropy (CE) loss, we show that the iterates of normalized steepest descent (NSD) defined with respect to (w.r.t.) any matrix entry-wise or Schatten norms converge to a solution that maximizes the margin defined w.r.t. the same norm, with a rate  $\mathcal{O}(\frac{1}{t^{1/2}})$ . This includes sign descent (entry-wise max-norm) [6], normalized gradient descent (entry-wise 2-norm) [21], and spectral descent (Schatten  $\infty$ -norm) [5] as special cases. To achieve this, we introduce a unified analysis framework that relates entry-wise and Schatten p-norms to the entry-wise max-norm, and construct a proxy function for the loss that closely traces both its value and gradient. We also show the same machinery applies to other multiclass losses such as the exponential loss [34] and the PairLogLoss [55].
2. Under the same setting, we utilize the same framework and the same proxy function to show that the same  $\mathcal{O}(\frac{1}{t^{1/2}})$  margin convergence rate holds for normalized momentum steepest descent (NMD). This includes the following algorithms in analogy to the ones above: sign momentum descent [6], normalized momentum gradient descent [12], and Muon [27]. The key step of the analysis is to use the proxy function to bound the sum-norm difference between the gradient and the momentum (i.e., the exponential moving averages (EMA) of the gradient), which translates to a bound on the dual norm through the fundamental norm-relationships used in the study of NSD.

75 The margin convergence rates of various algorithms are summarized in Table 1. Furthermore, we  
 76 extend the analysis to Adam (without the stability constant) and show its iterates maximize the  
 77 margin w.r.t. the matrix max-norm (proof details and numerical validations in App. G).

78 3. We experimentally verify our theoretical predictions across all considered algorithms. First, for  
 79 sign descent (SignGD) and Signum, we demonstrate that solutions favor the max-norm margin  
 80 over the 2-norm margin—the opposite behavior to normalized gradient descent (NGD) and  
 81 normalized momentum gradient descent (NMD-GD). Moreover, we show that both spectral  
 82 descent (Spectral-GD) and Muon favor the spectral-norm margin over the other norms.

## 83 2 Preliminaries

84 **Notations** Matrices, vectors, and scalars are denoted by  $\mathbf{A}$ ,  $\mathbf{a}$ , and  $a$  respectively. For matrix  $\mathbf{A}$ , we  
 85 denote its  $(i, j)$ -th entry as  $\mathbf{A}[i, j]$ , and for vector  $\mathbf{a}$ , its  $i$ -th entry as  $\mathbf{a}[i]$  or  $a_i$ . We consider entry-wise  
 86 matrix p-norms defined as  $\|\mathbf{A}\|_p = (\sum_{i,j} |\mathbf{A}[i, j]|^p)^{1/p}$ . Central to our results are: the infinity norm,  
 87 denoted as  $\|\mathbf{A}\|_{\max} := \|\mathbf{A}\|_{\infty} = \max_{i,j} |\mathbf{A}[i, j]|$  and called the **max-norm**, and the **entry-wise**  
 88 **1-norm**, denoted as  $\|\mathbf{A}\|_{\text{sum}} := \|\mathbf{A}\|_1 = \sum_{i,j} |\mathbf{A}[i, j]|$ . The entry-wise 1-norm is dual to the max-  
 89 norm. For vectors, the max-norm is equivalent to the infinity norm, denoted as  $\|\mathbf{a}\|_{\infty}$ , while we denote  
 90 the  $\ell_1$  norm as  $\|\mathbf{a}\|_1$ . We further denote the Schatten p-norm of  $\mathbf{A}$  as  $|||\mathbf{A}|||_p := (\sum_{i=1}^r \sigma_i^p)^{1/p}$ ,  
 91 where  $\sigma_1, \sigma_2, \dots, \sigma_r$  are the non-zero singular values of  $\mathbf{A}$ . Let  $r = \text{rank}(\mathbf{A})$ , then special cases of  
 92 Schatten p-norm include: nuclear norm  $|||\mathbf{A}|||_1 = \sum_{i=1}^r \sigma_i$ , Frobenius norm  $|||\mathbf{A}|||_2 = \sqrt{\sum_{i=1}^r \sigma_i^2}$ ,  
 93 and **spectral norm**  $|||\mathbf{A}|||_{\infty} = \sigma_1$ . To simplify the discussions, we sometimes write  $\|\mathbf{A}\|$  (dropping  
 94 subscripts) to refer to any entry-wise or Schatten p-norm with  $p \geq 1$ . We denote by  $\|\mathbf{A}\|_*$  the  
 95 dual-norm with respect to the standard matrix inner product  $\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr}(\mathbf{A}^\top \mathbf{B})$ . We denote the  
 96 gradient and its value at iteration  $t$  as  $\nabla := \nabla \mathcal{L}(\mathbf{W})$  and  $\nabla_t := \nabla \mathcal{L}(\mathbf{W}_t)$  respectively.

97 Let  $\mathbb{S} : \mathbb{R}^k \rightarrow \Delta^{k-1}$  the softmax map of  $k$ -dimensional vectors to the probability simplex  $\Delta^{k-1}$   
 98 such that for any  $\mathbf{a} \in \mathbb{R}^k$ , it holds that  $\mathbb{S}(\mathbf{a}) = [\frac{\exp(\mathbf{a}[c])}{\sum_{c \in [k]} \exp(\mathbf{a}[c])}]_{c=1}^k \in \Delta^{k-1}$ . Let  $\mathbb{S}_c(\mathbf{v})$  denote the  
 99  $c$ -th entry of  $\mathbb{S}(\mathbf{v})$ . Let  $\mathbb{S}'(\mathbf{a}) = \text{diag}(\mathbb{S}(\mathbf{a})) - \mathbb{S}(\mathbf{a})\mathbb{S}(\mathbf{a})^\top$  denote the softmax gradient, with  $\text{diag}(\cdot)$   
 100 a diagonal matrix. Finally, let  $\{\mathbf{e}_c\}_{c=1}^k$  be the standard basis vectors of  $\mathbb{R}^k$ , and indicator  $\delta_{ij}$  be such  
 101 that  $\delta_{ij} = 1$  if and only if  $i = j$ . For any integer  $k$ ,  $[k]$  denotes  $\{1, \dots, k\}$ .

102 **Setup** Consider a multiclass classification problem with training data  $\mathbf{h}_1, \dots, \mathbf{h}_n$  and labels  
 103  $y_1, \dots, y_n$ . Each datapoint  $\mathbf{h}_i \in \mathbb{R}^d$  is a vector in a  $d$ -dimensional embedding space (denote  
 104 data matrix  $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_n]^\top \in \mathbb{R}^{n \times d}$ ), and each label  $y_i \in [k]$  represents one of  $k$  classes. We  
 105 assume each class contains at least one datapoint. The classifier  $f_{\mathbf{W}} : \mathbb{R}^d \rightarrow \mathbb{R}$  is a linear model with  
 106 weight matrix  $\mathbf{W} \in \mathbb{R}^{k \times d}$ . The model outputs logits  $\ell_i = f_{\mathbf{W}}(\mathbf{h}_i) = \mathbf{W}\mathbf{h}_i$  for  $i \in [n]$ , which are  
 107 passed through the softmax map to produce class probabilities  $\hat{p}(c|\mathbf{h}_i) = \mathbb{S}_c(\ell_i)$ .

108 We train using empirical risk minimization (ERM):  $\mathcal{L}_{\text{ERM}}(\mathbf{W}) := -\frac{1}{n} \sum_{i \in [n]} \ell(\mathbf{W}\mathbf{h}_i; y_i)$ , where  
 109 the loss function  $\ell$  takes as input the logits of a datapoint and its label. The predominant choice in  
 110 classification is the CE loss

$$\mathcal{L}(\mathbf{W}) := -\frac{1}{n} \sum_{i \in [n]} \log(\mathbb{S}_{y_i}(\mathbf{W}\mathbf{h}_i)). \quad (1)$$

111 We focus our discussions on the CE loss due to its ubiquity in practice. However, our results hold for  
 112 other multiclass losses such as the exponential [34] and the PairLogLoss [55] (see App. F). Define  
 113 the maximum margin of the dataset w.r.t. any entry-wise or Schatten p-norm  $\|\cdot\|$  as

$$\gamma := \max_{\|\mathbf{W}\| \leq 1} \min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^\top \mathbf{W}\mathbf{h}_i. \quad (2)$$

114 **Optimization Methods** We study iterative algorithms that update the weight matrix via:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta_t \Delta_t.$$

115 For the NSD family [8], the update direction<sup>1</sup> w.r.t. the norm  $\|\cdot\|$  is:

$$\Delta_t := \arg \max_{\|\Delta\| \leq 1} \langle \nabla_t, \Delta \rangle. \quad (3)$$

<sup>1</sup>For  $p \in (1, \infty)$ , the norms  $\|\cdot\|_p$  and  $|||\cdot|||_p$  are strictly convex, thus there is a unique maximizer defining the update in Eqn. (3). For  $p = 1, \infty$  the maximizer is not necessarily unique and our results hold for any choice of  $\Delta_t$  in the set of maximizers; see e.g. Ziętak [64].

Note that this reduces to SignGD, Coordinate Descent (e.g., Nutini et al. [36]), or NGD when the max-norm (i.e.  $\|\cdot\|_\infty$ ), the entry-wise 1-norm (i.e.  $\|\cdot\|_{\text{sum}}$ ), or the Frobenius Euclidean-norm (i.e.  $\|\cdot\|_2$ ) is used, respectively. Concretely, the update directions for SignGD and NGD are:

$$\text{SignGD: } \Delta_t = \text{sign}(\nabla_t), \text{ and NGD: } \Delta_t = \nabla_t / \|\nabla_t\|_2,$$

where the  $\text{sign}(\cdot)$  and division  $\div$  operations are applied entry-wise. In the special case of spectral norm (i.e.  $\|\cdot\|_\infty$ ), this becomes the Spectral-GD, for which  $\Delta_t = U_t V_t^T$ , where  $U_t$  and  $V_t$  are the left/right singular matrices of  $\nabla_t$  respectively (i.e.,  $\nabla_t = U_t \Sigma_t V_t^T$  with singular values in  $\Sigma_t > 0$  arranged in non-increasing order). Finally, note that the Schatten 2-norm case reduces to NGD (as  $\|\cdot\|_2 = \|\cdot\|_2$ ).

We also consider the NMD family with the following update direction w.r.t. the norm  $\|\cdot\|$

$$\Delta_t := \arg \max_{\|\Delta\| \leq 1} \langle M_t, \Delta \rangle, \quad (4)$$

where the momentum  $M_t$  is computed as the EMA of the gradient given by

$$M_t = \beta_1 M_{t-1} + (1 - \beta_1) \nabla_t. \quad (5)$$

This form of momentum is also known as the heavy-ball or the SGDM-style momentum [39, 15, 31]. Thus, an NMD algorithm chooses the update direction (among all feasible directions in the unit  $\|\cdot\|$ -ball) that best aligns with the momentum instead of the gradient direction (as chosen by an NSD algorithm). Similar to above, when the max-norm and the Frobenius-norm are used, the resulting Signum and NMD-GD update directions are:

$$\text{Signum: } \Delta_t = \text{sign}(M_t), \text{ and NMD-MD: } \Delta_t = M_t / \|M_t\|_2.$$

When spectral norm is used in (4), this becomes Muon<sup>2</sup> for which the SVD is on  $M_t$  (i.e.  $M_t = \tilde{U}_t \tilde{\Sigma}_t \tilde{V}_t^T$ ) and the update direction is  $\Delta_t = \tilde{U}_t \tilde{V}_t^T$ . Note that Muon reduces to Spectral-GD when the momentum parameter  $\beta_1$  is set to 0 (similar reductions hold for Signum (to SignGD) and NMD-GD (to NGD) as well).

**Assumptions** Establishing the implicit bias of the above mentioned gradient-based optimization algorithms, requires the following assumptions. First, we assume data are linearly separable, ensuring the margin  $\gamma$  is strictly positive, an assumption routinely used in previous works [44, 40, 18, 35, 58].

**Assumption 1.** *There exists  $W \in \mathbb{R}^{k \times d}$  such that  $\min_{c \neq y_i} (e_{y_i} - e_c)^T W h_i > 0$  for all  $i \in [n]$ .*

In this work, we consider learning rate schedule  $\eta_t = \Theta(\frac{1}{t^a})$ , where  $a \in (0, 1]$ . Such schedule has been studied in the convergence and implicit bias of various optimization algorithms (e.g., Bottou et al. [7], Nacson et al. [35], and Sun et al. [46]) including Adam [60].

**Assumption 2.** *The learning rate schedule  $\{\eta_t\}$  is decreasing with respect to  $t$  and satisfies the following conditions:  $\lim_{t \rightarrow \infty} \eta_t = 0$  and  $\sum_{t=0}^{\infty} \eta_t = \infty$ .*

Assumption 3 can be satisfied by the above learning rate for a sufficiently large  $t$  as shown in Zhang et al. [60, Lemma C.1]. It is used in our analysis of NMD and Adam.

**Assumption 3.** *The learning rate schedule satisfies the following: let  $\beta \in (0, 1)$  and  $c_1 > 0$  be two constants, there exist time  $t_0 \in \mathbb{N}_+$  and constant  $c_2 = c_2(c_1, \beta) > 0$  such that  $\sum_{s=0}^t \beta^s (e^{c_1 \sum_{\tau=1}^s \eta_{s-\tau}} - 1) \leq c_2 \eta_t$  for all  $t \geq t_0$ .*

Finally, we assume that the 1-norm of the data is bounded. Similar assumptions were used in Ji and Telgarsky [22], Nacson et al. [35], Wu et al. [58], and Zhang et al. [60].

**Assumption 4.** *There exists constant  $B > 0$  such that  $\|h_i\|_1 \leq B$  for all  $i \in [n]$ .*

### 3 A Unified Framework with a Proxy Function

Analyzing margin convergence begins with studying loss convergence through second-order Taylor expansion of the CE loss (recall that  $S'(v) = \text{diag}(v) - vv^\top$ ):

$$\mathcal{L}(W + \Delta) = \mathcal{L}(W) + \langle \nabla \mathcal{L}(W), \Delta \rangle + \frac{1}{2n} \sum_{i \in [n]} h_i^\top \Delta^\top S'(W h_i) \Delta h_i + o(\|\Delta\|_F^3), \quad (6)$$

<sup>2</sup>The implementation in Jordan et al. [27] uses Nesterov-type momentum: Newton-Schulz iteration applied to  $\beta_1 M_t + \nabla_t$  instead of  $\beta_1 M_{t-1} + \nabla_t$  [30].

155 To bound the loss at  $\mathbf{W}_{t+1} = \mathbf{W}_t - \eta_t \Delta_t$ , we must bound both terms in (6). For NSD updates in Eq.  
 156 (3), the first term evaluates to  $-\eta_t \|\nabla \mathcal{L}(\mathbf{W})\|_*$  (recall that  $\|\cdot\|_*$  is the dual norm). This leads to two  
 157 key tasks: (1) Lower-bounding the dual gradient norm; (2) Upper-bounding the second-order term.

158 For the proof to proceed, these bounds should satisfy two desiderata: (1) They are expressible as the  
 159 same function of  $\mathbf{W}$ , call it  $\mathcal{G}(\mathbf{W})$ , up to constants. (2) The function  $\mathcal{G}(\mathbf{W})$  is a good proxy for the  
 160 loss for small values of the latter. The former helps with combining the terms, while the latter helps  
 161 with demonstrating descent. Next, we obtain these key bounds for the CE loss by determining the  
 162 appropriate proxy  $\mathcal{G}(\mathbf{W})$ .

163 Besides the need for a proxy  $\mathcal{G}(\mathbf{W})$ , we use the following facts about the sum-norm dominating any  
 164 entry-wise/Schatten  $p$ -norm. Concretely, for any matrix  $\mathbf{A}$  and any  $p \geq 1$ :

$$\|\mathbf{A}\|_{\max} \leq \|\mathbf{A}\|_p \leq \|\mathbf{A}\|_{\text{sum}}, \quad \text{and} \quad \|\mathbf{A}\|_{\max} \leq \|\mathbf{A}\|_p \leq \|\mathbf{A}\|_{\text{sum}}. \quad (7)$$

165 These relationships (proved in Lemma 11 in App. C) are crucial for unifying the analysis of NSD  
 166 and NMD algorithms w.r.t. either the entry-wise or the Schatten norms (details below).

167 **Construction of  $\mathcal{G}(\mathbf{W})$**  Before showing our construction for the CE loss, it is insightful to discuss  
 168 how previous works do this in the binary case with labels  $y_{b,i} \in \{\pm 1\}$ , classifier vector  $\mathbf{w} \in \mathbb{R}^d$   
 169 and binary margin  $\gamma_b := \max_{\|\mathbf{w}\| \leq 1} \min_{i \in [n]} y_{b,i} \mathbf{w}^\top \mathbf{h}_i$ . For exponential loss, Gunasekar et al.  
 170 [18] showed that  $\|\nabla \mathcal{L}(\mathbf{w})\| \geq \gamma_b \mathcal{L}(\mathbf{w})$ . For logistic loss  $\ell(t) = \log(1 + \exp(-t))$ , Zhang et al.  
 171 [60] proved  $\|\nabla \mathcal{L}(\mathbf{w})\|_1 \geq \gamma_b \mathcal{G}(\mathbf{w})$ , where  $\mathcal{G}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n |\ell'(y_{b,i} \mathbf{w}^\top \mathbf{h}_i)|$  and  $\ell'$  is the first-order  
 172 derivative. In both cases, one can take the common form  $\mathcal{G}_b(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n |\ell'(y_{b,i} \mathbf{w}^\top \mathbf{h}_i)|$ . The  
 173 proof relies on showing  $\gamma \leq \min_{\mathbf{r} \in \Delta^{n-1}} \|\mathbf{H}^T \mathbf{r}\|$  via Fenchel Duality [48, 18] and appropriately  
 174 choosing  $\mathbf{r}$ .

175 In the multiclass setting, where the loss function is vector-valued, it is unclear how to extend the binary  
 176 proof or definition of  $\mathcal{G}(\mathbf{W})$ . To this end, we realize that the key is in the proper manipulation of the  
 177 gradient inner product  $\langle \mathbf{A}, -\nabla \mathcal{L}(\mathbf{W}) \rangle$  (for arbitrary matrix  $\mathbf{A} \in \mathbb{R}^{k \times d}$ ). The CE gradient evaluates to  
 178  $\nabla \mathcal{L}(\mathbf{W}) = \frac{1}{n} \sum_{i=1}^n (\mathbf{e}_{y_i} - \mathbb{S}(\mathbf{W} \mathbf{h}_i)) \mathbf{h}_i^\top$  and using the fact that  $\mathbb{S}(\mathbf{W} \mathbf{h}_i) \in \Delta^{k-1}$ , it turns out that we  
 179 can express (details in Lemma 9):  $\langle \mathbf{A}, -\nabla \mathcal{L}(\mathbf{W}) \rangle = \frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} \mathbb{S}_c(\mathbf{W} \mathbf{h}_i) (\mathbf{e}_{y_i} - \mathbf{e}_c)^\top \mathbf{A} \mathbf{h}_i$ .  
 180 This motivates defining  $\mathcal{G}(\mathbf{W})$  as:

$$\mathcal{G}(\mathbf{W}) := \frac{1}{n} \sum_{i \in [n]} (1 - \mathbb{S}_{y_i}(\mathbf{W} \mathbf{h}_i)). \quad (8)$$

181 The lemma below, following from the inner-product calculation above and our definition of  $\mathcal{G}(\mathbf{W})$ ,  
 182 confirms this is the right choice. For convenience, denote  $s_{ic} := \mathbb{S}_c(\mathbf{W} \mathbf{h}_i)$ , for  $i \in [n], c \in [k]$ .

183 **Lemma 1** (Lower bounding the gradient dual-norm). *For any  $\mathbf{W} \in \mathbb{R}^{k \times d}$  and any entry-wise or  
 184 Schatten  $p$ -norm  $\|\cdot\|$  with  $p \geq 1$ , it holds that  $\|\nabla \mathcal{L}(\mathbf{W})\|_* \geq \gamma \cdot \mathcal{G}(\mathbf{W})$ , where  $\|\cdot\|_*$  is the dual-norm.*

185 The lemma completes the first task: lower bounding the gradient's dual norm. Importantly, the factor  
 186 in front of  $\mathcal{G}(\mathbf{W})$  is the margin  $\gamma$  w.r.t. the norm  $\|\cdot\|$ , which is crucial in the forthcoming analysis.

187  **$\mathcal{G}(\mathbf{W})$  and second-order term** We now show how to bound the second-order term in (6). For this,  
 188 we establish the following essential lemma whose proof relies on the relationships in (7).

189 **Lemma 2.** *For any entry-wise or Schatten  $p$ -norm  $\|\cdot\|$  with  $p \geq 1$ , any  $\mathbf{s} \in \Delta^{k-1}$  in the  $k$ -dimensional  
 190 simplex, any index  $c \in [k]$ , and  $\mathbf{v} \in \mathbb{R}^k$ , it holds that*

$$\mathbf{v}^\top (\text{diag}(\mathbf{s}) - \mathbf{s} \mathbf{s}^\top) \mathbf{v} \leq 4(1 - s_c) \|\mathbf{v} \mathbf{v}^\top\|.$$

191 *Proof.* Let  $\mathbf{S} := \text{diag}(\mathbf{s}) - \mathbf{s} \mathbf{s}^\top$  and  $q \geq 1$  such that  $1/p + 1/q = 1$ . By norm duality, it holds that

$$\mathbf{v}^\top \mathbf{S} \mathbf{v} = \text{tr}(\mathbf{S} \mathbf{v} \mathbf{v}^\top) \leq \|\mathbf{S}\|_q \|\mathbf{v} \mathbf{v}^\top\| \leq \|\mathbf{S}\|_{\text{sum}} \|\mathbf{v} \mathbf{v}^\top\|,$$

192 where  $\|\cdot\|_q$  is the dual of  $\|\cdot\|$  and the second inequality is by (7). Direct calculation yields  $\|\mathbf{S}\|_{\text{sum}} =$   
 193  $2 \sum_{c \in [k]} s_c (1 - s_c)$ . The advertised bound then follows by noting the following  $\sum_{c \in [k]} s_c (1 - s_c) \leq$   
 194  $2(1 - s_{c'})$  for any  $c' \in [k]$  (verified in Lemma 13 in App. C).  $\square$



Next, we apply the above lemma with  $\mathbf{v} \leftarrow \Delta \mathbf{h}_i$  and  $c \leftarrow y_i$ , and further use the inequalities:  
 $\|\mathbf{v}\mathbf{v}^T\|_p = \|\mathbf{v}\|_p^2 \leq \|\Delta\|_p^2 \|\mathbf{h}\|_q^2$  for entry-wise norms and  $\|\mathbf{v}\mathbf{v}^T\|_p = \|\mathbf{v}\|_2^2 \leq \|\Delta\|_\infty^2 \|\mathbf{h}\|_2^2 \leq \|\Delta\|_p^2 \|\mathbf{h}\|_2^2$  for Schatten norms. Together with Ass. 4, this upper bounds the second-order term in the CE loss expansion in terms of the proxy function:

$$2B^2 \|\Delta\|^2 \cdot \frac{1}{n} \sum_{i \in [n]} (1 - \mathbb{S}_{y_i}(\mathbf{W}\mathbf{h}_i)).$$

**Properties of  $\mathcal{G}(\mathbf{W})$**  We now show that  $\mathcal{G}(\mathbf{W})$  meets the second desiderata: being a good proxy for the loss  $\mathcal{L}(\mathbf{W})$ . This is rooted in the elementary relationships between  $\mathcal{G}(\mathbf{W})$  and  $\mathcal{L}(\mathbf{W})$ , which are used in the various parts of the proof. Below, we summarize these key relationships.

**Lemma 3** (Properties of  $\mathcal{G}(\mathbf{W})$  and  $\mathcal{L}(\mathbf{W})$ ). *Let  $\mathbf{W} \in \mathbb{R}^{k \times d}$ . The followings hold: (i) Under Ass. 4,  $2B \cdot \mathcal{G}(\mathbf{W}) \geq \|\nabla \mathcal{L}(\mathbf{W})\|_*$ ; (ii)  $1 \geq \frac{\mathcal{G}(\mathbf{W})}{\mathcal{L}(\mathbf{W})} \geq 1 - \frac{n\mathcal{L}(\mathbf{W})}{2}$ ; (iii) If  $\mathbf{W}$  satisfies  $\mathcal{L}(\mathbf{W}) \leq \frac{\log 2}{n}$  or  $\mathcal{G}(\mathbf{W}) \leq \frac{1}{2n}$ , then  $\mathcal{L}(\mathbf{W}) \leq 2\mathcal{G}(\mathbf{W})$ .*

Lemma 3 (i) extends Lemma 1 by establishing a sandwich relationship between  $\mathcal{G}(\mathbf{W})$  and the gradient's dual norm. The lemma's statements (ii) and (iii) show that  $\mathcal{G}(\mathbf{W})$  can substitute for the loss - it lower bounds  $\mathcal{L}(\mathbf{W})$  and serves as an upper bound when either  $\mathcal{L}(\mathbf{W})$  or  $\mathcal{G}(\mathbf{W})$  is sufficiently small. Specifically, the ratio  $\mathcal{G}(\mathbf{W})/\mathcal{L}(\mathbf{W})$  converges to 1 as the loss decreases, with the convergence rate depending on the rate of loss decrease. The key property (ii) may seem algebraically complex, but it turns out (details in Lemma 18 in App. C) that both sides of the sandwich relationship follow from the elementary fact that  $\forall x > 0 : 1 - x \leq e^{-x} \leq 1 - x + x^2/2$ .

## 4 Implicit Bias of Normalized Steepest Descent

We now leverage our construction of  $\mathcal{G}(\mathbf{W})$  to show that the margin of NSD's iterates converges to the data margin defined w.r.t. the same entry-wise or Schatten p-norm that is used to define the algorithm (refer to eqns. (2) and (3) for the definitions of margin and NSD). We only highlight the key steps in the proof and defer details to App. D.

**NSD Descent** We start by showing a descent property. By applying Lemmas 1 and 2 to lower and upper bound the first and second order terms in Eq. (6) yields:

$$\mathcal{L}(\mathbf{W}_{t+1}) \leq \mathcal{L}(\mathbf{W}_t) - \gamma \eta_t \mathcal{G}(\mathbf{W}_t) + 2\eta_t^2 B^2 \mathcal{G}(\mathbf{W}_t) \sup_{\zeta \in [0,1]} \frac{\mathcal{G}(\mathbf{W}_t - \zeta \eta_t \Delta_t)}{\mathcal{G}(\mathbf{W}_t)}.$$

Algebraic manipulations of the definition of  $\mathcal{G}(\mathbf{W})$  and the relationships in (7) allow us to bound the ratio in the right hand side.

**Lemma 4** (Ratio of  $\mathcal{G}(\mathbf{W})$ ). *For any  $\psi \in [0, 1]$ , we have the following:  $\frac{\mathcal{G}(\mathbf{W} - \psi \eta \Delta)}{\mathcal{G}(\mathbf{W})} \leq e^{2B\eta\psi\|\Delta\|_{\max}} \leq e^{2B\eta\psi\|\Delta\|}$  (note that the second inequality is by (7)).*

From this and  $\|\Delta_t\| \leq 1$  for NSD, we obtain

$$\mathcal{L}(\mathbf{W}_{t+1}) \leq \mathcal{L}(\mathbf{W}_t) - \gamma \eta_t (1 - \alpha_{s_1} \eta_t) \mathcal{G}(\mathbf{W}_t), \quad (9)$$

where  $\alpha_{s_1} = 2B^2 e^{2B\eta_0} / \gamma$ . Given a decay learning rate of the form  $\eta_t = \Theta(\frac{1}{t^a})$ , we can conclude that the loss starts to monotonically decrease after some time.

**NSD Unnormalized Margin** We now use the descent property in (9) to lower bound the unnormalized margin. An intermediate result towards this is recognizing that sufficiently small loss  $\mathcal{L}(\mathbf{W}) \leq \frac{\log 2}{n}$  guarantees  $\mathbf{W}$  separates the data (Lemma 19 in App. C). The descent property ensures that NSD iterates will eventually achieve this loss threshold, thereby guaranteeing separability. The main result of this section, shows that eventually the iterates achieve separability with a substantial (unnormalized) margin.

**Lemma 5** (NSD Unnormalized Margin). *Assume there exists  $\tilde{t}$  such that  $\mathcal{L}(\mathbf{W}_t) \leq \frac{\log 2}{n}, \forall t > \tilde{t}$ . Then, it holds that for all  $t \geq \tilde{t}$  ( $\alpha_{s_2} = 2B^2 e^{2B\eta_0}$ )*

$$\min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i \geq \gamma \sum_{s=\tilde{t}}^{t-1} \eta_s \frac{\mathcal{G}(\mathbf{W}_s)}{\mathcal{L}(\mathbf{W}_s)} - \alpha_{s_2} \sum_{s=\tilde{t}}^{t-1} \eta_s^2. \quad (10)$$

234 **NSD Margin Convergence** Proceeding from Eq. (10) requires showing the convergence of the  
 235 ratio  $\frac{\mathcal{G}(\mathbf{W})}{\mathcal{L}(\mathbf{W})}$ . The two key ingredients are given in Lemma 3 (ii) and (iii). Lemma 3 (ii) suggests that  
 236 it is sufficient to study the convergence of  $\mathcal{L}(\mathbf{W})$ , which is captured in (9). However, to obtain an  
 237 explicit rate via (9), we need to rewrite  $\mathcal{G}(\mathbf{W}_t)$  in terms of  $\mathcal{L}(\mathbf{W}_t)$ . This is where Lemma 3 (iii) helps.  
 238 Putting them together, we arrive at the following theorem (see Thm. 3 and Cor. 1 for details).

239 **Theorem 1.** *Suppose that Ass. 1, 2, and 4 hold. Set learning rate  $\eta_t = \Theta(\frac{1}{t^{1/2}})$ . The following holds*  
 240 *for the margin gap of NSD's iterates*

$$\gamma - \frac{\min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i}{\|\mathbf{W}_t\|} \leq \mathcal{O}\left(\frac{\log t + n}{t^{1/2}}\right).$$

241 **Remark 1.** *For margin convergence rates of NSD, Nacson et al. [35] showed a rate of  $\mathcal{O}(\frac{\log t}{t^{1/2}})$*   
 242 *in the binary setting, limited to the entry-wise  $p$ -norms and the exponential loss. Compared to*  
 243 *this, our results hold for the more practical setting of multiclass data and CE loss. To the best of*  
 244 *our knowledge, this is the first non-asymptotic result on the implicit bias of spectral-GD for linear*  
 245 *multiclass separable data, and it holds for other  $p$ -norms as well. Upon completion of this work, we*  
 246 *became aware of an update on the arXiv version of Tsilivis et al. [49], which includes an extension*  
 247 *of their previous results to steepest descent w.r.t. the spectral norm. In comparison to ours, their*  
 248 *gradient-flow analysis applies to homogeneous neural networks with the restriction of infinitesimal*  
 249 *step-sizes. Moreover, it does not include normalization nor momentum (like Muon, which we analyze),*  
 250 *and the convergence is (asymptotic) to a KKT point of a spectral-norm margin maximization problem.*

## 251 5 Implicit Bias of Normalized Momentum Steepest Descent

252 In this section, we study the implicit bias of NMD algorithms (proof details in App E). Similar to Sec.  
 253 4, its updates are defined w.r.t. either the entry-wise or the Schatten norm  $\|\cdot\|$ . The analysis relies on  
 254 the relationships in (7) and we show the same proxy function  $\mathcal{G}(\mathbf{W})$  naturally appears. Given that  
 255 the NMD updates satisfy  $\|\Delta\| \leq 1$ , the second-order term in (6) is bounded in the same way as NSD.  
 256 The main difference is in bounding the first-order term as shown by the following lemma.

257 **Lemma 6.** *Let  $\Omega_t := \mathbf{M}_t - \nabla_t$ . It holds for all  $t \geq 0$  that*

$$\langle \nabla_t, \mathbf{W}_{t+1} - \mathbf{W}_t \rangle \leq 2\eta \|\Omega_t\|_* - \eta \gamma \mathcal{G}(\mathbf{W}_t).$$

258 Given the relationships in (7) hold for any  $p \geq 1$ , we can bound the dual norm of  $\|\Omega_t\|_*$  via its  
 259 sum norm (i.e.  $\|\Omega_t\|_* \leq \|\Omega_t\|_{\text{sum}}$ ). Given the goal is to bound all the terms in the Taylor expansion  
 260 (6) via the proxy function  $\mathcal{G}(\mathbf{W}_t)$ , an natural next step is to bound  $\|\Omega_t\|_{\text{sum}}$  using the same proxy  
 261 function. To do this, we decompose the proxy function **per-class-wise**, and apply the per-class  
 262 proxy functions to bound the entries of  $\Omega_t$  associated with their corresponding classes. Concretely,  
 263 we write the function  $\mathcal{G}(\mathbf{W})$  in two equivalent ways:  $\mathcal{G}(\mathbf{W}) = \sum_{c \in [k]} \frac{1}{n} \sum_{i \in [n], y_i = c} (1 - s_{iy_i}) =$   
 264  $\sum_{c \in [k]} \frac{1}{n} \sum_{i \in [n], y_i \neq c} s_{ic}$ , which motivate the following definitions of the per-class proxy functions:

$$\mathcal{G}_c(\mathbf{W}) := 1/n \sum_{i \in [n], y_i = c} (1 - s_{iy_i}), \quad \text{and} \quad \mathcal{Q}_c(\mathbf{W}) := 1/n \sum_{i \in [n], y_i \neq c} s_{ic}.$$

265 Next, we bound the entries in each row of  $\Omega_t$  (thus belonging to the same class) via the corresponding  
 266 proxy functions  $\mathcal{G}_c(\mathbf{W}_t)$  and  $\mathcal{Q}_c(\mathbf{W}_t)$  to arrive at the following lemma. Its proof utilizes the nice  
 267 properties of softmax map given in Lemma 15 in App. B.

268 **Lemma 7.** *Suppose that Ass. 1, 2, 3, and 4 hold. Let  $c \in [k]$  and  $j \in [d]$ . There exists time  $t_0$  such*  
 269 *that for all  $t \geq t_0$  and for  $\alpha_M := B(1 - \beta_1)c_2$ :*

$$|\mathbf{M}_t[c, j] - (1 - \beta_1^{t+1}) \nabla \mathcal{L}(\mathbf{W}_t)[c, j]| \leq \alpha_M \eta_t (\mathcal{G}_c(\mathbf{W}_t) + \mathcal{Q}_c(\mathbf{W}_t)).$$

270 Given the result in Lemma 7, we can show that  $|\Omega_t[c, j]| \leq \beta_1^{t+1} |\nabla \mathcal{L}(\mathbf{W}_t)[c, j]| + \alpha_M \eta_t \mathcal{T}_c(\mathbf{W}_t)$ ,  
 271 where  $\mathcal{T}_c(\mathbf{W}_t)$  is defined to be  $\mathcal{T}_c(\mathbf{W}_t) := \mathcal{G}_c(\mathbf{W}_t) + \mathcal{Q}_c(\mathbf{W}_t)$ . Then, we sum over indices  $c \in [k]$  and  
 272  $j \in [d]$  and apply  $\|\nabla\|_{\text{sum}} \leq 2B \cdot \mathcal{G}(\mathbf{W})$  (from Lemma 3 (i)) to obtain:

273 **Lemma 8.** *It holds for all  $t \geq 0$  that  $\|\Omega_t\|_{\text{sum}} \leq 2B\beta_1^{t/2} \mathcal{G}(\mathbf{W}_t) + 2\alpha_M d \eta_t \mathcal{G}(\mathbf{W}_t)$ .*

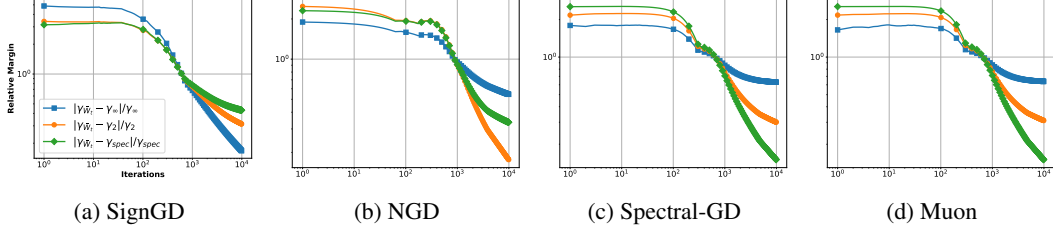


Figure 1: **(a)** We normalize the iterates of SignGD w.r.t. the max-norm (denoted as  $\bar{W}_t$ ), compute the margin (denoted as  $\gamma_{\bar{W}_t}$ ), then plot its difference to data margins  $\gamma_{\|\cdot\|_\infty}$ ,  $\gamma_{\|\cdot\|_2}$ , and  $\gamma_{\|\cdot\|_{\text{spec}}}$  (note that the margin difference is further divided by the corresponding data margin for comparisons). SignGD favors the margin defined w.r.t. the max-norm. **(b, c, and d)** Same as (a) with SignGD (max-norm) replaced by NGD (2-norm), Spectral-GD (spectral-norm), and Muon (spectral-norm) respectively. NGD favors the 2-norm margin, while Spectral-GD and Muon favor the spectral-norm margin.

274 This completes the bound on the first-order term for NMD algorithms via the proxy  $\mathcal{G}(\bar{W})$ . The rest  
 275 proof follows similar steps as NSD. We note that without the above per-class decomposition, an extra  
 276  $k$ -factor would appear in the second term of the bound on  $\|\Omega_t\|_{\text{sum}}$  (and thus also show up in the  
 277 final rate). We state the main theorem for NMD algorithms.

278 **Theorem 2.** *Under the setting of Lem. 7, the margin gap of NMD with  $\eta_t = \Theta(\frac{1}{t^{1/2}})$  is  $O(\frac{d \log t + dn}{t^{1/2}})$ .*

279 **Remark 2.** Wang et al. [53] studied implicit bias of un-normalized GD with momentum, and showed  
 280 its iterates converge asymptotically to the max 2-norm margin solution. In contrast, our rates are  
 281 non-asymptotic and cover a much wider family of algorithms converging to non-Euclidean geometric  
 282 margins (w.r.t. entry-wise/Schatten norms). Note the convergence rate of NMD matches that of NSD  
 283 (Thm. 1) up to a factor of  $d$ . It could be interesting to remove this dependence in a future work.

284 **Implicit Bias of Adam** Finally, observing that Adam [28] (without the stability constant, i.e., eqns  
 285 (33a), (33b), and (33c) in App. G) shares the same form of momentum as NMD and the (entry-wise)  
 286 updates are bounded by some constant as shown in Zhang et al. [63] and Xie and Li [59]. Thus, our  
 287 analysis extends to Adam. Concretely, a similar proof strategy can be adapted once a bound on the  
 288 second gradient moment via the proxy function is established (Lemma 35). In App. G, we prove a  
 289  $O(\frac{d \log(t) + nd}{t^{1/3}})$  max-norm margin convergence rate for Adam (details in Thm. 5 and Cor. 2).

## 290 6 Experiments

291 We generate synthetic multiclass separable data as follows:  $k = 10$  class centers are sampled  
 292 from a standard normal distribution; within each class, data is sampled from normal distribution  
 293  $\mathcal{N}(0, \sigma^2 I)$ ,  $\sigma = 0.1$ . We set  $d = 25$ , sample 50 data points for each class, and ensure that margin  
 294 is positive (thus data is separable). We run different algorithms to minimize CE loss using  $\eta_t = \frac{\eta_0}{t^a}$   
 295 ( $\eta_0 = 0.1$  for SignGD and NGD;  $\eta_0 = 0.05$  for Spectral-GD and Muon), where (based on our  
 296 theorems)  $a$  is set to  $1/2$ . We apply truncated SVD on the gradient and momentum for Spectral-GD  
 297 and Muon respectively. Data margins w.r.t. different norms are found via CVXPY [14]. We denote  
 298 max-margin classifiers defined w.r.t. the 2-norm, the max-norm, and the spectral-norm as  $V_2$ ,  $V_\infty$ ,  
 299 and  $V_{\text{spec}}$  respectively. Based on the margin-gap results in Figure 1, we observe that SignGD, NGD,  
 300 and Spectral-GD favor max-norm, 2-norm, and spectral-norm margin respectively. Besides this, the  
 301 behavior of Muon is very similar to that of Spectral-GD (in agreement with our theories). Figure  
 302 2 further confirms that the iterates of these algorithms correlate well with the corresponding max  
 303 margin separators. Experiments on Signum, NMD-GD, and Adam are provided in App. A and App.  
 304 G. The experiments are run using an AMD-Ryzen-9-5900X processor.

## 305 7 Related Works

306 Starting with GD, the foundational result by Soudry et al. [44] showed that gradient descent op-  
 307 timization of logistic loss on linearly separable data converges in direction to the  $L_2$  max-margin  
 308 classifier at a rate  $O(1/\log(t))$ . Contemporaneous work by Ji and Telgarsky [22] generalized this by



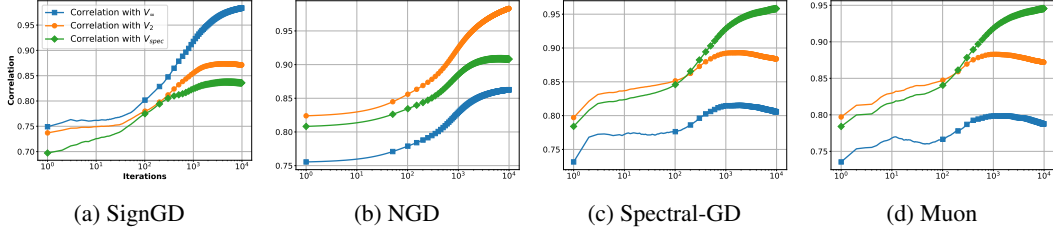


Figure 2: **(a)** Correlations between the iterates of SignGD ( $\mathbf{W}_t$ ) and max margin separators  $\mathbf{V}_\infty$ ,  $\mathbf{V}_2$ , and  $\mathbf{V}_{\text{spec}}$  against iterations (correlation defined as  $\frac{\langle \mathbf{W}_t, \mathbf{V} \rangle}{\|\mathbf{W}_t\|_2 \|\mathbf{V}\|_2}$ ). **(b, c, and d)** Same as (a) with SignGD replaced by NGD, Spectral-GD, and Muon respectively. SignGD and NGD correlate well with  $\mathbf{V}_\infty$  and  $\mathbf{V}_2$  respectively, while Spectral-GD and Muon correlate well with  $\mathbf{V}_{\text{spec}}$ .

relaxing the data separability requirement. Ji et al. [25] later connected these findings to earlier work on regularization paths of logistic loss minimization [41], which enabled extensions to other loss functions (e.g., those with polynomial tail decay). More recently, Wu et al. [58] extends these results to the large step size regime with the same  $O(1/\log(t))$  rate. The relatively slow convergence rate to the max-margin classifier motivated investigation into adaptive step-sizes. Nacson et al. [35] showed that NGD with decaying step-size  $\eta_t = 1/\sqrt{t}$  achieves  $L_2$ -margin convergence at rate  $O(1/\sqrt{t})$ . This rate was improved to  $O(1/t)$  by Ji and Telgarsky [24] using constant step-sizes, and further to  $O(1/t^2)$  through a specific momentum formulation [26]. Besides linear classifications, implicit bias of GD has been studied for least squares [17, 18, 3], homogeneous [33, 23, 57] or non-homogeneous neural networks [10], and matrix factorization [17]; see Vardi [50] for a survey. Moving onto the multiclass setting, Ravi et al. [40] extended the implicit bias result of Soudry et al. [44] to multiclass classification for losses with exponential tails, including CE, multiclass exponential, and PairLogLoss. Their approach leverages a framework of Wang and Scott [56] that allows multiclass losses and separability conditions to be written in margin-based forms similar to binary cases.

Beyond GD, Gunasekar et al. [18] and Nacson et al. [35] showed that steepest descent w.r.t. entry-wise p-norms yields updates that in the limit maximize the margin w.r.t the same norm. Sun et al. [45] showed that the iterates of mirror descent with the potential function chosen as the p-th power of the p-norm converge to the classifier that maximizes the margin w.r.t. the p-norm. In both cases, the convergence rate is slow at  $O(1/\log(t))$ . Wang et al. [54] further improved the rates for both steepest descent and mirror descent when  $p \in (1, 2]$ . Note that all these results apply only to the exponential loss. More recently, Tsilivis et al. [49] showed that the iterates of steepest descent algorithms converge to a KKT point of a generalized margin maximization problem in homogeneous neural networks. Moreover, the implicit bias of Adam (with or without the stability constant) has been studied in both linear and non-linear settings. Wang et al. [52] demonstrated the normalized iterates of Adam (with non-negligible stability constant) converge to a KKT point of a  $L_2$ -margin maximization problem for homogeneous neural networks. Zhang et al. [60] studied the implicit bias of Adam without the stability constant on (linearly) binary separable data. They showed that unlike GD, the Adam’s iterates converge to a solution that maximizes the margin w.r.t the  $L_\infty$ -norm. The study of excluding the stability constant is also the focus of another recent work on the implicit bias of AdamW [59], where the authors again establish that convergence aligns with the  $L_\infty$  geometry.

## 8 Conclusion

We have characterized the margin convergence rates of Spectral-GD and Muon for multiclass linear separable data. Given they are special cases of NSD and NMD w.r.t the spectral norm, the analysis is done on a wider scale by studying NSD/NMD w.r.t any entry-wise or Schatten p-norms. Thus, the rates also hold for optimizers of other geometries, such as the sign-descent (max-norm) or gradient-descent (2-norm) family. We further extend the analysis to Adam using the same framework. Future directions include removing the factor- $d$  from the bound of NMD, obtaining a tighter convergence rate for Adam, and studying other related algorithms such as Shampoo that involves non-diagonal preconditioners. It is also important to extend our results to nonlinear models such as diagonal neural nets [37], self-attention [47] and homogeneous neural nets [33, 49], helping further bridge the gap to deep learning practices.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Kang An, Yuxing Liu, Rui Pan, Shiqian Ma, Donald Goldfarb, and Tong Zhang. Asgo: Adaptive structured gradient optimization. *arXiv preprint arXiv:2503.20762*, 2025.
- [3] Navid Azizan, Sahin Lale, and Babak Hassibi. Stochastic mirror descent on overparameterized nonlinear models. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12): 7717–7727, 2021.
- [4] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- [5] Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.
- [6] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd: Compressed optimisation for non-convex problems. *arXiv preprint arXiv:1802.04434*, 2018.
- [7] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.
- [8] Stephen Boyd and Lieven Vandenbergh. *Convex optimization*. Cambridge university press, 2004.
- [9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [10] Yuhang Cai, Kangjie Zhou, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter L Bartlett. Implicit bias of gradient descent for non-homogeneous deep networks. *arXiv preprint arXiv:2502.16075*, 2025.
- [11] David E Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned spectral descent for deep learning. *Advances in neural information processing systems*, 28, 2015.
- [12] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized sgd. In *International conference on machine learning*, pages 2260–2268. PMLR, 2020.
- [13] George E Dahl, Frank Schneider, Zachary Nado, Naman Agarwal, Chandramouli Shama Sastry, Philipp Hennig, Sourabh Medapati, Runa Eschenhagen, Priya Kasimbeg, Daniel Suo, et al. Benchmarking neural network training algorithms. *arXiv preprint arXiv:2306.07179*, 2023.
- [14] Steven Diamond and Stephen Boyd. Cvxpy: A python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.
- [15] Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE, 2015.
- [16] Roger Grosse and James Martens. A kronecker-factored approximate fisher matrix for convolution layers. In *International Conference on Machine Learning*, pages 573–582. PMLR, 2016.
- [17] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

- [18] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.
- [19] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shiron Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [20] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018.
- [21] Elad Hazan, Kfir Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. *Advances in neural information processing systems*, 28, 2015.
- [22] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on learning theory*, pages 1772–1798. PMLR, 2019.
- [23] Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.
- [24] Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.
- [25] Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pages 2109–2136. PMLR, 2020.
- [26] Ziwei Ji, Nathan Srebro, and Matus Telgarsky. Fast margin maximization via dual acceleration. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2021.
- [27] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL <https://kellerjordan.github.io/posts/muon/>.
- [28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [29] Jiaxiang Li and Mingyi Hong. A note on the convergence of muon and further. *arXiv preprint arXiv:2502.02900*, 2025.
- [30] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.
- [31] Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems*, 33:18261–18271, 2020.
- [32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [33] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.
- [34] Indraneel Mukherjee and Robert E Schapire. A theory of multiclass boosting. *Advances in Neural Information Processing Systems*, 23, 2010.
- [35] Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.
- [36] Julie Nutini, Mark Schmidt, Issam Laradji, Michael Friedlander, and Hoyt Koepke. Coordinate descent converges faster with the gauss-southwell rule than random selection. In *International Conference on Machine Learning*, pages 1632–1641. PMLR, 2015.

- [37] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.
- [38] Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos. *arXiv preprint arXiv:2502.07529*, 2025.
- [39] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.
- [40] Hrithik Ravi, Clayton Scott, Daniel Soudry, and Yutong Wang. The implicit bias of gradient descent on separable multiclass data. *arXiv preprint arXiv:2411.01350*, 2024.
- [41] Saharon Rosset, Ji Zhu, and Trevor J. Hastie. Margin maximizing loss functions. In *NIPS*, 2003.
- [42] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018.
- [43] Hao-Jun Michael Shi, Tsung-Hsien Lee, Shintaro Iwasaki, Jose Gallego-Posada, Zhijing Li, Kaushik Rangadurai, Dheevatsa Mudigere, and Michael Rabbat. A distributed data-parallel pytorch implementation of the distributed shampoo optimizer for training neural networks at-scale. *arXiv preprint arXiv:2309.06497*, 2023.
- [44] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.
- [45] Haoyuan Sun, Kwangjun Ahn, Christos Thrampoulidis, and Navid Azizan. Mirror descent maximizes generalized margin and can be implemented efficiently. *Advances in Neural Information Processing Systems*, 35:31089–31101, 2022.
- [46] Haoyuan Sun, Khashayar Gatmiry, Kwangjun Ahn, and Navid Azizan. A unified approach to controlling implicit regularization via mirror descent. *Journal of Machine Learning Research*, 24(393):1–58, 2023.
- [47] Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines, 2023.
- [48] Matus Telgarsky. Margins, shrinkage, and boosting. In *International Conference on Machine Learning*, pages 307–315. PMLR, 2013.
- [49] Nikolaos Tsilivis, Gal Vardi, and Julia Kempe. Flavors of margin: Implicit bias of steepest descent in homogeneous neural networks. *arXiv preprint arXiv:2410.22069*, 2024.
- [50] Gal Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6):86–93, 2023.
- [51] Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321*, 2024.
- [52] Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *International Conference on Machine Learning*, pages 10849–10858. PMLR, 2021.
- [53] Bohan Wang, Qi Meng, Huishuai Zhang, Ruoyu Sun, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. Does momentum change the implicit regularization on separable data? *Advances in Neural Information Processing Systems*, 35:26764–26776, 2022.
- [54] Guanghui Wang, Zihao Hu, Vidya Muthukumar, and Jacob D Abernethy. Faster margin maximization rates for generic optimization methods. *Advances in Neural Information Processing Systems*, 36:62488–62518, 2023.

- 488 [55] Nan Wang, Zhen Qin, Le Yan, Honglei Zhuang, Xuanhui Wang, Michael Bendersky, and  
489 Marc Najork. Rank4class: a ranking formulation for multiclass classification. *arXiv preprint*  
490 *arXiv:2112.09727*, 2021.
- 491 [56] Yutong Wang and Clayton Scott. Unified binary and multiclass margin-based classification.  
492 *Journal of Machine Learning Research*, 25(143):1–51, 2024.
- 493 [57] Jingfeng Wu, Peter L Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for  
494 logistic loss: Non-monotonicity of the loss improves optimization efficiency. *arXiv preprint*  
495 *arXiv:2402.15926*, 2024.
- 496 [58] Jingfeng Wu, Vladimir Braverman, and Jason D Lee. Implicit bias of gradient descent for  
497 logistic regression at the edge of stability. *Advances in Neural Information Processing Systems*,  
498 36, 2024.
- 499 [59] Shuo Xie and Zhiyuan Li. Implicit bias of adamw: L-infinity norm constrained optimization.  
500 *arXiv preprint arXiv:2404.04454*, 2024.
- 501 [60] Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of adam on separable data. *arXiv*  
502 *preprint arXiv:2406.10650*, 2024.
- 503 [61] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding  
504 deep learning requires rethinking generalization, 2017.
- 505 [62] Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhiquan Luo. Why  
506 transformers need adam: A hessian perspective. *Advances in Neural Information Processing*  
507 *Systems*, 37:131786–131823, 2024.
- 508 [63] Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik P Kingma, Yinyu  
509 Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. *arXiv*  
510 *preprint arXiv:2406.16793*, 2024.
- 511 [64] Krystyna Ziętak. On the characterization of the extremal points of the unit sphere of matrices.  
512 *Linear Algebra and its Applications*, 106:57–75, 1988.



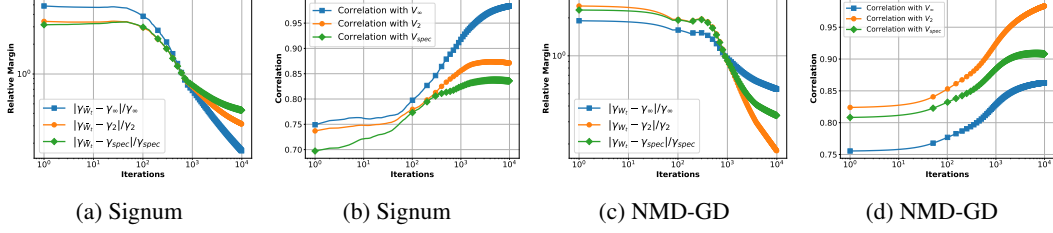


Figure 3: Implicit bias of Signum and NMD-GD on multiclass separable data. (a) Relative margin gap of Signum’s iterates against iterations. (b) Correlation of Signum’s iterates to  $V_{\infty}$ ,  $V_2$ , and  $V_{\text{spec}}$  against iterations. See Figure 1 and 2 for the definitions of relative margin and correlation. (c) and (d) Same as (a) and (b) with Signum replaced by NMD-GD.

## A Additional Experiments

We present additional experiments on Signum and NMD-GD in this section. Based on Figure 3, their margin convergence properties are very similar to those of SignGD and NGD respectively (see Sec. 6 in the main text for the experimental setup).

## B Facts about CE loss and Softmax

Lemma 9 is on the gradient of the cross-entropy loss. It will be used for showing the form of  $\mathcal{G}(\mathbf{W})$  in (8) lower bounds  $\|\nabla \mathcal{L}(\mathbf{W})\|$  in Lemma 16.

**Lemma 9** (Gradient). *Let CE loss*

$$\mathcal{L}(\mathbf{W}) := -\frac{1}{n} \sum_{i \in [n]} \log(\mathbb{S}_{y_i}(\mathbf{W}\mathbf{h}_i)).$$

For any  $\mathbf{W}$ , it holds

- $\nabla \mathcal{L}(\mathbf{W}) = -\frac{1}{n} \sum_{i \in [n]} (\mathbf{e}_{y_i} - \mathbf{s}_i) \mathbf{h}_i^\top = -\frac{1}{n} (\mathbf{Y} - \mathbf{S}) \mathbf{H}^\top$
- $\mathbb{1}_k^\top \nabla \mathcal{L}(\mathbf{W}) = 0$
- For any matrix  $\mathbf{A} \in \mathbb{R}^{k \times d}$ ,

$$\begin{aligned} \langle \mathbf{A}, -\nabla \mathcal{L}(\mathbf{W}) \rangle &= \frac{1}{n} \sum_i (1 - s_{iy_i}) \left( \mathbf{e}_{y_i}^\top \mathbf{A} \mathbf{h}_i - \frac{\sum_{c \neq y_i} s_{ic} \mathbf{e}_c^\top \mathbf{A} \mathbf{h}_i}{(1 - s_{iy_i})} \right) \\ &= \frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} s_{ic} (\mathbf{e}_{y_i} - \mathbf{e}_c)^\top \mathbf{A} \mathbf{h}_i \end{aligned} \quad (11)$$

where we simplify  $\mathbf{S} := \mathbb{S}(\mathbf{W}\mathbf{H}) = [\mathbf{s}_1, \dots, \mathbf{s}_n] \in \mathbb{R}^{k \times n}$ . The last statement yields

$$\langle \mathbf{A}, -\nabla \mathcal{L}(\mathbf{W}) \rangle \geq \frac{1}{n} \sum_{i \in [n]} (1 - s_{iy_i}) \cdot \min_{c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^\top \mathbf{A} \mathbf{h}_i. \quad (12)$$

*Proof.* First bullet is by direct calculation. Second bullet uses the fact that  $\mathbb{1}^\top (\mathbf{y}_i - \mathbf{s}_i) = 1 - 1 = 0$  since  $\mathbb{1}^\top \mathbf{s}_i = 1$ . The third bullet follows by direct calculation and writing  $\mathbf{s}_i^\top \mathbf{A} \mathbf{h}_i = (\sum_c s_{ic} \mathbf{e}_c)^\top \mathbf{A} \mathbf{h}_i = \sum_c s_{ic} \mathbf{e}_c^\top \mathbf{A} \mathbf{h}_i$ .  $\square$

Lemma 10 is on the Taylor expansion of the loss. It will be used in showing the descent properties of NSD and NMD.

**Lemma 10** (Hessian). *Let perturbation  $\Delta \in \mathbb{R}^{k \times d}$  and denote  $\mathbf{W}' = \mathbf{W} + \Delta$ . Then,*

$$\begin{aligned} \mathcal{L}(\mathbf{W}') &= \mathcal{L}(\mathbf{W}) - \frac{1}{n} \sum_{i \in [n]} \langle (\mathbf{e}_{y_i} - \mathbb{S}(\mathbf{W}\mathbf{h}_i)) \mathbf{h}_i^\top, \Delta \rangle \\ &\quad + \frac{1}{2n} \sum_{i \in [n]} \mathbf{h}_i^\top \Delta^\top (\text{diag}(\mathbb{S}(\mathbf{W}\mathbf{h}_i)) - \mathbb{S}(\mathbf{W}\mathbf{h}_i) \mathbb{S}(\mathbf{W}\mathbf{h}_i)^\top) \Delta \mathbf{h}_i + o(\|\Delta\|^3). \end{aligned} \quad (13)$$

532 *Proof.* Define function  $\ell_y : \mathbb{R}^k \rightarrow \mathbb{R}$  parameterized by  $y \in [k]$  as follows:

$$\ell_y(\mathbf{l}) := -\log(\mathbb{S}_y(\mathbf{l})).$$

533 From Lemma 9,

$$\nabla \ell_y(\mathbf{l}) = -(\mathbf{e}_y - \mathbb{S}(\mathbf{l})).$$

534 Thus,

$$\nabla^2 \ell_y(\mathbf{l}) = \nabla \mathbb{S}(\mathbf{l}) = \text{diag}(\mathbb{S}(\mathbf{l})) - \mathbb{S}(\mathbf{l})\mathbb{S}(\mathbf{l})^\top$$

535 Combining these the second-order taylor expansion of  $\ell_y$  writes as follows for any  $\mathbf{l}, \boldsymbol{\delta} \in \mathbb{R}^k$ :

$$\ell_y(\mathbf{l} + \boldsymbol{\delta}) = \ell_y(\mathbf{l}) - (\mathbf{e}_y - \mathbb{S}(\mathbf{l}))^\top \boldsymbol{\delta} + \frac{1}{2} \boldsymbol{\delta}^\top (\text{diag}(\mathbb{S}(\mathbf{l})) - \mathbb{S}(\mathbf{l})\mathbb{S}(\mathbf{l})^\top) \boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|^3).$$

536 To evaluate this with respect to a change on the classifier parameters, set  $\mathbf{l} = \mathbf{W}\mathbf{h}$  and  $\boldsymbol{\delta} = \boldsymbol{\Delta}\mathbf{h}$  for  
537  $\boldsymbol{\Delta} \in \mathbb{R}^{k \times d}$ . Denoting  $\mathbf{W}' = \mathbf{W} + \boldsymbol{\Delta}$ , we then have

$$\ell_y(\mathbf{W}') = \ell_y(\mathbf{W}) - \langle (\mathbf{e}_y - \mathbb{S}(\mathbf{l}))\mathbf{h}^\top, \boldsymbol{\Delta} \rangle + \frac{1}{2} \mathbf{h}^\top \boldsymbol{\Delta}^\top (\text{diag}(\mathbb{S}(\mathbf{l})) - \mathbb{S}(\mathbf{l})\mathbb{S}(\mathbf{l})^\top) \boldsymbol{\Delta} \mathbf{h} + o(\|\boldsymbol{\Delta}\|^3).$$

538 This shows the desired since  $n\mathcal{L}(\mathbf{W}) := \sum_{i \in [n]} \ell_{y_i}(\mathbf{W}\mathbf{h}_i)$  and we can further obtain

$$\ell_y(\mathbf{W}') = \ell_y(\mathbf{W}) - \langle (\mathbf{e}_y - \mathbb{S}(\mathbf{l}))\mathbf{h}^\top, \boldsymbol{\Delta} \rangle + \frac{1}{2} \mathbf{h}^\top \boldsymbol{\Delta}^\top (\text{diag}(\mathbb{S}(\mathbf{l}')) - \mathbb{S}(\mathbf{l}')\mathbb{S}(\mathbf{l}')^\top) \boldsymbol{\Delta} \mathbf{h}, \quad (14)$$

539 where  $\mathbf{l}' = \mathbf{l} + \zeta \boldsymbol{\delta}$  for some  $\zeta \in [0, 1]$ .  $\square$

540 We prove the relationships in (7), which are useful for unifying the analysis of entry-wise and Schatten  
541 norms.

542 **Lemma 11.** For any matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  and any entry-wise or Schatten  $p$ -norm  $\|\cdot\|$  with  $p \geq 1$ , it  
543 holds that

$$\|\mathbf{A}\|_{\max} \leq \|\mathbf{A}\| \leq \|\mathbf{A}\|_{\text{sum}}.$$

544 *Proof.* The entry-wise  $p$ -norm case is trivial. Here, we focus the Schatten  $p$ -norm case. Note that  
545  $\|\mathbf{A}\|_2$  coincides with the entrywise 2-norm  $\|\mathbf{A}\|_2$ , but in general Schatten norms are different from  
546 entry-wise norms. On the other hand, Schatten norms preserve the ordering of norms. Specifically,  
547 for any  $p \geq 1$ , it holds:

$$\|\mathbf{A}\|_\infty = \sigma_1 \leq \|\mathbf{A}\|_p = \left( \sum_{i=1}^r \sigma_i^p \right)^{1/p} \leq \sum_{i=1}^r \sigma_i = \|\mathbf{A}\|_1. \quad (15)$$

548 It is also well-known that

$$\|\mathbf{A}\|_\infty = \max_{\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1} \mathbf{u}^\top \mathbf{A} \mathbf{v} \geq \max_{i,j} |\mathbf{A}[i,j]| = \|\mathbf{A}\|_{\max} \quad (16)$$

549 where the inequality follows by selecting  $\mathbf{u} = \text{sign}(\mathbf{A}[i',j']) \cdot \mathbf{e}_{i'}$  and  $\mathbf{v} = \mathbf{e}_{j'}$  for  $(i',j')$  such that  
550  $|\mathbf{A}[i',j']| = \|\mathbf{A}\|_{\max}$  and  $\mathbf{e}_{i'}, \mathbf{e}_{j'}$  corresponding basis vectors.

551 Using this together with duality, it also holds that

$$\|\mathbf{A}\|_1 \leq \|\mathbf{A}\|_{\text{sum}}. \quad (17)$$

552 This follows from the following sequence of inequalities

$$\|\mathbf{A}\|_1 = \max_{\|\mathbf{B}\|_\infty \leq 1} \langle \mathbf{A}, \mathbf{B} \rangle \leq \|\mathbf{A}\|_{\text{sum}} \cdot \max_{\|\mathbf{B}\|_\infty \leq 1} \|\mathbf{B}\|_{\max} \leq \|\mathbf{A}\|_{\text{sum}} \cdot \max_{\|\mathbf{B}\|_\infty \leq 1} \|\mathbf{B}\|_\infty \leq \|\mathbf{A}\|_{\text{sum}}, \quad (18)$$

553 where the first inequality follows from generalized Cauchy-Schwartz and the second inequality by  
554 (16).  $\square$

555 Lemma 12 is used in bounding the second order term in the Taylor expansion of  $\mathcal{L}(\mathbf{W})$ .

556 **Lemma 12.** For any entry-wise or Schatten  $p$ -norm  $\|\cdot\|$  with  $p \geq 1$ , any  $\mathbf{s} \in \Delta^{k-1}$  in the  $k$ -  
 557 dimensional simplex, any index  $c \in [k]$ , and  $\mathbf{v} \in \mathbb{R}^k$ , it holds that

$$\mathbf{v}^\top (\text{diag}(\mathbf{s}) - \mathbf{s}\mathbf{s}^\top) \mathbf{v} \leq 4(1 - s_c) \|\mathbf{v}\mathbf{v}^\top\|.$$

558 *Proof.* See main text. □

559 Lemma 13 is used in the proof of Lemma 12.

560 **Lemma 13.** For any  $\mathbf{s} \in \Delta^{k-1}$  in the  $k$ -dimensional simplex and any index  $c \in [k]$  it holds that

$$\sum_{c'} s_{c'}(1 - s_{c'}) \leq 2(1 - s_c).$$

561 *Proof.* With a bit of algebra and using  $\sum_{c' \neq c} s_{c'} = 1 - s_c$  the claim becomes equivalent to

$$\sum_{c' \neq c} s_{c'}^2 + s_c^2 - 2s_c + 1 \geq 0.$$

562 Since this holds true, the lemma holds. □

563 **Lemma 14.** For any  $\mathbf{s} \in \Delta^{k-1}$  in the  $k$ -dimensional simplex, any index  $c \in [k]$ , any  $\Delta \in \mathbb{R}^{k \times d}$ ,  
 564 and any  $\mathbf{h} \in \mathbb{R}^d$ , it holds:

$$\mathbf{h}^\top \Delta^\top (\text{diag}(\mathbf{s}) - \mathbf{s}\mathbf{s}^\top) \Delta \mathbf{h} \leq 4B^2 \|\Delta\|^2 (1 - s_c).$$

565 *Proof.* We let  $\mathbf{v} := \Delta \mathbf{h}$ . For any Schatten  $p$ -norm, we have

$$\|\mathbf{v}\mathbf{v}^\top\| = \|\mathbf{v}\|_2^2 \leq \|\Delta\|_\infty^2 \|\mathbf{h}\|_2^2 \leq \|\Delta\|^2 \|\mathbf{h}\|_2^2 \leq B^2 \|\Delta\|^2.$$

566 For any entry-wise  $p$ -norm, we have

$$\|\Delta \mathbf{h}\|^p = \|\mathbf{v}\|^p = \sum_j |e_j^\top \Delta \mathbf{h}|^p \leq \sum_j \|e_j^\top \Delta\|_p^p \|\mathbf{h}\|^p = \|\mathbf{h}\|_*^p \sum_{ij} |\Delta[i, j]|^p = \|\mathbf{h}\|_*^p \|\Delta\|^p.$$

567 This implies

$$\|\mathbf{v}\mathbf{v}^\top\| = \|\mathbf{v}\|^2 = \|\Delta \mathbf{h}\|^2 \leq \|\Delta\|^2 \|\mathbf{h}\|_*^2 \leq B^2 \|\Delta\|^2.$$

568 Combine these results and apply Lemma 12, we obtain the desired. □

569 The following lemma summarizes the properties of the softmax map that will be used in the proof of  
 570 Lemma 25 and 35.

571 **Lemma 15.** For any  $\mathbf{v}, \mathbf{v}', \mathbf{q}, \mathbf{q}' \in \mathbb{R}^k$  and  $c \in [k]$ , the following inequalities hold:

572 (i)  $|\frac{\mathbb{S}_c(\mathbf{v}')}{\mathbb{S}_c(\mathbf{v})} - 1| \leq e^{2\|\mathbf{v} - \mathbf{v}'\|_\infty} - 1$

573 (ii)  $|\frac{1 - \mathbb{S}_c(\mathbf{v}')}{1 - \mathbb{S}_c(\mathbf{v})} - 1| \leq e^{2\|\mathbf{v} - \mathbf{v}'\|_\infty} - 1$

574 (iii)  $|\frac{\mathbb{S}_c(\mathbf{v}')\mathbb{S}_c(\mathbf{q}')}{\mathbb{S}_c(\mathbf{v})\mathbb{S}_c(\mathbf{q})} - 1| \leq e^{2(\|\mathbf{v}' - \mathbf{v}\|_\infty + \|\mathbf{q}' - \mathbf{q}\|_\infty)} - 1$

575 (iv)  $|\frac{\mathbb{S}_c(\mathbf{v}')(1 - \mathbb{S}_c(\mathbf{q}'))}{\mathbb{S}_c(\mathbf{v})(1 - \mathbb{S}_c(\mathbf{q}))} - 1| \leq e^{2(\|\mathbf{v}' - \mathbf{v}\|_\infty + \|\mathbf{q}' - \mathbf{q}\|_\infty)} - 1$

576 (v)  $|\frac{(1 - \mathbb{S}_c(\mathbf{v}'))(1 - \mathbb{S}_c(\mathbf{q}'))}{(1 - \mathbb{S}_c(\mathbf{v}))(1 - \mathbb{S}_c(\mathbf{q}))} - 1| \leq e^{2(\|\mathbf{v}' - \mathbf{v}\|_\infty + \|\mathbf{q}' - \mathbf{q}\|_\infty)} - 1$

577 *Proof.* We prove each inequality:

578 (i) First, observe that

$$\begin{aligned} \left| \frac{\mathbb{S}_c(\mathbf{v}')}{\mathbb{S}_c(\mathbf{v})} - 1 \right| &= \left| \frac{e^{v'_c} \sum_{i \in [k]} e^{v_i}}{e^{v_c} \sum_{i \in [k]} e^{v'_i}} - 1 \right| \\ &= \left| \frac{\sum_{i \in [k]} e^{v'_c + v_i} - \sum_{i \in [k]} e^{v_c + v'_i}}{\sum_{i \in [k]} e^{v_c + v'_i}} \right| \\ &\leq \frac{\sum_{i \in [k]} |e^{v'_c + v_i} - e^{v_c + v'_i}|}{\sum_{i \in [k]} e^{v_c + v'_i}} \end{aligned}$$

579 For any  $i \in [k]$ , we have  $\frac{|e^{v'_c + v_i} - e^{v_c + v'_i}|}{e^{v_c + v'_i}} = |e^{v'_c - v_c + v_i - v'_i} - 1| \leq e^{|v'_c - v_c + v_i - v'_i|} - 1 \leq e^{2\|\mathbf{v} - \mathbf{v}'\|_\infty} - 1$ .

580 This implies  $\sum_{i \in [k]} |e^{v'_c + v_i} - e^{v_c + v'_i}| \leq (e^{2\|\mathbf{v} - \mathbf{v}'\|_\infty} - 1) \sum_{i \in [k]} e^{v_c + v'_i}$ , from which we obtain the  
581 desired inequality.

582 (ii) For the second inequality:

$$\begin{aligned} \left| \frac{1 - \mathbb{S}_c(\mathbf{v}')}{1 - \mathbb{S}_c(\mathbf{v})} - 1 \right| &= \left| \frac{1 - \frac{e^{v'_c}}{\sum_{i \in [k]} e^{v'_i}}}{1 - \frac{e^{v_c}}{\sum_{i \in [k]} e^{v_i}}} - 1 \right| \\ &= \left| \frac{(\sum_{j \in [k], j \neq c} e^{v'_j})(\sum_{i \in [k]} e^{v_i})}{(\sum_{j \in [k], j \neq c} e^{v_j})(\sum_{i \in [k]} e^{v'_i})} - 1 \right| \\ &= \left| \frac{\sum_{j \in [k], j \neq c} \sum_{i \in [k]} [e^{v'_j + v_i} - e^{v_j + v'_i}]}{\sum_{j \in [k], j \neq c} \sum_{i \in [k]} e^{v_j + v'_i}} \right| \\ &\leq \frac{\sum_{j \in [k], j \neq c} \sum_{i \in [k]} |e^{v'_j + v_i} - e^{v_j + v'_i}|}{\sum_{j \in [k], j \neq c} \sum_{i \in [k]} e^{v_j + v'_i}} \end{aligned}$$

583 For any  $j \in [k]$ ,  $j \neq c$ , and  $i \in [k]$ , we have  $\frac{|e^{v'_j + v_i} - e^{v_j + v'_i}|}{e^{v_j + v'_i}} \leq e^{2\|\mathbf{v} - \mathbf{v}'\|_\infty} - 1$ . This implies that

584  $\sum_{j \in [k], j \neq c} \sum_{i \in [k]} |e^{v'_j + v_i} - e^{v_j + v'_i}| \leq (e^{2\|\mathbf{v} - \mathbf{v}'\|_\infty} - 1) \sum_{j \in [k], j \neq c} \sum_{i \in [k]} e^{v_j + v'_i}$ , from which the  
585 result follows.

586 (iii) For the third inequality:

$$\begin{aligned} \left| \frac{\mathbb{S}_c(\mathbf{v}')\mathbb{S}_c(\mathbf{q}')}{\mathbb{S}_c(\mathbf{v})\mathbb{S}_c(\mathbf{q})} - 1 \right| &= \left| \frac{\frac{e^{v'_c}}{\sum_{i \in [k]} e^{v'_i}} \frac{e^{q'_c}}{\sum_{i \in [k]} e^{q'_i}}}{\frac{e^{v_c}}{\sum_{i \in [k]} e^{v_i}} \frac{e^{q_c}}{\sum_{i \in [k]} e^{q_i}}} - 1 \right| \\ &= \left| \frac{\frac{e^{v'_c}}{\sum_{i \in [k]} e^{v'_i}} \frac{e^{q'_c}}{\sum_{i \in [k]} e^{q'_i}}}{\frac{e^{v_c}}{\sum_{i \in [k]} e^{v_i}} \frac{e^{q_c}}{\sum_{i \in [k]} e^{q_i}}} - \frac{\frac{e^{v_c}}{\sum_{i \in [k]} e^{v'_i}} \frac{e^{q_c}}{\sum_{i \in [k]} e^{q'_i}}}{\frac{e^{v_c}}{\sum_{i \in [k]} e^{v'_i}} \frac{e^{q_c}}{\sum_{i \in [k]} e^{q'_i}}} \right| \\ &= \left| \frac{e^{v'_c} e^{q'_c} \sum_{i \in [k]} e^{v_i} \sum_{j \in [k]} e^{q_j}}{e^{v_c} e^{q_c} \sum_{i \in [k]} e^{v'_i} \sum_{j \in [k]} e^{q'_j}} - \frac{e^{v_c} e^{q_c} \sum_{i \in [k]} e^{v'_i} \sum_{j \in [k]} e^{q'_j}}{e^{v_c} e^{q_c} \sum_{i \in [k]} e^{v'_i} \sum_{j \in [k]} e^{q'_j}} \right| \\ &= \left| \frac{\sum_{i \in [k]} \sum_{j \in [k]} [e^{v'_c + v_i + q'_c + q_j} - e^{v_c + v'_i + q_c + q'_j}]}{\sum_{i \in [k]} \sum_{j \in [k]} e^{v_c + v'_i + q_c + q'_j}} \right| \\ &\leq \frac{\sum_{i \in [k]} \sum_{j \in [k]} |e^{v'_c + v_i + q'_c + q_j} - e^{v_c + v'_i + q_c + q'_j}|}{\sum_{i \in [k]} \sum_{j \in [k]} e^{v_c + v'_i + q_c + q'_j}} \end{aligned}$$

587 For any  $i \in [k]$  and  $j \in [k]$ ,  $\frac{|e^{v'_c + v_i + q'_c + q_j} - e^{v_c + v'_i + q_c + q'_j}|}{e^{v_c + v'_i + q_c + q'_j}} = |e^{v'_c - v_c + v_i - v'_i + q'_c - q_c + q_j - q'_j} - 1| \leq$

588  $e^{|v'_c - v_c + v_i - v'_i + q'_c - q_c + q_j - q'_j|} - 1 \leq e^{2(\|\mathbf{v}' - \mathbf{v}\|_\infty + \|\mathbf{q}' - \mathbf{q}\|_\infty)} - 1$ . Then, rearranging and summing  
589 over  $i$  and  $j$  leads to the result.

590 (iv) For the fourth inequality:

$$\begin{aligned}
\left| \frac{\mathbb{S}_c(\mathbf{v}')(1 - \mathbb{S}_c(\mathbf{q}'))}{\mathbb{S}_c(\mathbf{v})(1 - \mathbb{S}_c(\mathbf{q}))} - 1 \right| &= \left| \frac{\frac{e^{v'_c}}{\sum_{s \in [k]} e^{v'_s}} (1 - \frac{e^{q'_c}}{\sum_{t \in [k]} e^{q'_t}})}{\frac{e^{v_c}}{\sum_{s \in [k]} e^{v_s}} (1 - \frac{e^{q_c}}{\sum_{t \in [k]} e^{q_t}})} - 1 \right| \\
&= \left| \frac{\frac{e^{v'_c}}{\sum_{s \in [k]} e^{v'_s}} \frac{\sum_{i \in [k], i \neq c} e^{q'_i}}{\sum_{t \in [k]} e^{q'_t}}}{\frac{e^{v_c}}{\sum_{s \in [k]} e^{v_s}} \frac{\sum_{i \in [k], i \neq c} e^{q_i}}{\sum_{t \in [k]} e^{q_t}}} - 1 \right| \\
&= \left| \frac{\sum_{i \in [k], i \neq c} \sum_{t \in [k]} \sum_{s \in [k]} \frac{e^{v'_c + q'_i + v_s + q_t}}{e^{v_c + q_i + v'_s + q'_t}} - 1 \right| \\
&\leq \frac{\sum_{i \in [k], i \neq c} \sum_{t \in [k]} \sum_{s \in [k]} |e^{v'_c + q'_i + v_s + q_t} - e^{v_c + q_i + v'_s + q'_t}|}{\sum_{i \in [k], i \neq c} \sum_{t \in [k]} \sum_{s \in [k]} e^{v_c + q_i + v'_s + q'_t}}
\end{aligned}$$

591 For each  $i \in [k], i \neq c$ ,  $s \in [k]$ , and  $t \in [k]$ , we obtain  $\frac{|e^{v'_c + q'_i + v_s + q_t} - e^{v_c + q_i + v'_s + q'_t}|}{e^{v_c + q_i + v'_s + q'_t}} \leq$   
592  $e^{2(\|\mathbf{v}' - \mathbf{v}\|_\infty + \|\mathbf{q}' - \mathbf{q}\|_\infty)} - 1$ . Then, rearranging and summing over  $i$ ,  $s$ , and  $t$  leads to the result.

593 (v) Finally, for the fifth inequality:

$$\begin{aligned}
\left| \frac{(1 - \mathbb{S}_c(\mathbf{v}'))(1 - \mathbb{S}_c(\mathbf{q}'))}{(1 - \mathbb{S}_c(\mathbf{v}))(1 - \mathbb{S}_c(\mathbf{q}))} - 1 \right| &= \left| \frac{(1 - \frac{e^{v'_c}}{\sum_{s \in [k]} e^{v'_s}})(1 - \frac{e^{q'_c}}{\sum_{t \in [k]} e^{q'_t}})}{(1 - \frac{e^{v_c}}{\sum_{s \in [k]} e^{v_s}})(1 - \frac{e^{q_c}}{\sum_{t \in [k]} e^{q_t}})} - 1 \right| \\
&= \left| \frac{\frac{\sum_{j \in [k], j \neq c} e^{v'_j}}{\sum_{s \in [k]} e^{v'_s}} \frac{\sum_{i \in [k], i \neq c} e^{q'_i}}{\sum_{t \in [k]} e^{q'_t}}}{\frac{\sum_{j \in [k], j \neq c} e^{v_j}}{\sum_{s \in [k]} e^{v_s}} \frac{\sum_{i \in [k], i \neq c} e^{q_i}}{\sum_{t \in [k]} e^{q_t}}} - 1 \right| \\
&= \left| \frac{\sum_{j \in [k], j \neq c} \sum_{i \in [k], i \neq c} \sum_{t \in [k]} \sum_{s \in [k]} \frac{e^{v'_j + q'_i + v_s + q_t}}{e^{v_j + q_i + v'_s + q'_t}} - 1 \right| \\
&\leq \frac{\sum_{j \in [k], j \neq c} \sum_{i \in [k], i \neq c} \sum_{t \in [k]} \sum_{s \in [k]} |e^{v'_j + q'_i + v_s + q_t} - e^{v_j + q_i + v'_s + q'_t}|}{\sum_{j \in [k], j \neq c} \sum_{i \in [k], i \neq c} \sum_{t \in [k]} \sum_{s \in [k]} e^{v_j + q_i + v'_s + q'_t}}.
\end{aligned}$$

594 For each  $j \in [k]$  ( $j \neq c$ ),  $i \in [k]$  ( $i \neq c$ ),  $s \in [k]$ , and  $t \in [k]$ , we have

$$\begin{aligned}
\frac{|e^{v'_j + q'_i + v_s + q_t} - e^{v_j + q_i + v'_s + q'_t}|}{e^{v_j + q_i + v'_s + q'_t}} &= |e^{v'_j - v_j + q'_i - q_i + v_s - v'_s + q_t - q'_t} - 1| \\
&\leq e^{|v'_j - v_j| + |q'_i - q_i| + |v_s - v'_s| + |q_t - q'_t|} - 1 \\
&\leq e^{2(\|\mathbf{v}' - \mathbf{v}\|_\infty + \|\mathbf{q}' - \mathbf{q}\|_\infty)} - 1
\end{aligned}$$

595 Then, rearranging and summing over  $j$ ,  $i$ ,  $s$ , and  $t$  leads to the result.  $\square$

## 596 C Lemmas on Loss and Proxy Function

597 Lemma 16 shows that  $\mathcal{G}(\mathbf{W})$  upper and lower bound the dual norm of the loss gradient.

598 **Lemma 16** ( $\mathcal{G}(\mathbf{W})$  as proxy to the loss-gradient norm). *Under Assumption 4. For any  $\mathbf{W} \in \mathbb{R}^{k \times d}$ ,*  
599 *it holds that*

$$2B \cdot \mathcal{G}(\mathbf{W}) \geq \|\nabla \mathcal{L}(\mathbf{W})\|_* \geq \gamma \cdot \mathcal{G}(\mathbf{W}).$$



600 *Proof.* First, we prove the lower bound. By duality and direct application of (12)

$$\begin{aligned}\|\nabla\mathcal{L}(\mathbf{W})\|_* &= \max_{\|\mathbf{A}\|\leq 1} \langle \mathbf{A}, -\nabla\mathcal{L}(\mathbf{W}) \rangle \\ &\geq \max_{\|\mathbf{A}\|\leq 1} \frac{1}{n} \sum_{i \in [n]} (1 - s_{iy_i}) \min_{c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{A} \mathbf{h}_i \\ &\geq \frac{1}{n} \sum_{i \in [n]} (1 - s_{iy_i}) \cdot \max_{\|\mathbf{A}\|\leq 1} \min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{A} \mathbf{h}_i.\end{aligned}$$

601 Second, for the upper bound, it holds by triangle inequality and relationships (7) that

$$\|\nabla\mathcal{L}(\mathbf{W})\|_* \leq \|\nabla\mathcal{L}(\mathbf{W})\|_{\text{sum}} \leq \frac{1}{n} \sum_{i \in [n]} \|\nabla\ell_i(\mathbf{W})\|_{\text{sum}},$$

602 where  $\ell_i(\mathbf{W}) = -\log(\mathbb{S}_{y_i}(\mathbf{W}\mathbf{h}_i))$ . Recall that

$$\nabla\ell_i(\mathbf{W}) = -(\mathbf{e}_{y_i} - \mathbb{S}_{y_i}(\mathbf{W}\mathbf{h}_i))\mathbf{h}_i^\top,$$

603 and, for two vectors  $\mathbf{v}, \mathbf{u}$ :  $\|\mathbf{u}\mathbf{v}^\top\|_{\text{sum}} = \|\mathbf{u}\|_1 \|\mathbf{v}\|_1$ . Combining these and noting that

$$\|\mathbf{e}_{y_i} - \mathbb{S}_{y_i}(\mathbf{W}\mathbf{h}_i)\|_1 = 2(1 - s_{y_i})$$

604 together with using the assumption  $\|\mathbf{h}_i\| \leq B$  yields the advertised upper bound.  $\square$

605 Built upon Lemma 16, we obtain a simple bound on the loss difference at two points.

606 **Lemma 17.** For any  $\mathbf{W}, \mathbf{W}_0 \in \mathbb{R}^{k \times d}$ , suppose that  $\mathcal{L}(\mathbf{W})$  is convex, we have

$$|\mathcal{L}(\mathbf{W}) - \mathcal{L}(\mathbf{W}_0)| \leq 2B\|\mathbf{W} - \mathbf{W}_0\|.$$

607 *Proof.* By convexity of  $\mathcal{L}$ , we have

$$\mathcal{L}(\mathbf{W}_0) - \mathcal{L}(\mathbf{W}) \leq \langle \nabla\mathcal{L}(\mathbf{W}_0), \mathbf{W}_0 - \mathbf{W} \rangle \leq \|\nabla\mathcal{L}(\mathbf{W}_0)\|_* \|\mathbf{W}_0 - \mathbf{W}\| \leq 2B\|\mathbf{W}_0 - \mathbf{W}\|,$$

608 where the last inequality is by Lemma 16. Similarly, we can also show that  $\mathcal{L}(\mathbf{W}) - \mathcal{L}(\mathbf{W}_0) \leq$   
609  $2B\|\mathbf{W}_0 - \mathbf{W}\|$ .  $\square$

610 Lemma 18 shows the close relationships between  $\mathcal{G}(\mathbf{W})$  and  $\mathcal{L}(\mathbf{W})$ . The proxy  $\mathcal{G}(\mathbf{W})$  not only  
611 lower bounds  $\mathcal{L}(\mathbf{W})$ , but also upper bounds  $\mathcal{L}(\mathbf{W})$  up to a factor depending on  $\mathcal{L}(\mathbf{W})$ . Moreover,  
612 the rate of convergence  $\frac{\mathcal{G}(\mathbf{W})}{\mathcal{L}(\mathbf{W})}$  depends on the rate of decrease in the loss.

613 **Lemma 18** ( $\mathcal{G}(\mathbf{W})$  as proxy to the loss). Let  $\mathbf{W} \in \mathbb{R}^{k \times d}$ , we have

614 (i)  $1 \geq \frac{\mathcal{G}(\mathbf{W})}{\mathcal{L}(\mathbf{W})} \geq 1 - \frac{n\mathcal{L}(\mathbf{W})}{2}$

615 (ii) Suppose that  $\mathbf{W}$  satisfies  $\mathcal{L}(\mathbf{W}) \leq \frac{\log 2}{n}$  or  $\mathcal{G}(\mathbf{W}) \leq \frac{1}{2n}$ , then  $\mathcal{L}(\mathbf{W}) \leq 2\mathcal{G}(\mathbf{W})$ .

616 *Proof.* (i) Denote for simplicity  $s_i := s_{iy_i} = \mathbb{S}_{y_i}(\mathbf{W}\mathbf{h}_i)$ , thus  $\mathcal{L}(\mathbf{W}) = \frac{1}{n} \sum_{i \in [n]} \log(1/s_i)$  and  
617  $\mathcal{G}(\mathbf{W}) = \frac{1}{n} \sum_{i \in [n]} (1 - s_i)$ . For the upper bound, simply use the fact that  $e^{x-1} \geq x$ , for all  $x \in [0, 1]$ ,  
618 thus  $\log(1/s_i) \geq 1 - s_i$  for all  $i \in [n]$ .

619 The lower bound can be proved using the exact same arguments in the proof of Zhang et al. [60,  
620 Lemma C.7] for the binary case. For completeness, we provide an alternative elementary proof. It  
621 suffices to prove for  $n = 1$  that for  $s \in (0, 1)$ :

$$1 - s \geq \log(1/s) - \frac{1}{2} \log^2(1/s). \quad (19)$$

622 The general case follows by summing over  $s = s_i$  and using  $\sum_{i \in [n]} \log^2(1/s_i) \leq$   
623  $\left(\sum_{i \in [n]} \log(1/s_i)\right)^2$  since  $\log(1/s_i) > 0$ . For (19), let  $x = \log(1/s) > 0$ . The inequality be-  
624 comes  $e^{-x} \leq 1 - x + x^2/2$ , which holds for  $x > 0$  by the second-order Taylor expansion of  $e^{-x}$   
625 around 0.

626 (ii) The sufficiency of  $\mathcal{L}(\mathbf{W}) \leq \frac{\log 2}{n}$  (to guarantee that  $\mathcal{L}(\mathbf{W}) \leq 2\mathcal{G}(\mathbf{W})$ ) follows from (i) and  
 627  $\mathcal{L}(\mathbf{W}) \leq \frac{\log 2}{n} \leq \frac{1}{n}$ . The inequality  $\log(\frac{1}{x}) \leq 2(1-x)$  holds when  $x \in [0.2032, 1]$ . This translates  
 628 to the following sufficient condition on  $s_{iy_i}$

$$s_i = \frac{e^{\ell_i[y_i]}}{\sum_{c \in [k]} e^{\ell_i[c]}} = \frac{1}{1 + \sum_{c \in [k], c \neq y_i} e^{\ell_i[c] - \ell_i[y_i]}} \geq 0.2032.$$

629 Under the assumption  $\mathcal{G}(\mathbf{W}) \leq \frac{1}{2n}$ , we have  $1 - s_i \leq \sum_{i \in [n]} (1 - s_i) = n\mathcal{G}(\mathbf{W}) \leq \frac{1}{2}$ , from which  
 630 we obtain  $s_i \geq \frac{1}{2} \geq 0.2032$  for all  $i \in [n]$ .  $\square$

631 Lemma 19 shows that the data becomes separable when the loss is small. It is used in deriving the  
 632 lower bound on the un-normalized margin.

633 **Lemma 19** (Low  $\mathcal{L}(\mathbf{W})$  implies separability). *Suppose that there exists  $\mathbf{W} \in \mathbb{R}^{k \times d}$  such that*  
 634  $\mathcal{L}(\mathbf{W}) \leq \frac{\log 2}{n}$ , *then we have*

$$(\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W} \mathbf{h}_i \geq 0, \quad \text{for all } i \in [n] \text{ and for all } c \in [k] \text{ such that } c \neq y_i. \quad (20)$$

635 *Proof.* We rewrite the loss into the form:

$$\mathcal{L}(\mathbf{W}) = -\frac{1}{n} \sum_{i \in [n]} \log\left(\frac{e^{\ell_i[y_i]}}{\sum_{c \in [k]} e^{\ell_i[c]}}\right) = \frac{1}{n} \sum_{i \in [n]} \log\left(1 + \sum_{c \neq y_i} e^{-(\ell_i[y_i] - \ell_i[c])}\right).$$

636 Fix any  $i \in [n]$ , by the assumption that  $\mathcal{L}(\mathbf{W}) \leq \frac{\log 2}{n}$ , we have the following:

$$\log\left(1 + \sum_{c \neq y_i} e^{-(\ell_i[y_i] - \ell_i[c])}\right) \leq n\mathcal{L}(\mathbf{W}) \leq \log(2).$$

637 This implies:

$$e^{-\min_{c \neq y_i} (\ell_i[y_i] - \ell_i[c])} = \max_{c \neq y_i} e^{-(\ell_i[y_i] - \ell_i[c])} \leq \sum_{c \neq y_i} e^{-(\ell_i[y_i] - \ell_i[c])} \leq 1.$$

638 After taking log on both sides, we obtain the following:  $\ell_i[y_i] - \ell_i[c] = (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W} \mathbf{h}_i \geq 0$  for  
 639 any  $c \in [k]$  such that  $c \neq y_i$ .  $\square$

640 Lemma 20 shows that the ratio of  $\mathcal{G}(\mathbf{W})$  at two points can be bounded by exponentiating the max-  
 641 norm of their differences. It is used in handling the second order term in the Taylor expansion of the  
 642 loss.

643 **Lemma 20** (Ratio of  $\mathcal{G}(\mathbf{W})$ ). *For any  $\psi \in [0, 1]$ , we have the following:*

$$\frac{\mathcal{G}(\mathbf{W} - \psi\eta\Delta)}{\mathcal{G}(\mathbf{W})} \leq e^{2B\psi\eta\|\Delta\|_{\max}} \leq e^{2B\psi\eta\|\Delta\|}.$$

644 *Proof.* Note that the second inequality is by relationships (7). Here, we only prove the first inequality.  
 645 By the definition of  $\mathcal{G}(\mathbf{W})$ , we have:

$$\frac{\mathcal{G}(\mathbf{W} - \psi\eta\Delta)}{\mathcal{G}(\mathbf{W})} = \frac{\sum_{i \in [n]} (1 - \mathbb{S}_{y_i}((\mathbf{W} - \psi\eta\Delta)\mathbf{h}_i))}{\sum_{i \in [n]} (1 - \mathbb{S}_{y_i}(\mathbf{W}\mathbf{h}_i))}.$$

646 For any  $c \in [k]$  and  $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^k$ , we have:

$$\begin{aligned} \frac{1 - \mathbb{S}_c(\mathbf{v}')}{1 - \mathbb{S}_c(\mathbf{v})} &= \frac{1 - \frac{e^{v'_c}}{\sum_{i \in [k]} e^{v'_i}}}{1 - \frac{e^{v_c}}{\sum_{i \in [k]} e^{v_i}}} \\ &= \frac{\frac{\sum_{j \in [k], j \neq c} e^{v'_j}}{\sum_{i \in [k]} e^{v'_i}}}{\frac{\sum_{j \in [k], j \neq c} e^{v_j}}{\sum_{i \in [k]} e^{v_i}}} \\ &= \frac{\sum_{j \in [k], j \neq c} \sum_{i \in [k]} e^{v'_j + v_i}}{\sum_{j \in [k], j \neq c} \sum_{i \in [k]} e^{v_j + v'_i}} \\ &\leq e^{2\|\mathbf{v} - \mathbf{v}'\|_{\infty}}. \end{aligned}$$

647 The last inequality is because  $\frac{e^{v'_j+v_i}}{e^{v_j+v'_i}} \leq e^{|v'_j-v_j|+|v_i-v'_i|} \leq e^{2\|v-v'\|_\infty}$ , which implies that  
 648  $\sum_{j \in [k], j \neq c} \sum_{i \in [k]} e^{v'_j+v_i} \leq e^{2\|v-v'\|_\infty} \sum_{j \in [k], j \neq c} \sum_{i \in [k]} e^{v_j+v'_i}$ . Next, we specialize this result to  
 649  $v' = (\mathbf{W} - \psi\eta\Delta)\mathbf{h}_i$ ,  $v = \mathbf{W}\mathbf{h}_i$ , and  $c = y_i$  for any  $i \in [n]$  to obtain:

$$\frac{1 - \mathbb{S}_{y_i}((\mathbf{W} - \psi\eta\Delta)\mathbf{h}_i)}{1 - \mathbb{S}_{y_i}(\mathbf{W}\mathbf{h}_i)} \leq e^{2\eta\psi\|\Delta\mathbf{h}_i\|_\infty} \leq e^{2B\eta\psi\|\Delta\|_{\max}}.$$

650 Then, we rearrange and sum over  $i \in [n]$  to obtain:  $\sum_{i \in [n]} (1 - \mathbb{S}_{y_i}((\mathbf{W} - \psi\eta\Delta)\mathbf{h}_i)) \leq$   
 651  $e^{2B\eta\psi\|\Delta\|_{\max}} \sum_{i \in [n]} (1 - \mathbb{S}_{y_i}(\mathbf{W}\mathbf{h}_i))$ , from which the desired inequality follows. The second  
 652 inequality in the lemma statement follows from the relationship (7).  $\square$

## 653 D Implicit Bias of Normalized Steepest Descent

654 **Proof Overview** We consider a decay learning rate schedule of the form  $\eta_t = \Theta(\frac{1}{t^a})$  where  
 655  $a \in (0, 1]$ . The first step is to show that the **loss monotonically decreases** after certain time and  
 656 the rate depends on  $\mathcal{G}(\mathbf{W})$ . To obtain this, we apply Lemma 16 and Lemma 12 to upper bound the  
 657 first-order and second-order terms in the Taylor expansion of the loss (21), respectively. Next, we use  
 658 the decrease in loss to derive a lower bound on the unnormalized margin which involves the ratio  
 659  $\frac{\mathcal{G}(\mathbf{W})}{\mathcal{L}(\mathbf{W})}$ . A crucial step involved is to find a time  $\bar{t}_2$  such that separability (32) holds for all  $t \geq \bar{t}_2$ , and  
 660 the existence of  $\bar{t}_2$  is guaranteed by loss monotonicity such that the condition  $\mathcal{L}(\mathbf{W}_t) \leq \frac{\log 2}{n}$  will be  
 661 satisfied for sufficiently large  $t$ 's.

662 Then, we argue that the ratio  $\frac{\mathcal{G}(\mathbf{W}_t)}{\mathcal{L}(\mathbf{W}_t)}$  converges to 1 exponentially fast (recalling that  $1 \geq \frac{\mathcal{G}(\mathbf{W}_t)}{\mathcal{L}(\mathbf{W}_t)} \geq$   
 663  $1 - \frac{n\mathcal{L}(\mathbf{W}_t)}{2}$ ) by showing the loss  $\mathcal{L}(\mathbf{W}_t)$  decreases exponentially fast. We first choose a time  $t_1$  after  
 664  $t_0$  (recall that  $t_0$  is the time that satisfies Assumption 3) such that  $\mathcal{L}(\mathbf{W}_{t+1}) \leq \mathcal{L}(\mathbf{W}_t) - \frac{\eta_t\gamma}{2}\mathcal{G}(\mathbf{W}_t)$   
 665 for all  $t \geq t_1$ . Next, we lower bound  $\mathcal{G}(\mathbf{W}_t)$  using  $\mathcal{L}(\mathbf{W}_t)$ . By Lemma 18, there are two sufficient  
 666 conditions (namely,  $\mathcal{L}(\mathbf{W}_t) \leq \frac{\log 2}{n} =: \tilde{\mathcal{L}}$  or  $\mathcal{G}(\mathbf{W}_t) \leq \frac{1}{2n}$ ) that guarantee  $\mathcal{L}(\mathbf{W}_t) \leq 2\mathcal{G}(\mathbf{W}_t)$ . We  
 667 choose a time  $t_2$  (after  $t_1$ ) that is sufficiently large such that there exists  $t^* \in [t_1, t_2]$  for which we  
 668 have  $\mathcal{G}(\mathbf{W}_{t^*}) \leq \frac{\tilde{\mathcal{L}}}{2} \leq \frac{1}{2n}$ . This not only guarantees that  $\mathcal{L}(\mathbf{W}_{t^*}) \leq 2\mathcal{G}(\mathbf{W}_{t^*})$  at time  $t^*$ , but also  
 669 (crucially due to monotonicity) implies that  $\mathcal{L}(\mathbf{W}_t) \leq \mathcal{L}(\mathbf{W}_{t^*}) \leq 2\mathcal{G}(\mathbf{W}_{t^*}) \leq \frac{\log 2}{n}$  for all  $t \geq t_2$ .  
 670 Thus, we observe that the other sufficient condition  $\mathcal{L}(\mathbf{W}_t) \leq \frac{\log 2}{n}$  is satisfied, from which we  
 671 conclude that  $\mathcal{L}(\mathbf{W}_t) \leq 2\mathcal{G}(\mathbf{W}_t)$  for all  $t \geq t_2$ . We remark that the choice of  $t_2$  depends on  $\mathcal{L}(\mathbf{W}_{t_1})$   
 672 (whose magnitude is bounded using Lemma 17), and  $t_2$  can be used as  $\bar{t}_2$  above. To recap,  $t_1$  is the  
 673 time (after  $t_0$ ) after which the successive loss decrease is lower bounded by the product  $\eta_t\gamma\mathcal{G}(\mathbf{W}_t)$ ;  
 674  $t_2$  (after  $t_1$ ) is the time after which  $\mathcal{L}(\mathbf{W}_t) \leq \frac{\log 2}{n}$  (thus, both  $\mathcal{L}(\mathbf{W}_t) \leq 2\mathcal{G}(\mathbf{W}_t)$  and separability  
 675 condition (32) hold for all  $t \geq t_2$ ).

676 In this following, we break the proof of implicit bias of NSD into several parts following previous  
 677 arguments. Lemma 21 shows the descent properties of NSD. It is used in Lemma 22 to lower bound  
 678 the un-normalized margin, and in the proof of Theorem 3 to show the convergence of  $\frac{\mathcal{G}(\mathbf{W}_t)}{\mathcal{L}(\mathbf{W}_t)}$ .

679 **Lemma 21** (NSD Descent). *Under the same setting as Theorem 3, it holds for all  $t \geq 0$ ,*

$$\mathcal{L}(\mathbf{W}_{t+1}) \leq \mathcal{L}(\mathbf{W}_t) - \gamma\eta_t(1 - \alpha_{s_1}\eta_t)\mathcal{G}(\mathbf{W}_t),$$

680 where  $\alpha_{s_1}$  is some constant that depends on  $B$  and  $\gamma$ .

681 *Proof.* By Lemma 10, we let  $\mathbf{W}' = \mathbf{W}_{t+1}$ ,  $\mathbf{W} = \mathbf{W}_t$ ,  $\tilde{\Delta}_t = \mathbf{W}_{t+1} - \mathbf{W}_t$ , and define  $\mathbf{W}_{t,t+1,\zeta} :=$   
 682  $\mathbf{W}_t + \zeta(\mathbf{W}_{t+1} - \mathbf{W}_t)$ . We choose  $\zeta^*$  such that  $\mathbf{W}_{t,t+1,\zeta^*}$  satisfies (14), we have:

$$\begin{aligned} \mathcal{L}(\mathbf{W}_{t+1}) &= \mathcal{L}(\mathbf{W}_t) + \underbrace{\langle \nabla \mathcal{L}(\mathbf{W}_t), \tilde{\Delta}_t \rangle}_{\spadesuit_t} \\ &\quad + \frac{1}{2n} \sum_{i \in [n]} \underbrace{\mathbf{h}_i^\top \tilde{\Delta}_t^\top (\text{diag}(\mathbb{S}(\mathbf{W}_{t,t+1,\gamma}\mathbf{h}_i)) - \mathbb{S}(\mathbf{W}_{t,t+1,\zeta^*}\mathbf{h}_i)\mathbb{S}(\mathbf{W}_{t,t+1,\zeta^*}\mathbf{h}_i)^\top) \tilde{\Delta}_t \mathbf{h}_i}_{\clubsuit_t}. \end{aligned} \tag{21}$$

For the  $\spadesuit_t$  term, we have by Lemma 16:

$$\spadesuit_t = -\eta_t \|\nabla \mathcal{L}(\mathbf{W}_t)\|_* \leq -\eta_t \gamma G(\mathbf{W}_t).$$

For the  $\clubsuit_t$  term, we let  $\mathbf{v} = \tilde{\Delta}_t \mathbf{h}_i$  and  $\mathbf{s} = \mathbb{S}(\mathbf{W}_{t,t+1,\zeta^*} \mathbf{h}_i)$ , and apply Lemma 14 to obtain

$$\clubsuit_t \leq 4\|\tilde{\Delta}_t\|^2 \|\mathbf{h}_i\|_*^2 (1 - \mathbb{S}_{y_i}(\mathbf{W}_{t,t+1,\zeta^*} \mathbf{h}_i)) \leq 4\eta_t^2 B^2 (1 - \mathbb{S}_{y_i}(\mathbf{W}_{t,t+1,\zeta^*} \mathbf{h}_i)),$$

where in the second inequality we have used  $\|\tilde{\Delta}_t\| \leq \eta_t$  and  $\|\mathbf{h}_i\|_* \leq \|\mathbf{h}_i\|_1 \leq 1$ . Putting these two pieces together, we obtain

$$\begin{aligned} \mathcal{L}(\mathbf{W}_{t+1}) &\leq \mathcal{L}(\mathbf{W}_t) - \gamma \eta_t \mathcal{G}(\mathbf{W}_t) + 2\eta_t^2 B^2 \frac{1}{n} \sum_{i \in [n]} (1 - \mathbb{S}_{y_i}(\mathbf{W}_{t,t+1,\zeta^*} \mathbf{h}_i)) \\ &= \mathcal{L}(\mathbf{W}_t) - \gamma \eta_t \mathcal{G}(\mathbf{W}_t) + 2\eta_t^2 B^2 \mathcal{G}(\mathbf{W}_{t,t+1,\zeta^*}) \\ &\leq \mathcal{L}(\mathbf{W}_t) - \gamma \eta_t \mathcal{G}(\mathbf{W}_t) + 2\eta_t^2 B^2 \sup_{\zeta \in [0,1]} \mathcal{G}(\mathbf{W}_{t,t+1,\zeta}) \\ &= \mathcal{L}(\mathbf{W}_t) - \gamma \eta_t \mathcal{G}(\mathbf{W}_t) + 2\eta_t^2 B^2 \mathcal{G}(\mathbf{W}_t) \sup_{\zeta \in [0,1]} \frac{\mathcal{G}(\mathbf{W}_t + \zeta \tilde{\Delta}_t)}{\mathcal{G}(\mathbf{W}_t)} \\ &\stackrel{(a)}{\leq} \mathcal{L}(\mathbf{W}_t) - \gamma \eta_t \mathcal{G}(\mathbf{W}_t) + 2\eta_t^2 B^2 \mathcal{G}(\mathbf{W}_t) \sup_{\zeta \in [0,1]} e^{2B\zeta \|\tilde{\Delta}_t\|} \\ &\stackrel{(b)}{\leq} \mathcal{L}(\mathbf{W}_t) - \gamma \eta_t \mathcal{G}(\mathbf{W}_t) + 2\eta_t^2 B^2 e^{2B\eta_0} \mathcal{G}(\mathbf{W}_t), \end{aligned} \quad (22)$$

where (a) is by Lemma 20 and (b) is by  $\|\tilde{\Delta}_t\| \leq \eta_t$ . Letting  $\alpha_{s_1} = \frac{2B^2 e^{2B\eta_0}}{\gamma}$ , Eq. (22) simplifies to:

$$\mathcal{L}(\mathbf{W}_{t+1}) \leq \mathcal{L}(\mathbf{W}_t) - \gamma \eta_t (1 - \alpha_{s_1} \eta_t) \mathcal{G}(\mathbf{W}_t),$$

from which we observe that the loss starts to monotonically decrease after  $\eta_t$  satisfies  $\eta_t \leq \frac{1}{\alpha_{s_1}}$  for a decreasing learning rate schedule.  $\square$

For a decaying learning rate schedule, Lemma 21 implies that the loss monotonically decreases after a certain time. Thus, we know that the assumption of Lemma 22 can be satisfied. In the proof of Theorem 3, we will specify a concrete form of  $\tilde{t}$  in Lemma 22.

**Lemma 22** (NSD Unnormalized Margin). *Suppose that there exist  $\tilde{t}$  such that  $\mathcal{L}(\mathbf{W}_t) \leq \frac{\log 2}{n}$  for all  $t > \tilde{t}$ , then we have*

$$\min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i \geq \gamma \sum_{s=\tilde{t}}^{t-1} \eta_s \frac{\mathcal{G}(\mathbf{W}_s)}{\mathcal{L}(\mathbf{W}_s)} - \alpha_{s_2} \sum_{s=\tilde{t}}^{t-1} \eta_s^2,$$

where  $\alpha_{s_2}$  is some constant that depends on  $B$ .

*Proof.* We let  $\alpha_{s_2} = 2Be^{2B\eta_0}$ , then from (22), we have for  $t > \tilde{t}$ :

$$\begin{aligned} \mathcal{L}(\mathbf{W}_{t+1}) &\leq \mathcal{L}(\mathbf{W}_t) - \gamma \eta_t \mathcal{G}(\mathbf{W}_t) + \alpha_{s_2} \eta_t^2 \mathcal{G}(\mathbf{W}_t) \\ &= \mathcal{L}(\mathbf{W}_t) \left(1 - \gamma \eta_t \frac{\mathcal{G}(\mathbf{W}_t)}{\mathcal{L}(\mathbf{W}_t)} + \alpha_{s_2} \eta_t^2 \frac{\mathcal{G}(\mathbf{W}_t)}{\mathcal{L}(\mathbf{W}_t)}\right) \\ &\leq \mathcal{L}(\mathbf{W}_t) \exp\left(-\gamma \eta_t \frac{\mathcal{G}(\mathbf{W}_t)}{\mathcal{L}(\mathbf{W}_t)} + \alpha_{s_2} \eta_t^2 \frac{\mathcal{G}(\mathbf{W}_t)}{\mathcal{L}(\mathbf{W}_t)}\right) \\ &\leq \mathcal{L}(\mathbf{W}_{\tilde{t}}) \exp\left(-\gamma \sum_{s=\tilde{t}}^t \eta_s \frac{\mathcal{G}(\mathbf{W}_s)}{\mathcal{L}(\mathbf{W}_s)} + \alpha_{s_2} \sum_{s=\tilde{t}}^t \eta_s^2\right). \\ &\leq \frac{\log 2}{n} \exp\left(-\gamma \sum_{s=\tilde{t}}^t \eta_s \frac{\mathcal{G}(\mathbf{W}_s)}{\mathcal{L}(\mathbf{W}_s)} + \alpha_{s_2} \sum_{s=\tilde{t}}^t \eta_s^2\right), \end{aligned} \quad (23)$$

where the penultimate inequality uses Lemma 18, and the last inequality uses the assumption that  $\mathcal{L}(\mathbf{W}_t) \leq \frac{\log 2}{n}$  for all  $t \geq \tilde{t}$ . Then, we have for all  $t > \tilde{t}$ :

$$\begin{aligned} e^{-\min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i} &= \max_{i \in [n]} e^{-\min_{c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i} \\ &\stackrel{(a)}{\leq} \max_{i \in [n]} \frac{1}{\log 2} \log(1 + e^{-\min_{c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i}) \\ &\leq \max_{i \in [n]} \frac{1}{\log 2} \log(1 + \sum_{c \neq y_i} e^{-(\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i}) \leq \frac{n\mathcal{L}(\mathbf{W}_t)}{\log 2} \\ &\stackrel{(b)}{\leq} \exp(-\gamma \sum_{s=\tilde{t}}^{t-1} \eta_s \frac{\mathcal{G}(\mathbf{W}_s)}{\mathcal{L}(\mathbf{W}_s)} + \alpha_{s_2} \sum_{s=\tilde{t}}^{t-1} \eta_s^2). \end{aligned}$$

(a) is by the following: the assumption  $\mathcal{L}(\mathbf{W}_t) \leq \frac{\log 2}{n}$  implies that  $\min_{c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i \geq 0$  for all  $i \in [n]$  by Lemma 19. We also know the inequality  $\frac{\log(1+e^{-z})}{e^{-z}} \geq \log 2$  holds for any  $z \geq 0$ . Then, for any  $i \in [n]$ , we can set  $z = \min_{c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i$  to obtain the desired inequality; and (b) is by (23). Finally, taking log on both sides leads to the result.  $\square$

Next Lemma upper bounds the p-norm of NSD's iterates using learning rates. It is used in the proof of Theorem 3.

**Lemma 23** (NSD  $\|\mathbf{W}_t\|$ ). *For NSD, we have for any  $t > 0$  that*

$$\|\mathbf{W}_t\| \leq \|\mathbf{W}_0\| + \sum_{s=0}^{t-1} \eta_s.$$

*Proof.* By the NSD update rule (3), we have

$$\mathbf{W}_{t+1} = \mathbf{W}_0 - \sum_{s=0}^t \eta_s \Delta_s.$$

This leads to  $\|\mathbf{W}_t\| \leq \|\mathbf{W}_0\| + \sum_{s=0}^{t-1} \eta_s$  given  $\Delta_s \leq 1$  for all  $s \geq 0$ .  $\square$

The main step in the proof of Theorem 3 is to determine the time that satisfies the assumption in Lemma 22 and show the convergence of  $\frac{\mathcal{G}(\mathbf{W}_t)}{\mathcal{L}(\mathbf{W}_t)}$ . Then, Lemma 22 and Lemma 23 will be combined to obtain the final result.

**Theorem 3.** *Suppose that Assumption 1, 2, and 4 hold, then there exists  $t_{s_2} = t_{s_2}(n, \gamma, B, \mathbf{W}_0)$  such that NSD achieves the following for all  $t > t_{s_2}$*

$$\left| \frac{\min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i}{\|\mathbf{W}_t\|} - \gamma \right| \leq \mathcal{O}\left( \frac{\sum_{s=t_{s_2}}^{t-1} \eta_s e^{-\frac{\gamma}{4} \sum_{\tau=t_{s_2}}^{s-1} \eta_\tau} + \sum_{s=0}^{t_{s_2}-1} \eta_s + \sum_{s=t_{s_2}}^{t-1} \eta_s^2}{\sum_{s=0}^{t-1} \eta_s} \right).$$

*Proof. Determination of  $t_{s_1}$ .* In Lemma 21 we choose  $t_{s_1}$  such that  $\eta_t \leq \frac{1}{2\alpha_{s_1}}$  for all  $t \geq t_{s_1}$ .

Considering  $\eta_t = \Theta(\frac{1}{t^a})$  (where  $a \in (0, 1]$ ), we set  $t_{s_1} = (2\alpha_{s_1})^{\frac{1}{a}} = (\frac{4B^2 e^{2B\eta_0}}{\gamma})^{\frac{1}{a}}$ . Then, we have for all  $t \geq t_{s_1}$

$$\mathcal{L}(\mathbf{W}_{t+1}) \leq \mathcal{L}(\mathbf{W}_t) - \frac{\eta_t \gamma}{2} \mathcal{G}(\mathbf{W}_t). \quad (24)$$

Rearranging this equation and using non-negativity of the loss we obtain  $\gamma \sum_{s=t_{s_1}}^t \eta_s \mathcal{G}(\mathbf{W}_s) \leq 2\mathcal{L}(\mathbf{W}_{t_{s_1}})$ .

**Determination of  $t_{s_2}$ .** By Lemma 17, we can bound  $\mathcal{L}(\mathbf{W}_{t_{s_1}})$  as follows

$$|\mathcal{L}(\mathbf{W}_{t_{s_1}}) - \mathcal{L}(\mathbf{W}_0)| \leq 2B\|\mathbf{W}_{t_{s_1}} - \mathbf{W}_0\| \leq 2B \sum_{s=0}^{t_{s_1}-1} \eta_s \|\Delta_s\| \leq 2B \sum_{s=0}^{t_{s_1}-1} \eta_s,$$



where the last inequality is by  $\|\Delta_s\| \leq 1$  for all  $s \geq 0$ . Combining this with the result above and letting  $\tilde{\mathcal{L}} := \frac{\log 2}{n}$ , we obtain

$$\mathcal{G}(\mathbf{W}_{t^*}) = \min_{s \in [t_{s_1}, t_{s_2}]} \mathcal{G}(\mathbf{W}_s) \leq \frac{2\mathcal{L}(\mathbf{W}_0) + 4B \sum_{s=0}^{t_{s_1}-1} \eta_s}{\gamma \sum_{s=t_{s_1}}^{t_{s_2}} \eta_s} \leq \frac{\tilde{\mathcal{L}}}{2} \leq \frac{1}{2n},$$

from which we derive the sufficient condition on  $t_{s_2}$  to be  $\sum_{s=t_{s_1}}^{t_{s_2}} \eta_s \geq \frac{4\mathcal{L}(\mathbf{W}_0) + 8B \sum_{s=0}^{t_{s_1}-1} \eta_s}{\gamma \tilde{\mathcal{L}}}$ .

**Convergence of  $\frac{\mathcal{G}(\mathbf{W}_t)}{\mathcal{L}(\mathbf{W}_t)}$**  Given  $\mathcal{G}(\mathbf{W}_{t^*}) \leq \frac{\tilde{\mathcal{L}}}{2} \leq \frac{1}{2n}$ , we obtain that  $\mathcal{L}(\mathbf{W}_t) \leq \mathcal{L}(\mathbf{W}_{t^*}) \leq 2\mathcal{G}(\mathbf{W}_{t^*}) \leq \tilde{\mathcal{L}}$  for all  $t \geq t_{s_2}$ , where the first and second inequalities are due to monotonicity in the risk and Lemma 18, respectively. Thus, the other sufficient condition  $\mathcal{L}(\mathbf{W}_t) \leq \frac{\log 2}{n}$  in Lemma 18 is satisfied, from which we conclude that  $\mathcal{L}(\mathbf{W}_t) \leq 2\mathcal{G}(\mathbf{W}_t)$  for all  $t \geq t_{s_2}$ . Substituting this into (24), we obtain for all  $t > t_{s_2}$

$$\mathcal{L}(\mathbf{W}_t) \leq (1 - \frac{\gamma \eta_{t-1}}{4}) \mathcal{L}(\mathbf{W}_{t-1}) \leq \mathcal{L}(\mathbf{W}_{t_{s_2}}) e^{-\frac{\gamma}{4} \sum_{s=t_{s_2}}^{t-1} \eta_s} \leq \tilde{\mathcal{L}} e^{-\frac{\gamma}{4} \sum_{s=t_{s_2}}^{t-1} \eta_s}$$

Then, by Lemma 18, we obtain

$$\frac{\mathcal{G}(\mathbf{W}_t)}{\mathcal{L}(\mathbf{W}_t)} \geq 1 - \frac{n\mathcal{L}(\mathbf{W}_t)}{2} \geq 1 - \frac{n\tilde{\mathcal{L}} e^{-\frac{\gamma}{4} \sum_{s=t_{s_2}}^{t-1} \eta_s}}{2} \geq 1 - e^{-\frac{\gamma}{4} \sum_{s=t_{s_2}}^{t-1} \eta_s}. \quad (25)$$

728

**Margin Convergence** Finally, we combine Lemma 22, Lemma 23, and (25) to obtain

$$\begin{aligned} \left| \frac{\min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i}{\|\mathbf{W}_t\|} - \gamma \right| &\leq \frac{\gamma (\|\mathbf{W}_0\| + \sum_{s=t_{s_2}}^{t-1} \eta_s e^{-\frac{\gamma}{4} \sum_{\tau=t_{s_2}}^{s-1} \eta_\tau} + \sum_{s=0}^{t_{s_2}-1} \eta_s) + \alpha_{s_2} \sum_{s=t_{s_2}}^{t-1} \eta_s^2}{\|\mathbf{W}_0\| + \sum_{s=0}^{t-1} \eta_s} \\ &\leq \mathcal{O} \left( \frac{\sum_{s=t_{s_2}}^{t-1} \eta_s e^{-\frac{\gamma}{4} \sum_{\tau=t_{s_2}}^{s-1} \eta_\tau} + \sum_{s=0}^{t_{s_2}-1} \eta_s + \sum_{s=t_{s_2}}^{t-1} \eta_s^2}{\sum_{s=0}^{t-1} \eta_s} \right) \end{aligned}$$

730

□

Next, we explicitly upper bound  $t_{s_2}$  in Theorem 3 to derive the margin convergence rates of NSD.

**Corollary 1.** Consider learning rate schedule of the form  $\eta_t = \Theta(\frac{1}{t^a})$  where  $a \in (0, 1]$ , under the same setting as Theorem 3, then we have for SignGD

$$\left| \frac{\min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i}{\|\mathbf{W}_t\|} - \gamma \right| = \begin{cases} \mathcal{O}(\frac{t^{1-2a} + n}{t^{1-a}}) & \text{if } a < \frac{1}{2} \\ \mathcal{O}(\frac{\log t + n}{t^{1/2}}) & \text{if } a = \frac{1}{2} \\ \mathcal{O}(\frac{n}{t^{1-a}}) & \text{if } \frac{1}{2} < a < 1 \\ \mathcal{O}(\frac{n}{\log t}) & \text{if } a = 1 \end{cases}$$

*Proof.* Recall that  $t_{s_1} = (\frac{4B^2 e^{2B\eta_0}}{\gamma})^{\frac{1}{a}} =: C_{s_1}$ , and the condition on  $t_{s_2}$  is  $\sum_{s=t_{s_1}}^{t_{s_2}} \eta_s \geq \frac{4\mathcal{L}(\mathbf{W}_0) + 8B \sum_{s=0}^{t_{s_1}-1} \eta_s}{\gamma \tilde{\mathcal{L}}}$ , where  $\tilde{\mathcal{L}} = \frac{\log 2}{n}$ . We can apply integral approximations to the terms that involve sums of learning rates to obtain

$$t_{s_2} \leq C_{s_2} n^{\frac{1}{1-a}} t_{s_1} + C_{s_3} n^{\frac{1}{1-a}} \mathcal{L}(\mathbf{W}_0)^{\frac{1}{1-a}}.$$

Given  $t_{s_1}$  is some constant, this further implies that

$$\sum_{s=0}^{t_{s_2}-1} \eta_s = \mathcal{O}(t_{s_2}^{1-a}) = \mathcal{O}(n + n\mathcal{L}(\mathbf{W}_0)).$$

Next, we focus on the term  $\sum_{s=t_{s_2}}^{t-1} \eta_s^2$ . For  $a > \frac{1}{2}$ , this term can be bounded by some constant.

For  $a < \frac{1}{2}$ , we have  $\sum_{s=t_{s_2}}^{t-1} \eta_s^2 = \mathcal{O}(t^{1-2a})$ , and it evaluates to  $\mathcal{O}(\log t)$  for  $a = \frac{1}{2}$ . Finally,

we have that  $\sum_{s=0}^{t-1} \eta_s = \mathcal{O}(t^{1-a})$  for  $a < 1$  and  $\sum_{s=0}^{t-1} \eta_s = \mathcal{O}(\log t)$  for  $a = 1$ . The term

$\sum_{s=t_{s_2}}^{t-1} \eta_s e^{-\frac{\gamma}{4} \sum_{\tau=t_{s_2}}^{s-1} \eta_\tau}$  is bounded by some constant as shown in Zhang et al. [60, Corollary 4.7]. □

## 743 E Implicit Bias of Normalized Momentum Steepest Descent

744 Recall that  $\|\cdot\|$  refer to either entry-wise or Schatten p-norm with its dual norm denoted as  $\|\cdot\|_*$ .

745 **Lemma 24.** Consider the following  $\mathbf{W}^\dagger := \mathbf{W} - \eta\Delta$ , where  $\Delta \in \mathbb{R}^{k \times d}$  is defined in (4). Let  
 746  $\mathbf{M} \in \mathbb{R}^{k \times d}$  be any matrix. It holds:

$$\langle \nabla \mathcal{L}(\mathbf{W}), \mathbf{W}^\dagger - \mathbf{W} \rangle \leq 2\eta \|\Omega\|_{\text{sum}} - \eta\gamma\mathcal{G}(\mathbf{W}),$$

747 where  $\Omega$  is defined to be  $\Omega := \mathbf{M} - \nabla \mathcal{L}(\mathbf{W})$ .

748 *Proof.* We define  $\Omega := \mathbf{M} - \nabla \mathcal{L}(\mathbf{W})$  to obtain

$$\begin{aligned} \langle \nabla \mathcal{L}(\mathbf{W}), \mathbf{W}^\dagger - \mathbf{W} \rangle &= \langle \nabla \mathcal{L}(\mathbf{W}) - \mathbf{M}, \mathbf{W}^\dagger - \mathbf{W} \rangle + \langle \mathbf{M}, \mathbf{W}^\dagger - \mathbf{W} \rangle \\ &= -\eta \langle \nabla \mathcal{L}(\mathbf{W}) - \mathbf{M}, \Delta \rangle - \eta \langle \mathbf{M}, \Delta \rangle \\ &\stackrel{(a)}{\leq} \eta \|\nabla \mathcal{L}(\mathbf{W}) - \mathbf{M}\|_* \|\Delta\| - \eta \|\mathbf{M}\|_* \\ &\stackrel{(b)}{\leq} \eta \|\mathbf{M} - \nabla \mathcal{L}(\mathbf{W})\|_* - \eta \|\mathbf{M} - \nabla \mathcal{L}(\mathbf{W}) + \nabla \mathcal{L}(\mathbf{W})\|_* \\ &\stackrel{(c)}{\leq} \eta \|\Omega\|_{\text{sum}} - \eta \|\Omega - (-\nabla \mathcal{L}(\mathbf{W}))\|_* \\ &\stackrel{(d)}{\leq} \eta \|\Omega\|_{\text{sum}} - \eta (\|\nabla \mathcal{L}(\mathbf{W})\|_* - \|\Omega\|_*) \\ &= 2\eta \|\Omega\|_{\text{sum}} - \eta \|\nabla \mathcal{L}(\mathbf{W})\|_* \\ &\stackrel{(e)}{\leq} 2\eta \|\Omega\|_{\text{sum}} - \eta\gamma\mathcal{G}(\mathbf{W}), \end{aligned}$$

749 where (a) is by Cauchy Schwarz inequality and  $\langle \mathbf{M}, \Delta \rangle = \|\mathbf{M}\|_*$ , (b) is by  $\|\Delta\| \leq 1$ , (c) is via  
 750 Lemma 11, (d) is by reverse triangle inequality, and (e) is via Lemma 16.  $\square$

751 The following Lemma bounds the entries of the momentum  $(\mathbf{M}_t)$  of NMD in terms of the product of  
 752  $\eta_t$  with the sum of  $\mathcal{G}_c(\mathbf{W}_t)$  and  $\mathcal{Q}_c(\mathbf{W}_t)$ .

753 **Lemma 25.** Suppose that Ass. 1, 2, 3, and 4 hold. Let  $c \in [k]$  and  $j \in [d]$ . There exists time  $t_0$  such  
 754 that for all  $t \geq t_0$ :

$$|\mathbf{M}_t[c, j] - (1 - \beta_1^{t+1})\nabla \mathcal{L}(\mathbf{W}_t)[c, j]| \leq \alpha_M \eta_t (\mathcal{G}_c(\mathbf{W}_t) + \mathcal{Q}_c(\mathbf{W}_t)),$$

755 where  $\alpha_M := B(1 - \beta_1)c_2$ .

756 *Proof.* For any fixed  $c \in [k]$  and  $j \in [d]$ ,

$$\begin{aligned} |\mathbf{M}_t[c, j] - (1 - \beta_1^{t+1})\nabla \mathcal{L}(\mathbf{W}_t)[c, j]| &= \left| \sum_{\tau=0}^t (1 - \beta_1)\beta_1^\tau (\nabla \mathcal{L}(\mathbf{W}_{t-\tau})[c, j] - \nabla \mathcal{L}(\mathbf{W}_t)[c, j]) \right| \\ &\leq \sum_{\tau=0}^t (1 - \beta_1)\beta_1^\tau \underbrace{|\nabla \mathcal{L}(\mathbf{W}_{t-\tau})[c, j] - \nabla \mathcal{L}(\mathbf{W}_t)[c, j]|}_{\clubsuit}. \end{aligned} \tag{26}$$

757 We first notice that for any  $\mathbf{W} \in \mathbb{R}^{k \times d}$ , we have  $\nabla \mathcal{L}(\mathbf{W})[c, j] = \mathbf{e}_c^T \nabla \mathcal{L}(\mathbf{W}) \mathbf{e}_j =$   
 758  $-\frac{1}{n} \sum_{i \in [n]} \mathbf{e}_c^T (\mathbf{e}_{y_i} - \mathbb{S}(\mathbf{W} \mathbf{h}_i)) \mathbf{h}_i^T \mathbf{e}_j = -\frac{1}{n} \sum_{i \in [n]} \mathbf{e}_c^T (\mathbf{e}_{y_i} - \mathbb{S}(\mathbf{W} \mathbf{h}_i)) h_{ij}$ . Then, the gradient

759 difference term becomes

$$\begin{aligned}
\clubsuit &= \left| -\frac{1}{n} \sum_{i \in [n]} e_c^T (e_{y_i} - \mathbb{S}(\mathbf{W}_{t-\tau} \mathbf{h}_i)) h_{ij} + \frac{1}{n} \sum_{i \in [n]} e_c^T (e_{y_i} - \mathbb{S}(\mathbf{W}_t \mathbf{h}_i)) h_{ij} \right| \\
&= \left| \frac{1}{n} \sum_{i \in [n]} e_c^T (\mathbb{S}(\mathbf{W}_{t-\tau} \mathbf{h}_i) - \mathbb{S}(\mathbf{W}_t \mathbf{h}_i)) h_{ij} \right| \\
&= \left| \frac{1}{n} \sum_{i \in [n]} (\mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i) - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i)) h_{ij} \right| \\
&\leq B \frac{1}{n} \sum_{i \in [n]} |\mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i) - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i)| \\
&= B \underbrace{\frac{1}{n} \sum_{i \in [n], y_i \neq c} |\mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i) - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i)|}_{\clubsuit_1} + B \underbrace{\frac{1}{n} \sum_{i \in [n], y_i = c} |\mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i) - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i)|}_{\clubsuit_2}
\end{aligned}$$

760 Next, we link the  $\clubsuit_1$  and  $\clubsuit_2$  terms with  $\mathcal{G}(\mathbf{W})$ . Starting with the first term, we obtain:

$$\begin{aligned}
\clubsuit_1 &= \frac{1}{n} \sum_{i \in [n], y_i \neq c} |\mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i)| \left| \frac{\mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i)}{\mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i)} - 1 \right| \\
&\stackrel{(a)}{\leq} \frac{1}{n} \sum_{i \in [n], y_i \neq c} |\mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i)| (e^{2\|\mathbf{W}_{t-\tau} - \mathbf{W}_t\|_\infty} - 1) \\
&\stackrel{(b)}{\leq} \frac{1}{n} \sum_{i \in [n], y_i \neq c} |\mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i)| (e^{2B\|\mathbf{W}_{t-\tau} - \mathbf{W}_t\|_{\max}} - 1) \\
&\stackrel{(c)}{\leq} (e^{2B \sum_{s=1}^{\tau} \eta_{t-s}} - 1) \left( \frac{1}{n} \sum_{i \in [n], y_i \neq c} |\mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i)| \right) \\
&\stackrel{(d)}{\leq} (e^{2B \sum_{s=1}^{\tau} \eta_{t-s}} - 1) \mathcal{Q}_c(\mathbf{W}_t),
\end{aligned}$$

761 where (a) is by Lemma 15, (b) is by  $\|\mathbf{h}_i\|_1 \leq B$  for all  $i \in [n]$ , (c) is by (3) and triangle inequality,  
762 and (d) is by  $\|\Delta_{t-s}\|_{\max} \leq \|\Delta_{t-s}\| \leq 1$  (for any entry-wise or Schatten p-norm) and the definition  
763 of  $\mathcal{G}(\mathbf{W}_t)$ . For the second term, we obtain:

$$\begin{aligned}
\clubsuit_2 &= \frac{1}{n} \sum_{i \in [n], y_i = c} |\mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i) - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i)| \\
&= \frac{1}{n} \sum_{i \in [n], y_i = c} |\mathbb{S}_{y_i}(\mathbf{W}_{t-\tau} \mathbf{h}_i) - 1 + 1 - \mathbb{S}_{y_i}(\mathbf{W}_t \mathbf{h}_i)| \\
&= \frac{1}{n} \sum_{i \in [n], y_i = c} (1 - \mathbb{S}_{y_i}(\mathbf{W}_t \mathbf{h}_i)) \left| \frac{\mathbb{S}_{y_i}(\mathbf{W}_{t-\tau} \mathbf{h}_i) - 1}{1 - \mathbb{S}_{y_i}(\mathbf{W}_t \mathbf{h}_i)} + 1 \right| \\
&= \frac{1}{n} \sum_{i \in [n], y_i = c} (1 - \mathbb{S}_{y_i}(\mathbf{W}_t \mathbf{h}_i)) \left| \frac{1 - \mathbb{S}_{y_i}(\mathbf{W}_{t-\tau} \mathbf{h}_i)}{1 - \mathbb{S}_{y_i}(\mathbf{W}_t \mathbf{h}_i)} - 1 \right| \\
&\stackrel{(e)}{\leq} \frac{1}{n} \sum_{i \in [n], y_i = c} (1 - \mathbb{S}_{y_i}(\mathbf{W}_t \mathbf{h}_i)) (e^{2\|\mathbf{W}_{t-\tau} - \mathbf{W}_t\|_\infty} - 1) \\
&\stackrel{(f)}{\leq} (e^{2B \sum_{s=1}^{\tau} \eta_{t-s}} - 1) \mathcal{G}_c(\mathbf{W}_t),
\end{aligned}$$

764 where (e) is by Lemma 15, and (f) is by the same approach taken for  $\clubsuit_1$ . Based on the upper bounds  
765 for  $\clubsuit_1$  and  $\clubsuit_2$ , we obtain the following:  $\clubsuit \leq 2B(e^{2\alpha B \sum_{s=1}^{\tau} \eta_{t-s}} - 1)(\mathcal{G}_c(\mathbf{W}_t) + \mathcal{Q}_c(\mathbf{W}_t))$ . Then,

we substitute this into (26) to obtain:

$$\begin{aligned} |\mathbf{M}_t[c, j] - (1 - \beta_1^{t+1})\nabla\mathcal{L}(\mathbf{W}_t)[c, j]| &\leq B(1 - \beta_1)(\mathcal{G}_c(\mathbf{W}_t) + \mathcal{Q}_c(\mathbf{W}_t)) \sum_{\tau=0}^t \beta_1^\tau (e^{2B\sum_{s=1}^\tau \eta_{t-s}} - 1) \\ &\stackrel{(g)}{\leq} B(1 - \beta_1)c_2\eta_t(\mathcal{G}_c(\mathbf{W}_t) + \mathcal{Q}_c(\mathbf{W}_t)), \end{aligned}$$

where (g) is by the Assumption 3.  $\square$

**Lemma 26.** Let  $\mathbf{\Omega}_t = \mathbf{M}_t - \nabla\mathcal{L}(\mathbf{W}_t)$ , where  $\mathbf{M}_t$  is defined in (5). Then, it holds

$$\|\mathbf{\Omega}_t\|_{\text{sum}} \leq 2B\beta_1^{t/2}\mathcal{G}(\mathbf{W}_t) + 2\alpha_M d\eta_t\mathcal{G}(\mathbf{W}_t),$$

where  $\alpha_M := B(1 - \beta_1)c_2$ .

*Proof.* For simplicity, we drop the subscripts  $t$ . Denote  $\mathcal{T}_c(\mathbf{W}) := \mathcal{G}_c(\mathbf{W}) + \mathcal{Q}_c(\mathbf{W})$ . Then, by Lemma 25, we have for any  $c \in [k]$  and  $j \in [d]$ :

$$\begin{aligned} \mathbf{M}[c, j] &= (1 - \beta_1^{t+1})\nabla\mathcal{L}(\mathbf{W})[c, j] + \alpha_M\eta\mathcal{T}_c(\mathbf{W})\epsilon_{m,c,j} \\ &= \nabla\mathcal{L}(\mathbf{W})[c, j] - \beta_1^{t+1}\nabla\mathcal{L}(\mathbf{W})[c, j] + \alpha_M\eta\mathcal{T}_c(\mathbf{W})\epsilon_{m,c,j}, \end{aligned}$$

where  $\alpha_M := B(1 - \beta_1)c_2$  and  $\epsilon_{m,c,j}$  is some constant s.t.  $|\epsilon_{m,c,j}| \leq 1$ . Recall that  $\mathbf{\Omega} := \mathbf{M} - \nabla\mathcal{L}(\mathbf{W})$ , then we have

$$\begin{aligned} |\mathbf{\Omega}[c, j]| &= |\mathbf{M}[c, j] - \nabla\mathcal{L}(\mathbf{W})[c, j]| \\ &= |-\beta_1^{t+1}\nabla\mathcal{L}(\mathbf{W})[c, j] + \alpha_M\eta\mathcal{T}_c(\mathbf{W})\epsilon_{m,c,j}| \\ &\leq \beta_1^{t+1}|\nabla\mathcal{L}(\mathbf{W})[c, j]| + \alpha_M\eta\mathcal{T}_c(\mathbf{W}). \end{aligned}$$

This implies the following:

$$\begin{aligned} \|\mathbf{\Omega}\|_{\text{sum}} &= \sum_{c,j} |\mathbf{\Omega}[c, j]| \leq \beta_1^{t+1} \sum_{c,j} |\nabla\mathcal{L}(\mathbf{W})[c, j]| + \alpha_M\eta \sum_{c,j} \mathcal{T}_c(\mathbf{W}) \\ &= \beta_1^{t+1}\|\nabla\mathcal{L}(\mathbf{W})\|_{\text{sum}} + 2\alpha_M d\eta\mathcal{G}(\mathbf{W}) \\ &\leq 2B\beta_1^{t/2}\mathcal{G}(\mathbf{W}) + 2\alpha_M d\eta\mathcal{G}(\mathbf{W}), \end{aligned}$$

where in the last inequality we have used Lemma 16.  $\square$

**Lemma 27.** Suppose that there exist  $\tilde{t}$  such that  $\mathcal{L}(\mathbf{W}_t) \leq \frac{\log 2}{n}$  for all  $t > \tilde{t}$ , then we have

$$\min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i \geq \gamma \sum_{s=\tilde{t}}^{t-1} \eta_s \frac{\mathcal{G}(\mathbf{W}_s)}{\mathcal{L}(\mathbf{W}_s)} - a_2 \sum \eta_s^2 - Q$$

where  $a_2 = (4\alpha_M + 2B^2e^{2B\eta_0})d$  and  $Q = 4B\eta_0 \frac{1}{1-\beta_1^{1/2}}$ .

*Proof.* We follow a similar approach as Lemma 21 to show the descent of NMD. Specifically, we apply Lemma 24 to bound the first-order term. For the Hessian term, we apply Lemma 14 and Lemma 20 similar to NSD. Then, we can obtain the following

$$\begin{aligned} \mathcal{L}(\mathbf{W}_{t+1}) &\leq \mathcal{L}(\mathbf{W}_t) - \eta_t\gamma\mathcal{G}(\mathbf{W}_t) + 2\eta_t\|\mathbf{\Omega}_t\|_{\text{sum}} + 2\eta_t^2B^2e^{2B\eta_0}\mathcal{G}(\mathbf{W}_t) \\ &\stackrel{(a)}{\leq} \mathcal{L}(\mathbf{W}_t) - \eta_t\gamma\mathcal{G}(\mathbf{W}_t) + 4B\beta_1^{t/2}\eta_t\mathcal{G}(\mathbf{W}_t) + 4\alpha_M\eta_t^2d\mathcal{G}(\mathbf{W}_t) + 2\eta_t^2B^2e^{2B\eta_0}\mathcal{G}(\mathbf{W}_t) \\ &\stackrel{(b)}{\leq} \mathcal{L}(\mathbf{W}_t) - \eta_t\gamma\mathcal{G}(\mathbf{W}_t) + a_1\beta_1^{t/2}\eta_t\mathcal{G}(\mathbf{W}_t) + a_2\eta_t^2d\mathcal{G}(\mathbf{W}_t) \\ &\leq \mathcal{L}(\mathbf{W}_{\tilde{t}}) \exp\left(-\gamma \sum_{s=\tilde{t}}^t \eta_s \frac{\mathcal{G}(\mathbf{W}_s)}{\mathcal{L}(\mathbf{W}_s)}\right) + a_1 \sum_{s=\tilde{t}}^t \beta_1^{s/2}\eta_s + a_2d \sum_{s=\tilde{t}}^t \eta_s^2 \\ &\stackrel{(c)}{\leq} \frac{\log 2}{n} \exp\left(-\gamma \sum_{s=\tilde{t}}^t \eta_s \frac{\mathcal{G}(\mathbf{W}_s)}{\mathcal{L}(\mathbf{W}_s)}\right) + a_2d \sum_{s=\tilde{t}}^t \eta_s^2 + Q, \end{aligned}$$

781 where (a) is by Lemma 24. In (b), we have defined  $a_1 := 4B$  and  $a_2 = (4\alpha_M + 2B^2e^{2B\eta_0})d$ . In  
 782 (c), we have used the assumption and defined  $Q := a_1\eta_0\frac{1}{1-\beta_1^{1/2}} \geq a_1\sum_{s=\tilde{t}}^t\beta_1^{s/2}\eta_s$ . The rest of the  
 783 proof follows the same steps as Lemma 22.  $\square$

784 **Theorem 4.** Suppose that Ass. 1, 2, 3, and 4 hold. Set learning rate  $\eta_t = \Theta(\frac{1}{t^{1/2}})$ . The margin gap  
 785 of NMD's iterates satisfy

$$\gamma - \frac{\min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i}{\|\mathbf{W}_t\|} \leq O\left(\frac{d \log t + dn}{t^{1/2}}\right).$$

786 *Proof.* Given the updates of NMD are normalized (i.e.,  $\|\Delta\| \leq 1$ ), we can obtain the following via  
 787 Lemma 27:

$$\begin{aligned} \gamma - \frac{\min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i}{\|\mathbf{W}_t\|} &\leq \frac{\gamma(\|\mathbf{W}_0\| + \sum_{s=0}^{t_2-1} \eta_s + \sum_{s=t_2}^{t-1} \eta_s e^{\frac{\gamma}{4} \sum_{\tau=t_2}^{s-1} \eta_\tau}) + a_2 d \sum_{s=t_2}^{t-1} \eta_s^2 + Q}{\|\mathbf{W}_0\| + \sum_{s=0}^{t-1} \eta_s} \\ &\leq O\left(\frac{\sum_{s=t_2}^{t-1} \eta_s e^{-\frac{\gamma}{4} \sum_{\tau=t_2}^{s-1} \eta_\tau} + \sum_{s=0}^{t_2-1} \eta_s + d \sum_{s=t_2}^{t-1} \eta_s^2}{\sum_{s=0}^{t-1} \eta_s}\right). \end{aligned}$$

788 Then, we follow the same approach as Corollary 1 for a decreasing learning rate of the form  
 789  $\eta_t = \Theta(\frac{1}{t^a})$ . Specifically, we have  $t_1 = \Theta(d^{1/a})$  and  $t_2 \leq C_1 n^{\frac{1}{1-a}} t_1 + C_2 n^{\frac{1}{1-a}} L(\mathbf{W}_0)^{\frac{1}{1-a}}$ . This  
 790 leads to

$$\sum_{s=0}^{t_2-1} \eta_s = O(t_2^{1-a}) = nt_1^{1-a} + nL(\mathbf{W}_0) + d \log(t).$$

791 Thus, we have the margin gap upper bounded by  $O(\frac{nd+d \log(t)}{t^{1/2}})$ .  $\square$

## 792 F Other multiclass loss functions

### 793 F.1 Exponential Loss

794 The multiclass exponential loss is given as

$$\mathcal{L}_{\text{exp}}(\mathbf{W}) := \frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} \exp(-(e_{y_i} - e_c)^\top \mathbf{W} \mathbf{h}_i).$$

795 The gradient of  $\mathcal{L}_{\text{exp}}(\mathbf{W})$  is

$$\nabla \mathcal{L}_{\text{exp}}(\mathbf{W}) = \frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} -\exp(-(e_{y_i} - e_c)^\top \mathbf{W} \mathbf{h}_i) (e_{y_i} - e_c) \mathbf{h}_i^\top.$$

796 Thus, for any matrix  $\mathbf{A} \in \mathbb{R}^{k \times d}$ , we have

$$\langle \mathbf{A}, -\nabla \mathcal{L}_{\text{exp}}(\mathbf{W}) \rangle = \frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} \exp(-(e_{y_i} - e_c)^\top \mathbf{W} \mathbf{h}_i) \cdot (e_{y_i} - e_c)^\top \mathbf{A} \mathbf{h}_i.$$

797 This motivates us to define  $\mathcal{G}(\mathbf{W})$  as

$$\mathcal{G}_{\text{exp}}(\mathbf{W}) = \frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} \exp(-(e_{y_i} - e_c)^\top \mathbf{W} \mathbf{h}_i),$$

798 from which we recognize that  $\mathcal{G}_{\text{exp}}(\mathbf{W}) = \mathcal{L}_{\text{exp}}(\mathbf{W})$ . Then, the proof follows similar steps as the  
 799 CE loss.



## 800 F.2 PairLogLoss

801 The PairLogLoss loss [55] is given as

$$\mathcal{L}_{\text{pll}}(\mathbf{W}) := \frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} \log(1 + \exp(-(e_{y_i} - e_c)^\top \mathbf{W} \mathbf{h}_i)) .$$

802 Note that  $\mathcal{L} = \frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} f((e_{y_i} - e_c)^\top \mathbf{W} \mathbf{h}_i)$  where  $f(t) := \log(1 + e^{-t})$  denotes the  
803 logistic loss. Therefore, the Taylor expansion of PLL writes:

$$\begin{aligned} \mathcal{L}_{\text{pll}}(\mathbf{W} + \Delta) &= \mathcal{L}(\mathbf{W}) + \frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} f'((e_{y_i} - e_c)^\top \mathbf{W} \mathbf{h}_i) \cdot (e_{y_i} - e_c)^\top \Delta \mathbf{h}_i \\ &\quad + \frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} f''((e_{y_i} - e_c)^\top \mathbf{W} \mathbf{h}_i) \cdot \mathbf{h}_i^\top \Delta^\top (e_{y_i} - e_c)(e_{y_i} - e_c)^\top \Delta \mathbf{h}_i + o(\|\Delta\|^3) . \end{aligned} \quad (27)$$

804 From the above, the gradient of the PLL loss is:

$$\begin{aligned} \nabla \mathcal{L}_{\text{pll}}(\mathbf{W}) &= \frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} f'((e_{y_i} - e_c)^\top \mathbf{W} \mathbf{h}_i) \cdot (e_{y_i} - e_c) \mathbf{h}_i^\top \\ &= \frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} \frac{-\exp(-(e_{y_i} - e_c)^\top \mathbf{W} \mathbf{h}_i)}{1 + \exp(-(e_{y_i} - e_c)^\top \mathbf{W} \mathbf{h}_i)} (e_{y_i} - e_c) \mathbf{h}_i^\top \end{aligned} \quad (28)$$

805 Thus, for any matrix  $\mathbf{A} \in \mathbb{R}^{k \times d}$ ,

$$\langle \mathbf{A}, -\nabla \mathcal{L}_{\text{pll}}(\mathbf{W}) \rangle = \frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} |f'((e_{y_i} - e_c)^\top \mathbf{W} \mathbf{h}_i)| \cdot (e_{y_i} - e_c)^\top \mathbf{A} \mathbf{h}_i . \quad (29)$$

806 This motivates us to define

$$\mathcal{G}_{\text{pll}}(\mathbf{W}) = \frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} |f'(-(e_{y_i} - e_c)^\top \mathbf{W} \mathbf{h}_i)| = \frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} \frac{\exp(-(e_{y_i} - e_c)^\top \mathbf{W} \mathbf{h}_i)}{1 + \exp(-(e_{y_i} - e_c)^\top \mathbf{W} \mathbf{h}_i)} \quad (30)$$

807 **Lemma 28** (Analogue of Lemma 16 for PLL). *For any  $\mathbf{W}$ , the PairLogLoss (PLL) satisfies:*

$$2B \cdot \mathcal{G}_{\text{pll}}(\mathbf{W}) \geq \|\nabla \mathcal{L}_{\text{pll}}(\mathbf{W})\| \geq \gamma \cdot \mathcal{G}_{\text{pll}}(\mathbf{W}) .$$

808 *Proof.* The lower bound follows immediately from (29) and expressing  $\|\nabla \mathcal{L}_{\text{pll}}(\mathbf{W})\|_* =$   
809  $\max_{\|\mathbf{A}\| \leq 1} \langle \mathbf{A}, -\nabla \mathcal{L}_{\text{pll}}(\mathbf{W}) \rangle$ . The lower bound follows from triangle inequality applied to (28):

$$\|\nabla \mathcal{L}_{\text{pll}}(\mathbf{W})\|_{\text{sum}} \leq \frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} |f'(-(e_{y_i} - e_c)^\top \mathbf{W} \mathbf{h}_i)| \|e_{y_i} - e_c\|_1 \|\mathbf{h}_i\|_1 \leq 2B \cdot \mathcal{G}(\mathbf{W}) ,$$

810 and use the relationships in (7), i.e.  $\|\nabla \mathcal{L}_{\text{pll}}(\mathbf{W})\| \leq \|\nabla \mathcal{L}_{\text{pll}}(\mathbf{W})\|_{\text{sum}}$  for any entry-wise or Schatten  
811  $p$ -norm with  $p \geq 1$ .  $\square$

812 For bounding with  $\mathcal{G}(\mathbf{W})$  the second-order term in the Taylor expansion of PLL, note the following.  
813 First, for all  $i \in [n], c \neq y_i$ :

$$\begin{aligned} \mathbf{h}_i^\top \Delta^\top (e_{y_i} - e_c)(e_{y_i} - e_c)^\top \Delta \mathbf{h}_i &= \langle (e_{y_i} - e_c)(e_{y_i} - e_c)^\top, \Delta \mathbf{h}_i \mathbf{h}_i^\top \Delta^\top \rangle \\ &\leq \|(e_{y_i} - e_c)(e_{y_i} - e_c)^\top\|_{\text{sum}} \|\Delta \mathbf{h}_i \mathbf{h}_i^\top \Delta^\top\|_{\text{max}} \\ &\leq \|e_{y_i} - e_c\|_1^2 \cdot (\|\Delta \mathbf{h}_i\|_\infty)^2 \\ &\leq 4 \cdot (\|\Delta\|_{\text{max}})^2 \cdot \|\mathbf{h}_i\|_1^2 \leq 4B^2 (\|\Delta\|_{\text{max}})^2 \\ &\leq 4B^2 \|\Delta\|^2 . \end{aligned}$$

814 Second, the (easy to check) property of logistic loss that  $f''(t) \leq |f'(t)|$ . Putting these together:

$$\frac{1}{n} \sum_{i \in [n]} \sum_{c \neq y_i} f''((\mathbf{e}_{y_i} - \mathbf{e}_c)^\top \mathbf{W} \mathbf{h}_i) \cdot \mathbf{h}_i^\top \Delta^\top (\mathbf{e}_{y_i} - \mathbf{e}_c) (\mathbf{e}_{y_i} - \mathbf{e}_c)^\top \Delta \mathbf{h}_i \leq 4B^2 \cdot \mathcal{G}(\mathbf{W}) \cdot (\|\Delta\|)^2.$$

815 Finally, we verify PLL satisfies Lemma 18.

816 **Lemma 29** (Analogue of Lemma 18 for PLL). *Let  $\mathbf{W} \in \mathbb{R}^{k \times d}$ , we have*

817 (i)  $1 \geq \frac{\mathcal{G}_{\text{pll}}(\mathbf{W})}{\mathcal{L}_{\text{pll}}(\mathbf{W})} \geq 1 - \frac{n\mathcal{L}_{\text{pll}}(\mathbf{W})}{2}$

818 (ii) *Suppose that  $\mathbf{W}$  satisfies  $\mathcal{L}_{\text{pll}}(\mathbf{W}) \leq \frac{\log 2}{n}$  or  $\mathcal{G}_{\text{pll}}(\mathbf{W}) \leq \frac{1}{2n}$ , then  $\mathcal{L}_{\text{pll}}(\mathbf{W}) \leq 2\mathcal{G}_{\text{pll}}(\mathbf{W})$ .*

819 *Proof.* (i) The upper bound follows by the well-known self-boundedness property of the logistic loss,  
820 namely  $|f'(t)| \leq f(t)$

821 To prove the upper bound, it suffices to prove for  $x > 0$ :

$$\frac{x}{1+x} \geq \log(1+x) - \frac{1}{2} \log^2(1+x). \quad (31)$$

822 The general case follows by summing over  $x_{ic} = \exp(-(\mathbf{e}_{y_i} - \mathbf{e}_c)^\top \mathbf{W} \mathbf{h}_i)$ ,  $i \in [n], c \neq y_i$  since  
823 then we have

$$\begin{aligned} \mathcal{G}(\mathbf{W}) &= \sum_{i \in [n]} \sum_{c \neq y_i} \frac{x_{ic}}{1+x_{ic}} \geq \sum_{i \in [n]} \sum_{c \neq y_i} \log(1+x_{ic}) - \frac{1}{2} \sum_{i \in [n]} \sum_{c \neq y_i} \log^2(1+x_{ic}) \\ &\geq \sum_{i \in [n]} \sum_{c \neq y_i} \log(1+x_{ic}) - \frac{1}{2} \left( \sum_{i \in [n]} \sum_{c \neq y_i} \log(1+x_{ic}) \right)^2, \end{aligned}$$

824 where the last line used  $\log(1+x_{ic}) \geq 0$ . For (19), let  $a = \log(1+x) > 0$ . The inequality becomes  
825  $e^{-a} \leq 1 - a + a^2/2$ , which holds for  $a > 0$  by the second-order Taylor expansion of  $e^{-a}$  around 0.

826 (ii) Denote  $\mathcal{L} := \mathcal{L}_{\text{pll}}$  and  $\mathcal{G} := \mathcal{G}_{\text{pll}}$ . Given  $\mathcal{L} \leq \frac{\log(2)}{n} \leq \frac{1}{n}$ , we have  $1 - \frac{n\mathcal{L}}{2} \geq \frac{1}{2}$ , then the first part  
827 follows from (i). For the second part, denote  $l_{ic} := (\mathbf{e}_{y_i} - \mathbf{e}_c)^\top \mathbf{W} \mathbf{h}_i$ ,  $i \in [n], c \neq y_i$ . For  $\mathcal{L} \leq 2\mathcal{G}$   
828 to hold, it is sufficient to show that  $\log(1 + e^{-l_{ic}}) \leq 2 \frac{e^{-l_{ic}}}{1+e^{-l_{ic}}}$  for all  $i \in [n], c \neq y_i$ . This holds true  
829 when  $l_{ic} \geq -1.366$ , which is clearly satisfied given the assumption  $\mathcal{G} \leq \frac{1}{2n}$  implying  $l_{ic} \geq 0$ .  $\square$

830 **Lemma 30** (Analogue of Lemma 20 for PLL). *For any  $\psi \in [0, 1]$ , we have the following:*

$$\frac{\mathcal{G}_{\text{pll}}(\mathbf{W} - \psi\eta\Delta)}{\mathcal{G}_{\text{pll}}(\mathbf{W})} \leq e^{2B\psi\|\Delta\|\mathbf{W}} + 2$$

831 *Proof.* For logistic loss  $f(z) = \log(1 + e^{-z})$ , for any  $z_1, z_2 \in \mathbb{R}$ , we have the following

$$\begin{aligned} \left| \frac{f'(z_1)}{f'(z_2)} \right| &= \left| \frac{1 + e^{z_2}}{1 + e^{z_1}} \right| = \left| \frac{1 + e^{z_2} - e^{z_1} + e^{z_1}}{1 + e^{z_1}} \right| \\ &= \left| \frac{e^{z_2} - e^{z_1}}{1 + e^{z_1}} + 1 \right| \leq \left| \frac{e^{z_2} - e^{z_1}}{1 + e^{z_1}} \right| + 1 \\ &\leq |e^{z_2-z_1} - 1| + 1 \\ &\leq e^{|z_2-z_1|} + 2. \end{aligned}$$

832 Denote  $x_{ic}^{\mathbf{W}} := (\mathbf{e}_{y_i} - \mathbf{e}_c)^\top \mathbf{W} \mathbf{h}_i$  and  $x_{ic}^{\mathbf{W}'} := (\mathbf{e}_{y_i} - \mathbf{e}_c)^\top (\mathbf{W} - \psi\eta\Delta) \mathbf{h}_i$ , then we have for  $i \in [n]$ ,  
833  $c \neq y_i$

$$\begin{aligned} \frac{f'(x_{ic}^{\mathbf{W}'})}{f'(x_{ic}^{\mathbf{W}})} &= \left| \frac{f'(x_{ic}^{\mathbf{W}'})}{f'(x_{ic}^{\mathbf{W}})} \right| \leq e^{|x_{ic}^{\mathbf{W}} - x_{ic}^{\mathbf{W}'}|} + 2 = e^{\psi\eta|(\mathbf{e}_c - \mathbf{e}_{y_i})^\top \Delta \mathbf{h}_i|} + 2 = e^{\psi\eta|\langle \Delta, (\mathbf{e}_c - \mathbf{e}_{y_i}) \mathbf{h}_i^\top \rangle|} + 2 \\ &\leq e^{\psi\eta\|\Delta\|_{\max} \|(\mathbf{e}_c - \mathbf{e}_{y_i}) \mathbf{h}_i^\top\|_{\text{sum}}} + 2 \\ &= e^{\psi\eta\|\Delta\|_{\max} \|\mathbf{e}_c - \mathbf{e}_{y_i}\|_{\text{sum}} \|\mathbf{h}_i\|_{\text{sum}}} + 2 \\ &\leq e^{2B\psi\eta\|\Delta\|_{\max}} + 2. \end{aligned}$$

834 This leads to  $\sum_{i \in [n]} \sum_{c \neq y_i} f'(x_{ic}^{\mathbf{W}'}) \leq (e^{2B\psi} \|\Delta \mathbf{W}\|_{\max} + 2) \sum_{i \in [n]} \sum_{c \neq y_i} f'(x_{ic}^{\mathbf{W}})$ . Rearrange  
835 and using the definition of  $\mathcal{G}_{pll}(\mathbf{W})$  and relationships in (7), we obtain the desired.  $\square$

836 **Lemma 31** (Analogue of Lemma 19 for PLL). *Suppose that there exists  $\mathbf{W} \in \mathbb{R}^{k \times d}$  such that*  
837  $\mathcal{L}_{pll}(\mathbf{W}) \leq \frac{\log 2}{n}$ , *then we have*

$$(e_{y_i} - e_c)^T \mathbf{W} \mathbf{h}_i \geq 0, \quad \text{for all } i \in [n] \text{ and for all } c \in [k] \text{ such that } c \neq y_i. \quad (32)$$

838 *Proof.* Denote  $x_{ic} = (e_{y_i} - e_c)^T \mathbf{W} \mathbf{h}_i$ . Then, by the assumption, we have for any  $i \in [n], c \neq y_i$

$$\log(1 + e^{-x_{ic}}) \leq \sum_{i \in [n]} \sum_{c \neq y_i} \log(1 + e^{-x_{ic}}) \leq \log(2).$$

839 This implies that  $x_{ic} \geq 0$  for all  $i \in [n], c \neq y_i$ .  $\square$

840 **Lemma 32** (Analogue of Lemma 17 for PLL). *For any  $\mathbf{W}, \mathbf{W}_0 \in \mathbb{R}^{k \times d}$ , suppose that  $\mathcal{L}(\mathbf{W})$  is*  
841 *convex, we have*

$$|\mathcal{L}_{pll}(\mathbf{W}) - \mathcal{L}_{pll}(\mathbf{W}_0)| \leq 2B \|\mathbf{W} - \mathbf{W}_0\|.$$

842 *Proof.* This lemma is a direct consequence of Lemma 28 and can be proved in the same way as  
843 Lemma 17.  $\square$

844 Thus, we have proved all the Lemmas for  $\mathcal{G}_{pll}(\mathbf{W})$  and its relationships to  $\mathcal{L}_{pll}(\mathbf{W})$  in analogous  
845 to those in section C. The proof of NSD ((3)) with PairLogLoss follow the same steps as with  
846 cross-entropy loss given in section D.

## 847 G Implicit Bias of Adam

848 We consider Adam without the stability constant ( $\epsilon$ ), which performs the following coordinate-wise  
849 updates for iteration  $t \geq 0$  and initialization  $\mathbf{W}_0$  [28]:

$$\mathbf{M}_t = \beta_1 \mathbf{M}_{t-1} + (1 - \beta_1) \nabla \mathcal{L}(\mathbf{W}_t) \quad (33a)$$

$$\mathbf{V}_t = \beta_2 \mathbf{V}_{t-1} + (1 - \beta_2) \nabla \mathcal{L}(\mathbf{W}_t)^2 \quad (33b)$$

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta_t \frac{\mathbf{M}_t}{\sqrt{\mathbf{V}_t}}, \quad (33c)$$

850 where  $\mathbf{M}_t$  and  $\mathbf{V}_t$  are the first and second moment estimates of the gradient (with momentum  
851 parameters  $\beta_1$  and  $\beta_2$ ) respectively. The squaring  $(\cdot)^2$  and dividing  $\cdot$  operations are applied *entry-*  
852 *wise*. In the special case of  $\beta_1 = \beta_2 = 0$ , this simplifies to the SignGD updates.

853 To study the implicit bias of Adam, we further make the following assumption, which ensures  
854 that all entries of the second moment buffer  $\mathbf{V}_t$  of Adam are bounded away from 0 for all  $t \geq 0$ .  
855 Previously used by Zhang et al. [60] in binary classification, this assumption is satisfied when the  
856 data distribution is continuous and non-degenerate. A similar assumption appears in [59].

857 **Assumption 5.** *The Adam initialization satisfies  $\nabla \mathcal{L}(\mathbf{W}_0)[c, j]^2 \geq \omega$  for all  $c \in [k]$  and  $j \in [d]$ .*

858 The proof of Adam follows the similar approach as NSD. The key challenge is to connect  $\mathbf{M}_t$  and  
859  $\mathbf{V}_t$  to a per-class decomposition of  $\mathcal{G}(\mathbf{W}_t)$ . The following Lemma in [60, Lemma 6.5] is useful. It  
860 provides an entry-wise bound on the ratio between the first moment and square root of the second  
861 moment.

862 **Lemma 33.** *Considering the Adam updates given in (5), (33b), and (33c), suppose that  $\beta_1 \leq \beta_2$  and*  
863 *set  $\alpha = \sqrt{\frac{\beta_2(1-\beta_1)^2}{(1-\beta_2)(\beta_2-\beta_1^2)^2}}$ , then we obtain  $\mathbf{M}_t[c, j] \leq \alpha \cdot \sqrt{\mathbf{V}_t[c, j]}$  for all  $c \in [k]$  and  $j \in [d]$ .*

864 The following Lemma bounds the first moment buffer ( $\mathbf{M}_t$ ) of Adam in terms of the product of  $\eta_t$   
865 with the sum of  $\mathcal{G}_c(\mathbf{W}_t)$  and  $\mathcal{Q}_c(\mathbf{W}_t)$ . It is used in the proof of Lemma 36.

866 **Lemma 34.** Let  $c \in [k]$ . Under the same setting as Theorem 5, there exists a time  $t_0$  such that the  
 867 following holds for all  $t \geq t_0$

$$|\mathbf{M}_t[c, j] - (1 - \beta_1^{t+1})\nabla\mathcal{L}(\mathbf{W}_t)[c, j]| \leq \alpha_M \eta_t (\mathcal{G}_c(\mathbf{W}_t) + \mathcal{Q}_c(\mathbf{W}_t)),$$

868 where  $j \in [d]$  and  $\alpha_M$  is some constant that depends on  $B$  and  $\beta_1$ .

869 *Proof.* The proof follows the same steps as Lemma 25 with  $\|\Delta\|$  replaced by  $\left\|\frac{M}{\sqrt{V}}\right\|_{\max}$ .  $\square$

870 The following Lemma bounds the first moment buffer ( $\mathbf{V}_t$ ) of Adam in terms of the product of  $\eta_t$  and  
 871 with  $\mathcal{G}_c(\mathbf{W}_t)$  and  $\mathcal{Q}_c(\mathbf{W}_t)$ . It is used in the proof of Lemma 36.

872 **Lemma 35.** Let  $c \in [k]$ . Under the same setting as Theorem 5, there exists a time  $t_0$  such that the  
 873 following holds for all  $t \geq t_0$

$$\left| \sqrt{\mathbf{V}_t[c, j]} - \sqrt{(1 - \beta_2^{t+1})} |\nabla\mathcal{L}(\mathbf{W}_t)[c, j]| \right| \leq \alpha_V \sqrt{\eta_t} (\mathcal{Q}_c(\mathbf{W}_t) + \mathcal{G}_c(\mathbf{W}_t)),$$

874 where  $j \in [d]$ , and  $\alpha_V$  is some constant that depends on  $B$  and  $\beta_2$ .

875 *Proof.* Consider any fixed  $c \in [k]$  and  $j \in [d]$ ,

$$\begin{aligned} |\mathbf{V}_t[c, j] - (1 - \beta_2^{t+1})\nabla\mathcal{L}(\mathbf{W}_t)[c, j]^2| &= \left| \sum_{\tau=0}^t (1 - \beta_2)\beta_2^\tau (\nabla\mathcal{L}(\mathbf{W}_{t-\tau})[c, j]^2 - \nabla\mathcal{L}(\mathbf{W}_t)[c, j]^2) \right| \\ &\leq \sum_{\tau=0}^t (1 - \beta_2)\beta_2^\tau \underbrace{|\nabla\mathcal{L}(\mathbf{W}_{t-\tau})[c, j]^2 - \nabla\mathcal{L}(\mathbf{W}_t)[c, j]^2|}_{\spadesuit}. \end{aligned} \quad (34)$$

876 For any  $\mathbf{W} \in \mathbb{R}^{k \times d}$ , recall that  $-\frac{1}{n} \sum_{i \in [n]} \mathbf{e}_c^T (\mathbf{e}_{y_i} - \mathbb{S}(\mathbf{W}\mathbf{h}_i)) h_{ij}$ . Then, we can obtain  
 877  $\nabla\mathcal{L}(\mathbf{W})[c, j]^2 = \frac{1}{n^2} \sum_{i \in [n]} \sum_{p \in [n]} h_{ij} h_{pj} (\delta_{cy_i} - \mathbb{S}_c(\mathbf{W}\mathbf{h}_i)) (\delta_{cy_p} - \mathbb{S}_c(\mathbf{W}\mathbf{h}_p))$  where  $\delta_{cy} = 1$   
 878 if and only if  $c = y$ . Next, we define the function  $f_{c,i,p}$  to be  $f_{c,i,p}(\mathbf{W}) := (\delta_{cy_i} - \mathbb{S}_c(\mathbf{W}\mathbf{h}_i)) (\delta_{cy_p} -$   
 879  $\mathbb{S}_c(\mathbf{W}\mathbf{h}_p))$ . Then, we have

$$\begin{aligned} |f_{c,i,p}(\mathbf{W}_{t-\tau}) - f_{c,i,p}(\mathbf{W}_t)| &= \delta_{cy_i} (\mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p) - \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_p)) + \delta_{cy_p} (\mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i) - \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i)) \\ &\quad + (\mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i) \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_p) - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i) \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p)) \end{aligned}$$

880 We can substitute this result into  $\spadesuit$  to obtain

$$\begin{aligned} \spadesuit &= \left| \frac{1}{n^2} \sum_{i \in [n]} \sum_{p \in [n]} h_{ij} h_{pj} (f_{c,i,p}(\mathbf{W}_{t-\tau}) - f_{c,i,p}(\mathbf{W}_t)) \right| \\ &\leq \frac{B^2}{n^2} \sum_{i \in [n]} \sum_{p \in [n]} |f_{c,i,p}(\mathbf{W}_{t-\tau}) - f_{c,i,p}(\mathbf{W}_t)| \\ &= B^2 \frac{1}{n^2} \underbrace{\sum_{i \in [n], y_i \neq c} \sum_{p \in [n], y_p \neq c} |f_{c,i,p}(\mathbf{W}_{t-\tau}) - f_{c,i,p}(\mathbf{W}_t)|}_{\spadesuit_1} \\ &\quad + B^2 \frac{1}{n^2} \underbrace{\sum_{i \in [n], y_i \neq c} \sum_{p \in [n], y_p = c} |f_{c,i,p}(\mathbf{W}_{t-\tau}) - f_{c,i,p}(\mathbf{W}_t)|}_{\spadesuit_2} \\ &\quad + B^2 \frac{1}{n^2} \underbrace{\sum_{i \in [n], y_i = c} \sum_{p \in [n], y_p \neq c} |f_{c,i,p}(\mathbf{W}_{t-\tau}) - f_{c,i,p}(\mathbf{W}_t)|}_{\spadesuit_3} \\ &\quad + B^2 \frac{1}{n^2} \underbrace{\sum_{i \in [n], y_i = c} \sum_{p \in [n], y_p = c} |f_{c,i,p}(\mathbf{W}_{t-\tau}) - f_{c,i,p}(\mathbf{W}_t)|}_{\spadesuit_4} \end{aligned}$$

881 We deal with the 4 terms  $\spadesuit_1, \spadesuit_2, \spadesuit_3$ , and  $\spadesuit_4$  separately. Starting with the first term, we have

$$\begin{aligned}
\spadesuit_1 &= \frac{1}{n^2} \sum_{i \in [n], y_i \neq c} \sum_{p \in [n], y_p \neq c} |\mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i) \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_p) - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i) \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p)| \\
&= \frac{1}{n^2} \sum_{i \in [n], y_i \neq c} \sum_{p \in [n], y_p \neq c} \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i) \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p) \left| \frac{\mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i) \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_p)}{\mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i) \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p)} - 1 \right| \\
&\stackrel{(a)}{\leq} \frac{1}{n^2} \sum_{i \in [n], y_i \neq c} \sum_{p \in [n], y_p \neq c} \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i) \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p) (e^{2(\|\mathbf{W}_{t-\tau} - \mathbf{W}_t\|_{\infty} \|\mathbf{h}_i\|_{\infty} + \|\mathbf{W}_{t-\tau} - \mathbf{W}_t\|_{\infty} \|\mathbf{h}_p\|_{\infty})} - 1) \\
&\stackrel{(b)}{\leq} \frac{1}{n^2} \sum_{i \in [n], y_i \neq c} \sum_{p \in [n], y_p \neq c} \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i) \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p) (e^{4B\|\mathbf{W}_{t-\tau} - \mathbf{W}_t\|_{\max}} - 1) \\
&\stackrel{(c)}{\leq} (e^{4B \sum_{s=1}^{\tau} \eta_{t-s}} \left\| \frac{\mathbf{M}_{t-s}}{\sqrt{\mathbf{V}_{t-s}}} \right\|_{\max} - 1) \frac{1}{n^2} \sum_{i \in [n], y_i \neq c} \sum_{p \in [n], y_p \neq c} \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i) \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p) \\
&\stackrel{(d)}{\leq} (e^{4B\alpha \sum_{s=1}^{\tau} \eta_{t-s}} - 1) \mathcal{Q}_c(\mathbf{W}_t)^2,
\end{aligned}$$

882 where (a) is by Lemma 15, (b) is by  $\|\mathbf{h}_i\|_1 \leq B$  for all  $i \in [n]$ , (c) is by (33c) and the triangle  
883 inequality, and (d) is by Lemma 33 and the definition of  $\mathcal{G}(\mathbf{W}_t)$ . For the second term, we have

$$\begin{aligned}
\spadesuit_2 &= \frac{1}{n^2} \sum_{i \in [n], y_i \neq c} \sum_{p \in [n], y_p = c} |(\mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i) - \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i)) \\
&\quad + (\mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i) \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_p) - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i) \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p))| \\
&= \frac{1}{n^2} \sum_{i \in [n], y_i \neq c} \sum_{p \in [n], y_p = c} |\mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i) (1 - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p)) - (1 - \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_p)) \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i)| \\
&= \frac{1}{n^2} \sum_{i \in [n], y_i \neq c} \sum_{p \in [n], y_p = c} \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i) (1 - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p)) \left| 1 - \frac{(1 - \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_p)) \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i)}{(1 - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p)) \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i)} \right| \\
&\leq (e^{4B\alpha \sum_{s=1}^{\tau} \eta_{t-s}} - 1) \mathcal{Q}_c(\mathbf{W}_t) \mathcal{G}_c(\mathbf{W}_t),
\end{aligned}$$

884 where the last inequality is by Lemma 15 and the same steps taken for  $\spadesuit_1$ . The third term can be  
885 derived similarly as the second term and we can obtain the same bound as follows:

$$\begin{aligned}
\spadesuit_3 &= \frac{1}{n^2} \sum_{i \in [n], y_i = c} \sum_{p \in [n], y_p \neq c} \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p) (1 - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i)) \left| 1 - \frac{(1 - \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_p)) \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i)}{(1 - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i)) \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p)} \right| \\
&\leq (e^{4B\alpha \sum_{s=1}^{\tau} \eta_{t-s}} - 1) \mathcal{Q}_c(\mathbf{W}_t) \mathcal{G}_c(\mathbf{W}_t).
\end{aligned}$$

886 For the fourth term, we obtain:

$$\begin{aligned}
\spadesuit_4 &= \frac{1}{n^2} \sum_{i \in [n], y_i = c} \sum_{p \in [n], y_p = c} |(1 - \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i)) (1 - \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_p)) - (1 - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i)) (1 - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p))| \\
&= \frac{1}{n^2} \sum_{i \in [n], y_i = c} \sum_{p \in [n], y_p = c} (1 - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i)) (1 - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p)) \left| \frac{(1 - \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_i)) (1 - \mathbb{S}_c(\mathbf{W}_{t-\tau} \mathbf{h}_p))}{(1 - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_i)) (1 - \mathbb{S}_c(\mathbf{W}_t \mathbf{h}_p))} - 1 \right| \\
&\leq (e^{4B\alpha \sum_{s=1}^{\tau} \eta_{t-s}} - 1) \mathcal{G}_c(\mathbf{W}_t)^2,
\end{aligned}$$

887 where the last inequality is by Lemma 15 and the same steps taken for  $\spadesuit_1$ . We combine the bounds  
888 for  $\spadesuit_1, \spadesuit_2, \spadesuit_3$ , and  $\spadesuit_4$  to obtain:  $\spadesuit \leq 4B^2 (e^{4B\alpha \sum_{s=1}^{\tau} \eta_{t-s}} - 1) (\mathcal{G}_c(\mathbf{W}_t) + \mathcal{Q}_c(\mathbf{W}_t))^2$ . Then, we  
889 substitute this into (34) to obtain:

$$\begin{aligned}
|\mathbf{V}_t[c, j] - (1 - \beta_2^{t+1}) \nabla \mathcal{L}(\mathbf{W}_t)[c, j]^2| &\leq B^2 (1 - \beta_2) (\mathcal{Q}_c(\mathbf{W}_t) + \mathcal{G}_c(\mathbf{W}_t))^2 \sum_{\tau=0}^t \beta_2^{\tau} (e^{4\alpha B \sum_{s=1}^{\tau} \eta_{t-s}} - 1) \\
&\leq B^2 (1 - \beta_2) c_2 \eta_t (\mathcal{Q}_c(\mathbf{W}_t) + \mathcal{G}_c(\mathbf{W}_t))^2,
\end{aligned}$$

890 where the last inequality is by the Assumption 3. The final result follows from the fact that  $|p - q|^2 \leq$   
891  $|p^2 - q^2|$  when both  $p$  and  $q$  are positive.  $\square$

892 The following Lemma bounds the term  $|\langle \nabla \mathcal{L}(\mathbf{W}_t), \frac{\mathbf{M}_t}{\sqrt{\mathbf{V}_t}} - \frac{\nabla \mathcal{L}(\mathbf{W}_t)}{|\nabla \mathcal{L}(\mathbf{W}_t)|} \rangle|$  using  $\mathcal{G}(\mathbf{W}_t)$ . It is used in  
 893 Lemma 37 to show the decrease in the risk. The proof is similar to that of Zhang et al. [60, Lemma  
 894 A.3], but here we need to carefully track the index  $c \in [k]$  using both  $\mathcal{G}_c(\mathbf{W})$  and  $\mathcal{Q}_c(\mathbf{W})$  to avoid  
 895  $k$  dependence. The final result crucially relies on the decomposition  $\mathcal{G}(\mathbf{W}_t) = \sum_{c \in [k]} \mathcal{T}_c(\mathbf{W}_t) =$   
 896  $\sum_{c \in [k]} \mathcal{Q}_c(\mathbf{W}_t)$ .

897 **Lemma 36.** *Under the same setting as Theorem 5, we have*

$$\underbrace{|\langle \nabla \mathcal{L}(\mathbf{W}_t), \frac{\mathbf{M}_t}{\sqrt{\mathbf{V}_t}} - \frac{\nabla \mathcal{L}(\mathbf{W}_t)}{|\nabla \mathcal{L}(\mathbf{W}_t)|} \rangle|}_{\clubsuit} \leq 4 \sqrt{\frac{\beta_1^{t+1}}{1 - \beta_2^{t+1}}} \|\nabla \mathcal{L}(\mathbf{W}_t)\|_{\text{sum}} + \frac{2d}{\sqrt{1 - \beta_2}} \left( \frac{6\alpha_V}{\sqrt{1 - \beta_2^{t+1}}} \sqrt{\eta_t} + 3\alpha_M \eta_t \right) \mathcal{G}(\mathbf{W}_t).$$

898 *Proof.* For simplicity, we drop the subscripts  $t$ . Denote  $\mathcal{T}_c(\mathbf{W}) := \mathcal{G}_c(\mathbf{W}) + \mathcal{Q}_c(\mathbf{W})$ . Then, by  
 899 Lemmas 34 and 35, we have for any  $c \in [k]$  and  $j \in [d]$ :

$$\mathbf{M}[c, j] = (1 - \beta_1^{t+1}) \nabla \mathcal{L}(\mathbf{W})[c, j] + \alpha_M \eta_t \mathcal{T}_c(\mathbf{W}) \epsilon_{m, c, j}, \quad (35)$$

$$\sqrt{\mathbf{V}[c, j]} = \sqrt{1 - \beta_2^{t+1}} |\nabla \mathcal{L}(\mathbf{W})[c, j]| + \alpha_V \sqrt{\eta_t} \mathcal{T}_c(\mathbf{W}) \epsilon_{v, c, j}, \quad (36)$$

900 where  $|\epsilon_{m, c, j}| \leq 1$  and  $|\epsilon_{v, c, j}| \leq 1$  are some residual terms. We denote  
 901  $\psi_{c, j} := \nabla \mathcal{L}(\mathbf{W})[c, j] \left( \frac{\mathbf{M}[c, j]}{\sqrt{\mathbf{V}[c, j]}} - \frac{\nabla \mathcal{L}(\mathbf{W})[c, j]}{|\nabla \mathcal{L}(\mathbf{W})[c, j]|} \right)$ , the set of index  $E_{c, j} := \{j \in [d] \mid$   
 902  $\sqrt{1 - \beta_2^{t+1}} |\nabla \mathcal{L}(\mathbf{W})[c, j]| \geq 2\alpha_V \sqrt{\eta_t} \mathcal{T}_c(\mathbf{W}) |\epsilon_{v, c, j}|\}$ , and its complement  $E_{c, j}^c = [d] \setminus E_{c, j}$ . The  
 903 goal is to bound  $|\psi_{c, j}|$  when  $j \in E_{c, j}^c$  or  $j \in E_{c, j}$  using  $\mathcal{T}_c(\mathbf{W})$ . We start with the indices in  $E_{c, j}^c$ :

$$\begin{aligned} \sum_{j \in E_{c, j}^c} |\psi_{c, j}| &\leq \sum_{j \in E_{c, j}^c} |\nabla \mathcal{L}(\mathbf{W})[c, j]| \left( \frac{|\mathbf{M}[c, j]|}{\sqrt{\mathbf{V}[c, j]}} + 1 \right) \\ &\stackrel{(a)}{\leq} \sum_{j \in E_{c, j}^c} |\nabla \mathcal{L}(\mathbf{W})[c, j]| \left( \frac{(1 - \beta_1^{t+1}) |\nabla \mathcal{L}(\mathbf{W})[c, j]| + \alpha_M \eta_t \mathcal{T}_c(\mathbf{W})}{\sqrt{1 - \beta_2} |\nabla \mathcal{L}(\mathbf{W})[c, j]|} + 1 \right) \\ &\stackrel{(b)}{\leq} \sum_{j \in E_{c, j}^c} \left( \frac{1 - \beta_1^{t+1}}{\sqrt{1 - \beta_2}} + 1 \right) \frac{2\alpha_V \sqrt{\eta_t} \mathcal{T}_c(\mathbf{W})}{\sqrt{1 - \beta_2^{t+1}}} + \frac{\alpha_M \eta_t \mathcal{T}_c(\mathbf{W})}{\sqrt{1 - \beta_2}} \\ &\leq \frac{d}{\sqrt{1 - \beta_2}} \left( \frac{4\alpha_V}{\sqrt{1 - \beta_2^{t+1}}} \sqrt{\eta_t} + \alpha_M \eta_t \right) \mathcal{T}_c(\mathbf{W}), \end{aligned}$$

904 where (a) is by (35),  $|\epsilon_{m, c, j}| \leq 1$ , and  $\mathbf{V}[c, j] \geq (1 - \beta_2) |\nabla \mathcal{L}(\mathbf{W})[c, j]|^2$ ; and (b) is by  $j \in E_{c, j}^c$  s.t.  
 905  $|\nabla \mathcal{L}(\mathbf{W})[c, j]| \leq \frac{2\alpha_V \sqrt{\eta_t} \mathcal{T}_c(\mathbf{W})}{\sqrt{1 - \beta_2^{t+1}}}$ . Next, we focus on the indices  $j \in E_{c, j}$ . In this case, we have

$$\begin{aligned} \psi_{c, j} &= \nabla \mathcal{L}(\mathbf{W})[c, j] \left( \frac{\mathbf{M}[c, j]}{\sqrt{1 - \beta_2^{t+1}} |\nabla \mathcal{L}(\mathbf{W})[c, j]| + \alpha_V \sqrt{\eta_t} \mathcal{T}_c(\mathbf{W}) \epsilon_{v, c, j}} - \frac{\nabla \mathcal{L}(\mathbf{W})[c, j]}{|\nabla \mathcal{L}(\mathbf{W})[c, j]|} \right) \\ &= \nabla \mathcal{L}(\mathbf{W})[c, j] \frac{\mathbf{M}[c, j] |\nabla \mathcal{L}(\mathbf{W})[c, j]| - (\sqrt{1 - \beta_2^{t+1}} |\nabla \mathcal{L}(\mathbf{W})[c, j]| + \alpha_V \sqrt{\eta_t} \mathcal{T}_c(\mathbf{W}) \epsilon_{v, c, j}) |\nabla \mathcal{L}(\mathbf{W})[c, j]|}{\underbrace{(\sqrt{1 - \beta_2^{t+1}} |\nabla \mathcal{L}(\mathbf{W})[c, j]| + \alpha_V \sqrt{\eta_t} \mathcal{T}_c(\mathbf{W}) \epsilon_{v, c, j}) |\nabla \mathcal{L}(\mathbf{W})[c, j]|}_{\clubsuit_1}}, \end{aligned}$$

906 where

$$\begin{aligned} \left| \clubsuit_1 \right| &= \left| (1 - \beta_1^{t+1} - \sqrt{1 - \beta_2^{t+1}}) \nabla \mathcal{L}(\mathbf{W})[c, j] |\nabla \mathcal{L}(\mathbf{W})[c, j]| + \right. \\ &\quad \left. \alpha_M \eta_t \mathcal{T}_c(\mathbf{W}) \epsilon_{m, c, j} |\nabla \mathcal{L}(\mathbf{W})[c, j]| - \alpha_V \sqrt{\eta_t} \mathcal{T}_c(\mathbf{W}) \epsilon_{v, c, j} \nabla \mathcal{L}(\mathbf{W})[c, j] \right| \\ &\stackrel{(c)}{\leq} \left| 1 - \beta_1^{t+1} - \sqrt{1 - \beta_2^{t+1}} \right| |\nabla \mathcal{L}(\mathbf{W})[c, j]|^3 + (\alpha_M \eta_t + \alpha_V \sqrt{\eta_t}) \mathcal{T}_c(\mathbf{W}) |\nabla \mathcal{L}(\mathbf{W})[c, j]|^2, \end{aligned}$$

907 and

$$\left| \spadesuit_2 \right| = \spadesuit_2 \stackrel{(d)}{\geq} \frac{1}{2} \sqrt{1 - \beta_2^{t+1}} |\nabla \mathcal{L}(\mathbf{W})[c, j]|^2.$$

908 Inequality (c) is by  $|\epsilon_{m,c,j}| \leq 1$  and  $|\epsilon_{v,c,j}| \leq 1$ , and (d) is by  $\alpha_V \sqrt{\eta_t} \mathcal{T}_c(\mathbf{W}) \epsilon_{v,c,j} \geq$   
 909  $-\frac{1}{2} \sqrt{1 - \beta_2^{t+1}} |\nabla \mathcal{L}(\mathbf{W})[c, j]|$  for any  $j \in E_{c,j}$ . Putting these two pieces together, we obtain

$$\begin{aligned} \sum_{j \in E_{c,j}} |\psi_{c,j}| &\leq \sum_{j \in E_{c,j}} \frac{|1 - \beta_1^{t+1} - \sqrt{1 - \beta_2^{t+1}}| |\nabla \mathcal{L}(\mathbf{W})[c, j]|^3 + (\alpha_M \eta_t + \alpha_V \sqrt{\eta_t}) \mathcal{T}_c(\mathbf{W}) |\nabla \mathcal{L}(\mathbf{W})[c, j]|^2}{\frac{1}{2} \sqrt{1 - \beta_2^{t+1}} |\nabla \mathcal{L}(\mathbf{W})[c, j]|^2} \\ &\stackrel{(e)}{\leq} \left( \sum_{j \in E_{c,j}} 4 \sqrt{\frac{\beta_1^{t+1}}{1 - \beta_2^{t+1}}} |\nabla \mathcal{L}(\mathbf{W})[c, j]| \right) + d \left( \frac{2\alpha_V}{\sqrt{1 - \beta_2^{t+1}}} \sqrt{\eta_t} + \frac{2\alpha_M}{\sqrt{1 - \beta_2^{t+1}}} \eta_t \right) \mathcal{T}_c(\mathbf{W}) \\ &\leq 4 \sqrt{\frac{\beta_1^{t+1}}{1 - \beta_2^{t+1}}} \|\nabla \mathcal{L}(\mathbf{W})[c, :]\|_{\text{sum}} + \frac{d}{\sqrt{1 - \beta_2}} \left( \frac{2\alpha_V}{\sqrt{1 - \beta_2^{t+1}}} \sqrt{\eta_t} + 2\alpha_M \eta_t \right) \mathcal{T}_c(\mathbf{W}), \end{aligned}$$

910 where (e) is by  $\sqrt{a} \leq \sqrt{a-b} + \sqrt{b}$  implying  $1 - \sqrt{1 - \beta_2^{t+1}} \leq \beta_1^{\frac{t+1}{2}}$ , and  $\nabla \mathcal{L}(\mathbf{W})[c, :]$  denotes  
 911 the  $c$ th row of  $\nabla \mathcal{L}(\mathbf{W})$ . Finally, we note that  $|\langle \nabla \mathcal{L}(\mathbf{W}), \frac{\mathbf{M}}{\sqrt{\mathbf{V}}} - \frac{\nabla \mathcal{L}(\mathbf{W})}{|\nabla \mathcal{L}(\mathbf{W})|} \rangle| = |\sum_{(c,j)} \psi_{c,j}| \leq$   
 912  $\sum_{(c,j)} |\psi_{c,j}|$ . Then, we obtain

$$\begin{aligned} \sum_{c,j} |\psi_{c,j}| &= \sum_{c \in [k]} \left( \sum_{j \in E_{c,j}^c} |\psi_{c,j}| + \sum_{j \in E_{c,j}} |\psi_{c,j}| \right) \\ &= \sum_{c \in [k]} 4 \sqrt{\frac{\beta_1^{t+1}}{1 - \beta_2^{t+1}}} \|\nabla \mathcal{L}(\mathbf{W})[c, :]\|_{\text{sum}} + \sum_{c \in [k]} \frac{d}{\sqrt{1 - \beta_2}} \left( \frac{2\alpha_V}{\sqrt{1 - \beta_2^{t+1}}} \sqrt{\eta_t} + 2\alpha_M \eta_t \right) \mathcal{T}_c(\mathbf{W}) \\ &\stackrel{(f)}{=} 4 \sqrt{\frac{\beta_1^{t+1}}{1 - \beta_2^{t+1}}} \|\nabla \mathcal{L}(\mathbf{W})\|_{\text{sum}} + \frac{2d}{\sqrt{1 - \beta_2}} \left( \frac{2\alpha_V}{\sqrt{1 - \beta_2^{t+1}}} \sqrt{\eta_t} + 2\alpha_M \eta_t \right) \mathcal{G}(\mathbf{W}), \end{aligned}$$

913 where (f) is by  $\sum_{c \in [k]} \mathcal{T}_c(\mathbf{W}) = \sum_{c \in [k]} \mathcal{Q}_c(\mathbf{W}) + \mathcal{G}_c(\mathbf{W}) = 2\mathcal{G}(\mathbf{W})$ . □

914 **Lemma 37** (Adam Descent). *Under the same setting as Theorem 5, set  $t_A := \frac{2 \log(\frac{\sqrt{1-\beta_2}}{4})}{\log(\beta_1)}$ , then we*  
 915 *have for all  $t \geq t_A$*

$$\mathcal{L}(\mathbf{W}_{t+1}) \leq \mathcal{L}(\mathbf{W}_t) - \eta_t \gamma (1 - \alpha_{a_1} \beta_1^{t/2} - \alpha_{a_2} d \eta_t^{\frac{1}{2}} - \alpha_{a_3} d \eta_t) \mathcal{G}(\mathbf{W}_t),$$

916 where  $\alpha_{a_1}$ ,  $\alpha_{a_2}$ , and  $\alpha_{a_3}$  are some constants that depend on  $B$ ,  $\gamma$ ,  $\beta_1$ , and  $\beta_2$ .

917 *Proof.* We follow the same notations and strategy of Lemma 21,  
 918 and recall the definitions  $\spadesuit_t = \langle \nabla \mathcal{L}(\mathbf{W}_t), \Delta_t \rangle$  and  $\clubsuit_t =$   
 919  $\mathbf{h}_i^\top \Delta_t^\top (\text{diag}(\mathbb{S}(\mathbf{W}_{t,t+1}, \gamma \mathbf{h}_i)) - \mathbb{S}(\mathbf{W}_{t,t+1}, \zeta^* \mathbf{h}_i) \mathbb{S}(\mathbf{W}_{t,t+1}, \zeta^* \mathbf{h}_i)^\top) \Delta_t \mathbf{h}_i$ . In the case of  
 920 Adam, we have  $\Delta_t = \frac{\mathbf{M}_t}{\sqrt{\mathbf{V}_t}}$ . We bound  $\spadesuit_t$  and  $\clubsuit_t$  separately. Starting with  $\spadesuit_t$ , we have for all



921  $t \geq t_A$

$$\begin{aligned}
\spadesuit_t &= -\eta_t \langle \nabla \mathcal{L}(\mathbf{W}_t), \frac{\mathbf{M}_t}{\sqrt{\mathbf{V}_t}} \rangle \\
&= -\eta_t \left( \langle \nabla \mathcal{L}(\mathbf{W}_t), \frac{\mathbf{M}_t}{\sqrt{\mathbf{V}_t}} - \frac{\nabla \mathcal{L}(\mathbf{W}_t)}{|\nabla \mathcal{L}(\mathbf{W}_t)|} \rangle + \langle \nabla \mathcal{L}(\mathbf{W}_t), \frac{\nabla \mathcal{L}(\mathbf{W}_t)}{|\nabla \mathcal{L}(\mathbf{W}_t)|} \rangle \right) \\
&\leq -\eta_t \|\nabla \mathcal{L}(\mathbf{W}_t)\|_{\text{sum}} + \eta_t \left| \langle \nabla \mathcal{L}(\mathbf{W}_t), \frac{\mathbf{M}_t}{\sqrt{\mathbf{V}_t}} - \frac{\nabla \mathcal{L}(\mathbf{W}_t)}{|\nabla \mathcal{L}(\mathbf{W}_t)|} \rangle \right| \\
&\stackrel{(a)}{\leq} -\eta_t \left( 1 - 4\sqrt{\frac{\beta_1^{t+1}}{1 - \beta_2^{t+1}}} \right) \|\nabla \mathcal{L}(\mathbf{W}_t)\|_{\text{sum}} + \frac{2d}{\sqrt{1 - \beta_2}} \left( \frac{6\alpha_V}{\sqrt{1 - \beta_2^{t+1}}} \eta_t^{\frac{3}{2}} + 3\alpha_M \eta_t^2 \right) \mathcal{G}(\mathbf{W}_t) \\
&\leq -\eta_t \left( 1 - 4\frac{\beta_1^{\frac{t}{2}}}{\sqrt{1 - \beta_2}} \right) \|\nabla \mathcal{L}(\mathbf{W}_t)\|_{\text{sum}} + \frac{12\alpha_V}{1 - \beta_2} d \eta_t^{3/2} \mathcal{G}(\mathbf{W}_t) + \frac{6\alpha_M}{\sqrt{1 - \beta_2}} d \eta_t^2 \mathcal{G}(\mathbf{W}_t) \\
&\stackrel{(b)}{\leq} -\eta_t \gamma \left( 1 - 4\frac{\beta_1^{\frac{t}{2}}}{\sqrt{1 - \beta_2}} \right) \mathcal{G}(\mathbf{W}_t) + \frac{12\alpha_V}{1 - \beta_2} d \eta_t^{3/2} \mathcal{G}(\mathbf{W}_t) + \frac{6\alpha_M}{\sqrt{1 - \beta_2}} d \eta_t^2 \mathcal{G}(\mathbf{W}_t),
\end{aligned}$$

922 where (a) is by Lemma 36, and (b) is by Lemma 16. For  $\clubsuit_t$ , we apply Lemma 12 to obtain

$$\clubsuit_t \leq 4\|\Delta_t \mathbf{h}_i\|_\infty^2 (1 - \mathbb{S}_{y_i}(\mathbf{W}_{t,t+1,\zeta^*} \mathbf{h}_i)) \leq 4\eta_t^2 \alpha^2 B^2 (1 - \mathbb{S}_{y_i}(\mathbf{W}_{t,t+1,\zeta^*} \mathbf{h}_i)),$$

923 where in the second inequality we have used  $\|\Delta_t \mathbf{h}_i\|_\infty \leq \|\Delta_t\|_{\max} \|\mathbf{h}_i\|_1$ ,  $\|\mathbf{h}_i\|_1 \leq B$ , and

924  $\|\Delta_t\|_{\max} = \eta_t \left\| \frac{\mathbf{M}_t}{\sqrt{\mathbf{V}_t}} \right\|_{\max} \leq \eta_t \alpha$  by Lemma 33 given  $t \geq t_A$  implying that  $1 \geq 4\frac{\beta_1^{\frac{t}{2}}}{\sqrt{1 - \beta_2}}$ . Combing  
925 this with Lemma 12, we obtain

$$\begin{aligned}
&\frac{1}{2n} \sum_{i \in [n]} \mathbf{h}_i^\top \Delta_t^\top (\text{diag}(\mathbb{S}(\mathbf{W}_{t,t+1,\gamma} \mathbf{h}_i)) - \mathbb{S}(\mathbf{W}_{t,t+1,\zeta^*} \mathbf{h}_i) \mathbb{S}(\mathbf{W}_{t,t+1,\zeta^*} \mathbf{h}_i)^\top) \Delta_t \mathbf{h}_i \\
&\leq \frac{1}{2n} \sum_{i \in [n]} 4\eta_t^2 \alpha^2 B^2 (1 - \mathbb{S}_{y_i}(\mathbf{W}_{t,t+1,\zeta^*} \mathbf{h}_i)) \leq 2\alpha^2 \eta_t^2 B^2 e^{2B\eta_0} \mathcal{G}(\mathbf{W}_t),
\end{aligned}$$

926 where the derivation of the second inequality can be found in the derivation of 22. Putting everything  
927 together, we obtain

$$\begin{aligned}
\mathcal{L}(\mathbf{W}_{t+1}) &\leq \mathcal{L}(\mathbf{W}_t) - \eta_t \gamma \mathcal{G}(\mathbf{W}_t) + 4\frac{\beta_1^{\frac{t}{2}}}{\sqrt{1 - \beta_2}} \gamma \eta_t \mathcal{G}(\mathbf{W}_t) + \frac{12\alpha_V}{1 - \beta_2} d \eta_t^{3/2} \mathcal{G}(\mathbf{W}_t) + \\
&\quad \left( \frac{6\alpha_M}{\sqrt{1 - \beta_2}} + 2\alpha^2 B^2 e^{2B\eta_0} \right) d \eta_t^2 \mathcal{G}(\mathbf{W}_t) \\
&= \mathcal{L}(\mathbf{W}_t) - \eta_t \gamma \left( 1 - \alpha_{a_1} \beta_1^{t/2} - \alpha_{a_2} d \eta_t^{\frac{1}{2}} - \alpha_{a_3} d \eta_t \right) \mathcal{G}(\mathbf{W}_t),
\end{aligned}$$

928 where we have defined  $\alpha_{a_1} := \frac{4}{\sqrt{1 - \beta_2}}$ ,  $\alpha_{a_2} := \frac{12\alpha_V}{\gamma(1 - \beta_2)}$ , and  $\alpha_{a_3} := \frac{6\alpha_M}{\gamma\sqrt{1 - \beta_2}} + \frac{2\alpha^2 B^2 e^{2B\eta_0}}{\gamma}$ .  $\square$

929 Built upon Lemma 37, we can further lower bound the unnormalized margin of Adam iterates for a  
930 sufficiently large  $t$ . The proof is similar to that of NSD (i.e., Lemma 22), which crucially depends  
931 on the separability condition obtained after achieving a low loss (Lemma 19). The time  $\tilde{t}_A$  will be  
932 specified in the proof of Theorem 5.

933 **Lemma 38** (Adam Unnormalized Margin). *Under the same setting as Theorem 5, suppose that there*  
934 *exist  $\tilde{t}$  such that  $\mathcal{L}(\mathbf{W}_t) \leq \frac{\log 2}{n}$  for all  $t > \tilde{t}$ , then we have for all  $t \geq \tilde{t}_A := \max\{\tilde{t}_A, \tilde{t}\}$*

$$\min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^\top \mathbf{W}_t \mathbf{h}_i \geq \gamma \sum_{s=\tilde{t}_A}^{t-1} \eta_s \frac{\mathcal{G}(\mathbf{W}_s)}{\mathcal{L}(\mathbf{W}_s)} - \alpha_{a_5} d \sum_{s=\tilde{t}_A}^{t-1} \eta_s^{\frac{3}{2}} - \alpha_{a_6} d \sum_{s=\tilde{t}_A}^{t-1} \eta_s^2 - \alpha_{a_7},$$

935 where  $\tilde{t}_A = \frac{2 \log(\frac{\sqrt{1 - \beta_2}}{4})}{\log(\beta_1)}$ , and  $\alpha_{a_5}$ ,  $\alpha_{a_6}$ , and  $\alpha_{a_7}$  are some constants that depend on  $B$ ,  $\beta_1$ , and  $\beta_2$ .

936 *Proof.* We denote  $\alpha_{a_4} := \frac{4}{\sqrt{1-\beta_2}}$ ,  $\alpha_{a_5} := \frac{12\alpha_V}{1-\beta_2}$ , and  $\alpha_{a_6} := \frac{6\alpha_M}{\sqrt{1-\beta_2}} + 2\alpha^2 B^2 e^{2B\eta_0}$ . Under the  
 937 assumption that  $\mathcal{L}(\mathbf{W}_t) \leq \frac{\log 2}{n}$  for all  $t \geq \tilde{t}$ , we have for all  $t \geq \tilde{t}_A := \max\{t_A, \tilde{t}\}$  (recall that  
 938  $t_A = \frac{2 \log(\frac{\sqrt{1-\beta_2}}{4})}{\log(\beta_1)}$ )

$$\begin{aligned} \mathcal{L}(\mathbf{W}_{t+1}) &\stackrel{(a)}{\leq} \mathcal{L}(\mathbf{W}_t) - \eta_t \gamma \mathcal{G}(\mathbf{W}_t) + \alpha_{a_4} \beta_1^{\frac{t}{2}} \gamma \eta_t \mathcal{G}(\mathbf{W}_t) + \alpha_{a_5} d \eta_t^{\frac{3}{2}} \mathcal{G}(\mathbf{W}_t) + \alpha_{a_6} d \eta_t^2 \mathcal{G}(\mathbf{W}_t) \\ &\stackrel{(b)}{\leq} \mathcal{L}(\mathbf{W}_t) \left(1 - \eta_t \gamma \frac{\mathcal{G}(\mathbf{W}_t)}{\mathcal{L}(\mathbf{W}_t)} + \alpha_{a_4} \beta_1^{\frac{t}{2}} \gamma \eta_t + \alpha_{a_5} d \eta_t^{\frac{3}{2}} + \alpha_{a_6} d \eta_t^2\right) \\ &\leq \mathcal{L}(\mathbf{W}_{\tilde{t}_A}) \exp\left(-\gamma \sum_{s=\tilde{t}_A}^t \eta_s \frac{\mathcal{G}(\mathbf{W}_s)}{\mathcal{L}(\mathbf{W}_s)} + \alpha_{a_4} \gamma \sum_{s=\tilde{t}_A}^t \beta_1^{\frac{s}{2}} \eta_s + \alpha_{a_5} d \sum_{s=\tilde{t}_A}^t \eta_s^{\frac{3}{2}} + \alpha_{a_6} d \sum_{s=\tilde{t}_A}^t \eta_s^2\right) \\ &\stackrel{(c)}{\leq} \frac{\log 2}{n} \exp\left(-\gamma \sum_{s=\tilde{t}_A}^t \eta_s \frac{\mathcal{G}(\mathbf{W}_s)}{\mathcal{L}(\mathbf{W}_s)} + \alpha_{a_5} d \sum_{s=\tilde{t}_A}^t \eta_s^{\frac{3}{2}} + \alpha_{a_6} d \sum_{s=\tilde{t}_A}^t \eta_s^2 + \alpha_{a_7}\right), \end{aligned}$$

939 where (a) is by Lemma 37, (b) is by  $\frac{\mathcal{G}(\mathbf{W}_t)}{\mathcal{L}(\mathbf{W}_t)} \leq 1$  (shown in Lemma 18), and (c) is by  $\mathcal{L}(\mathbf{W}_{\tilde{t}_A}) \leq \frac{\log 2}{n}$   
 940 and  $\alpha_{a_4} \gamma \sum_{s=\tilde{t}_A}^t \beta_1^{\frac{s}{2}} \eta_s \leq \frac{\alpha_{a_4} \gamma \eta_0}{1-\beta_1^{\frac{1}{2}}} =: \alpha_{a_7}$ . The rest of the proof follows the same arguments in  
 941 Lemma 22. Namely, the assumption  $\mathcal{L}(\mathbf{W}_t) \leq \frac{\log 2}{n}$  implies that  $\min_{c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i \geq 0$  for  
 942 all  $i \in [n]$ . This separability condition can be used further to show that for all  $t \geq \tilde{t}_A$

$$e^{-\min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i} \leq \exp\left(-\gamma \sum_{s=\tilde{t}_A}^{t-1} \eta_s \frac{\mathcal{G}(\mathbf{W}_s)}{\mathcal{L}(\mathbf{W}_s)} + \alpha_{a_5} d \sum_{s=\tilde{t}_A}^{t-1} \eta_s^{\frac{3}{2}} + \alpha_{a_6} d \sum_{s=\tilde{t}_A}^{t-1} \eta_s^2 + \alpha_{a_7}\right).$$

943 Taking the log on both sides leads to the final result.  $\square$

944 Next lemma upper bounds the max-norm of Adam iterates. It involves showing that the risk upper  
 945 bounds entry-wise second moment, which will become small after the risk starts to monotonically  
 946 decrease. Its proof can be found in Zhang et al. [60, Lemma 6.4]. Here, we only show the steps that  
 947 are specific in our settings.

948 **Lemma 39** (Adam  $\|\mathbf{W}_t\|_{\max}$ ). *Under the same setting as Theorem 5, suppose that there exists*  
 949  $\tilde{t}_B > \log(\frac{1}{\omega})$  *such that  $\mathcal{L}(\mathbf{W}_t) \leq \frac{1}{\sqrt{4B^2 + \alpha_V \eta_0}}$  for all  $t \geq \tilde{t}_B$ , then we have*

$$\|\mathbf{W}_t\|_{\max} \leq \alpha_{a_8} \sum_{s=0}^{\tilde{t}_B-1} \eta_s + \sum_{s=\tilde{t}_B}^{t-1} \eta_s + \|\mathbf{W}_0\|_{\max},$$

950 where  $\alpha_{a_8}$  is some constant that depends on  $B$ ,  $\beta_1$ , and  $\beta_2$ .

951 *Proof.* For any  $c \in [k]$  and  $j \in [d]$ , we have for all  $t \geq \tilde{t}_B$

$$\begin{aligned} \mathbf{V}_t[c, j] &\stackrel{(a)}{\leq} (1 - \beta_2^{t+1}) \nabla \mathcal{L}(\mathbf{W}_t)[c, j]^2 + \alpha_V \eta_t \mathcal{G}(\mathbf{W}_t)^2 \\ &\leq \nabla \mathcal{L}(\mathbf{W}_t)[c, j]^2 + \alpha_V \eta_t \mathcal{G}(\mathbf{W}_t)^2 \\ &\stackrel{(b)}{\leq} 4B^2 \mathcal{G}(\mathbf{W}_t)^2 + \alpha_V \eta_0 \mathcal{G}(\mathbf{W}_t)^2 \\ &\stackrel{(c)}{\leq} (4B^2 + \alpha_V \eta_0) \mathcal{L}(\mathbf{W}_t)^2 \stackrel{(d)}{\leq} 1, \end{aligned}$$

952 where (a) is by Lemma 34, (b) is by Lemma 16, (c) is by Lemma 18, and (d) is by the assumption.  
 953 This implies that for all  $t \geq \tilde{t}_B$

$$0 \geq \log(\mathbf{V}_t[c, j]) \geq \log(\beta_2^t (1 - \beta_2) \mathcal{L}(\mathbf{W}_t)[c, j]^2) \stackrel{(e)}{\geq} t \log(\beta_2) + \log(1 - \beta_2) + \log(\omega),$$

954 where (e) is by the Assumption 5. The rest proof follows the same arguments in Zhang et al. [60,  
 955 Lemma 6.4].  $\square$

**Theorem 5.** Suppose that Assumption 1, 2, 3, 4, and 5 hold, and  $\beta_1 \leq \beta_2$ , then there exists  $t_{a_2} = t_{a_2}(n, d, \gamma, B, \mathbf{W}_0, \beta_1, \beta_2, \omega)$  such that Adam achieves the following for all  $t > t_{a_2}$

$$\left| \frac{\min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i}{\|\mathbf{W}_t\|_{\max}} - \gamma \right| \leq \mathcal{O}\left(\frac{\sum_{s=t_{a_2}}^{t-1} \eta_s e^{-\frac{\gamma}{4} \sum_{\tau=t_{a_2}}^{s-1} \eta_\tau} + \sum_{s=0}^{t_{a_2}-1} \eta_s + d \sum_{s=t_{a_2}}^{t-1} \eta_s^{3/2}}{\sum_{s=0}^{t-1} \eta_s}\right).$$

*Proof. Determination of  $t_{a_1}$ .* Here, we consider learning rate schedule of the form  $\eta_t = \Theta(\frac{1}{t^a})$  where  $a \in (0, 1]$ . We choose  $t_{a_1}$  after  $(\max\{t_0, t_A, \log(\frac{1}{\omega})\})$  where  $t_0$  satisfies Assumption 3 and  $t_A = \frac{2 \log(\frac{\sqrt{1-\beta_2}}{4})}{\log \beta_1}$  such that the following conditions are met:  $\alpha_{a_1} \beta_1^{t/2} \leq \frac{1}{6}$ ,  $\alpha_{a_2} d \eta_t^{1/2} \leq \frac{1}{6}$ , and  $\alpha_{a_3} d \eta_t \leq \frac{1}{6}$ . Concretely, we can set  $t_{a_1} = \max\{\frac{-2 \log(6\alpha_{a_1})}{\log \beta_1}, (36\alpha_{a_2}^2 d^2)^{1/a}, (6\alpha_{a_3} d)^{1/a}\} = \Theta(d^{2/a})$ . Then, we have for all  $t \geq t_{a_1}$

$$\mathcal{L}(\mathbf{W}_{t+1}) \leq \mathcal{L}(\mathbf{W}_t) - \frac{\eta_t \gamma}{2} \mathcal{G}(\mathbf{W}_t). \quad (37)$$

Rearranging this equation and using non-negativity of the loss we obtain  $\gamma \sum_{s=t_{a_1}}^t \eta_s \mathcal{G}(\mathbf{W}_s) \leq 2\mathcal{L}(\mathbf{W}_{t_{a_1}})$ .

**Determination of  $t_{a_2}$ .** By Lemma 17, we can bound  $\mathcal{L}(\mathbf{W}_{t_{a_1}})$  as follows

$$|\mathcal{L}(\mathbf{W}_{t_{a_1}}) - \mathcal{L}(\mathbf{W}_0)| \leq 2B \|\mathbf{W}_{t_{a_1}} - \mathbf{W}_0\|_{\max} \leq 2B \sum_{s=0}^{t_{a_1}-1} \eta_s \left\| \frac{\mathbf{M}_s}{\sqrt{\mathbf{V}_s}} \right\|_{\max} \leq 2B\alpha \sum_{s=0}^{t_{a_1}-1} \eta_s,$$

where the last inequality is by Lemma 33. Combining this with the result above and letting  $\tilde{\mathcal{L}} = \min\{\frac{\log 2}{n}, \frac{1}{\sqrt{4B^2 + \alpha_V \eta_0}}\}$ , we obtain

$$\mathcal{G}(\mathbf{W}_{t^*}) = \min_{s \in [t_{a_1}, t_{a_2}]} \mathcal{G}(\mathbf{W}_s) \leq \frac{2\mathcal{L}(\mathbf{W}_0) + 4B\alpha \sum_{s=1}^{t_{a_1}-1} \eta_s}{\gamma \sum_{s=t_{a_1}}^{t_{a_2}} \eta_s} \leq \frac{\tilde{\mathcal{L}}}{2} \leq \frac{1}{2n},$$

from which we derive the sufficient condition on  $t_{a_2}$  to be  $\sum_{s=t_{a_1}}^{t_{a_2}} \eta_s \geq \frac{4\mathcal{L}(\mathbf{W}_0) + 8B\alpha \sum_{s=1}^{t_{a_1}-1} \eta_s}{\gamma \tilde{\mathcal{L}}}$ .

**Convergence of  $\frac{\mathcal{G}(\mathbf{W}_t)}{\mathcal{L}(\mathbf{W}_t)}$**  We follow the same arguments in the proof of NSD (Theorem 3) to conclude that

$$\frac{\mathcal{G}(\mathbf{W}_t)}{\mathcal{L}(\mathbf{W}_t)} \geq 1 - e^{-\frac{\gamma}{4} \sum_{s=t_{a_2}}^{t-1} \eta_s}. \quad (38)$$

We note that  $t_{a_2}$  satisfies the assumptions in Lemma 38 and Lemma 39.

**Margin Convergence** Finally, we combine Lemma 38, Lemma 39, and (38) to obtain

$$\begin{aligned} \left| \frac{\min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i}{\|\mathbf{W}_t\|_{\max}} - \gamma \right| &\leq \mathcal{O}\left(\frac{\sum_{s=t_{a_2}}^{t-1} \eta_s e^{-\frac{\gamma}{4} \sum_{\tau=t_{a_2}}^{s-1} \eta_\tau} + \sum_{s=0}^{t_{a_2}-1} \eta_s + d \sum_{s=t_{a_2}}^{t-1} \eta_s^{3/2} + d \sum_{s=t_{a_2}}^{t-1} \eta_s^2}{\sum_{s=0}^{t-1} \eta_s}\right) \\ &\leq \mathcal{O}\left(\frac{\sum_{s=t_{a_2}}^{t-1} \eta_s e^{-\frac{\gamma}{4} \sum_{\tau=t_{a_2}}^{s-1} \eta_\tau} + \sum_{s=0}^{t_{a_2}-1} \eta_s + d \sum_{s=t_{a_2}}^{t-1} \eta_s^{3/2}}{\sum_{s=0}^{t-1} \eta_s}\right) \end{aligned}$$

973

□

Similar to the case of NSD, we can derive the margin convergence rates for Adam.

**Corollary 2.** Consider learning rate schedule of the form  $\eta_t = \Theta(\frac{1}{t^a})$  where  $a \in (0, 1]$ , under the same setting as Theorem 5, then we have for Adam

$$\left| \frac{\min_{i \in [n], c \neq y_i} (\mathbf{e}_{y_i} - \mathbf{e}_c)^T \mathbf{W}_t \mathbf{h}_i}{\|\mathbf{W}_t\|_{\max}} - \gamma \right| = \begin{cases} \mathcal{O}\left(\frac{dt^{1-\frac{3a}{2}} + nd^{\frac{2(1-a)}{a}} + n\mathcal{L}(\mathbf{W}_0) + [\log(1/\omega)]^{1-a}}{t^{1-a}}\right) & \text{if } a < \frac{2}{3} \\ \mathcal{O}\left(\frac{d \log(t) + nd + n\mathcal{L}(\mathbf{W}_0) + [\log(1/\omega)]^{1/3}}{t^{1/3}}\right) & \text{if } a = \frac{2}{3} \\ \mathcal{O}\left(\frac{d + nd^{\frac{2(1-a)}{a}} + n\mathcal{L}(\mathbf{W}_0) + [\log(1/\omega)]^{1-a}}{t^{1-a}}\right) & \text{if } \frac{2}{3} < a < 1 \\ \mathcal{O}\left(\frac{d + n \log(d) + n\mathcal{L}(\mathbf{W}_0) + \log \log(1/\omega)}{\log t}\right) & \text{if } a = 1 \end{cases}$$

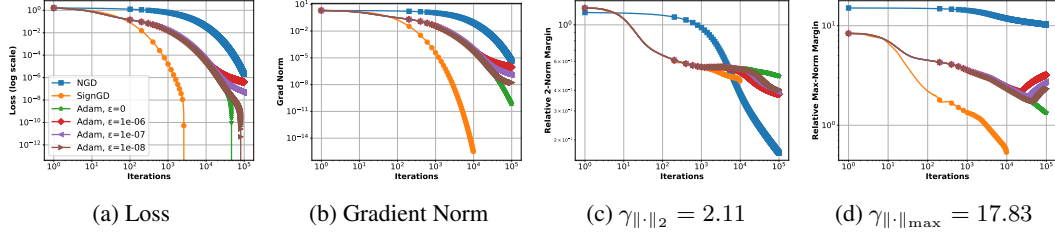


Figure 4: Implicit bias of NGD, SignGD, and Adam on multiclass separable data ( $k = 5, d = 25$ , and 50 data points in each class). **(a,b)** Loss and gradient 2-norm vs. iterations: SignGD converges faster than others. **(c)** We normalize the iterates w.r.t. 2-norm (aka Frobenius), compute the margin, then plot its difference to the dataset’s max-margin w.r.t. 2-norm (given in captions). Only NGD converges to the max 2-norm margin. **(d)** Same as (c) with 2-norm replaced by max-norm. Margins of SignGD/Adam (with  $\epsilon = 0$ ) converge to max-margin w.r.t max-norm. For SignGD, the training is stopped after  $10^4$  iterations due to the numerical instabilities caused by the small gradient norm.

977 *Proof.* Recall that  $t_{a_1} = \Theta(d^{2/a}) = C_{a_1} d^{2/a}$ , and the condition on  $t_{a_2}$  is  $\frac{2\mathcal{L}(\mathbf{W}_0) + 4B\alpha \sum_{s=1}^{t_{a_1}-1} \eta_s}{\gamma \sum_{s=t_{a_1}}^{t_{a_2}} \eta_s} \leq$   
978  $\frac{\tilde{\mathcal{L}}}{2}$ , where  $\tilde{\mathcal{L}} = \min\{\frac{\log 2}{n}, \frac{1}{\sqrt{4B^2 + \alpha_V \eta_0}}\}$ . Then, we apply integral approximations and the rest of the  
979 proof can be found in [60, Corollary 4.7 and Lemma C.1].  $\square$

980 **Remark 3.** These rates match exactly those in the binary case of Zhang et al. [60] with logarithmic  
981 dependence on the initialization parameter  $\omega$  (Ass. 5). This is only made possible through the  
982 fine-grained per-class bounding of the first and second moments using both  $\mathcal{G}_c(\mathbf{W})$  and  $\mathcal{Q}_c(\mathbf{W})$ .  
983 Note that Lemma 36 takes the same form as Zhang et al. [60, Lemma A.4]. However, without the tight  
984 per-class bound and the equivalent decomposition of  $\mathcal{G}(\mathbf{W})$  using either  $\mathcal{Q}_c(\mathbf{W})$  or  $\mathcal{G}_c(\mathbf{W})$ , an extra  
985 factor of  $k$  would appear. Interestingly, our rates for SignGD in Corollary 1 reveal a theoretical gap:  
986 Adam’s optimal choice  $a = \frac{2}{3}$  yields  $\mathcal{O}(\frac{d \log(t) + nd}{t^{1/3}})$  while SignGD achieves  $\mathcal{O}(\frac{\log(t) + n}{t^{1/2}})$  with  $a = \frac{1}{2}$ .  
987 Despite achieving tightness w.r.t. class-dimension ( $k$ ), this gap emerges from our entry-wise analysis  
988 of the  $\clubsuit$  term in Lemma 36 across the feature dimension ( $d$ ) using scalar functions  $\mathcal{G}_c(\mathbf{W})$ ,  $\mathcal{Q}_c(\mathbf{W})$ .  
989 Closing this theoretical gap—revealed through our NSD analysis—that also appears in the binary case  
990 [60], forms an important direction for future work.

991 **Numerical Validations** We test the margin converge of Adam with different stability constants ( $\epsilon$ )  
992 chosen from the set  $\{0, 10^{-6}, 10^{-7}, 10^{-8}\}$ . We make the following observations: **(1) SignGD/Adam**  
993 **vs NGD:** SignGD and Adam (with zero stability) iterates favor the max-norm margin over the 2-norm  
994 margin. The opposite is true for NGD (Figs. 4c, 4d); **(2) Speed of convergence:** For the learning  
995 rate schedules considered, the margin convergence of SignGD is faster than that of Adam (with zero  
996 stability constant), consistent with our theoretical results (Fig. 4d); **(3) Effect of Adam’s stability**  
997 **constant:** For Adam with non-zero stability constants, when the magnitude of the gradient (or the  
998 second moment) is above the stability constant, the convergence to the max-norm margin is favored  
999 over that to the 2-norm margin. However, when the gradient values approach or fall below the stability  
1000 constant, the max-norm margin starts to decrease while the 2-norm margin increases. This is shown in  
1001 Figs. 5a and 5b (compared against the case of zero stability constant in which the max-norm margin  
1002 keeps increasing). Moreover, the max-norm to 2-norm margin transition occurs earlier in training for  
1003 larger stability constants. The experiments in Figs 5d and 5c further confirm this two-phase behavior  
1004 given the correlations to the 2-norm separator  $\mathbf{V}_2$  rise while those to the max-norm separator  $\mathbf{V}_\infty$   
1005 plateau in the late training stage. These experiments confirm that the results of Wang et al. [52] also  
1006 hold in the multiclass setting provided that the stability constant is non-zero and the magnitudes of the  
1007 second moment are small (compared against the stability constant). Note, however, that the stability  
1008 constant is typically chosen to be very small ( $\epsilon \sim 10^{-8}$ ) and the training is not long enough until  
1009 all gradients are below this value. Thus, the max-norm margin convergence results are practically  
1010 relevant.

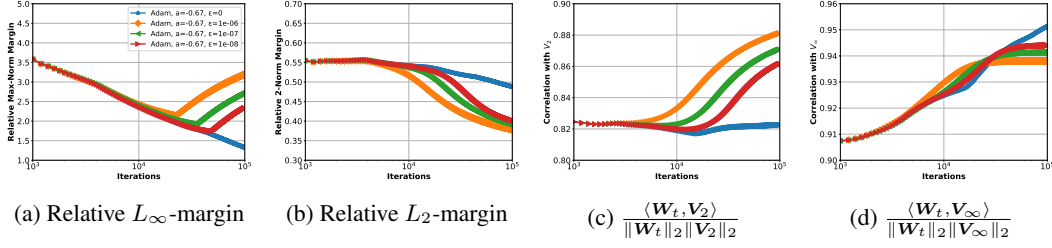


Figure 5: Effects of non-zero stability constants  $\epsilon$  for Adam. We focus on late training stage between iteration  $10^3$  to iteration  $10^5$ . **(a)** Gradient 2-norm vs. iterations. **(b)** Same quantity as Figure 4d. We observe the max-norm margin of non-zero stability constants decreases after gradient magnitudes approach or fall below the values of the constants (drawn as horizontal lines in (a)). **(c)** Same quantity as Figure 4c. After  $10^4$  iterations, the non-zero stability constants start to increase towards the max 2-norm margin. **(d, e)** Correlations between  $\mathbf{W}_t$  and max-margin classifiers  $\mathbf{V}_2$ ,  $\mathbf{V}_\infty$  against iterations. Considering correlations to  $\mathbf{V}_2$ , its value stays nearly constant for the zero stability constant, whereas rise after  $10^4$  iterations for non-zero ones. We also observe the transitions occur earlier for larger stability constants. Considering correlations to  $\mathbf{V}_\infty$ , the values rise and plateau for non-zero stability constants. However, its value keeps increasing for the zero stability constant.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: See Sec. 1.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: See Sec. 8.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution

is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: See Sec. 2.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: See Sec. 6.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.



- 1101 (b) If the contribution is primarily a new model architecture, the paper should describe  
1102 the architecture clearly and fully.
- 1103 (c) If the contribution is a new model (e.g., a large language model), then there should  
1104 either be a way to access this model for reproducing the results or a way to reproduce  
1105 the model (e.g., with an open-source dataset or instructions for how to construct  
1106 the dataset).
- 1107 (d) We recognize that reproducibility may be tricky in some cases, in which case  
1108 authors are welcome to describe the particular way they provide for reproducibility.  
1109 In the case of closed-source models, it may be that access to the model is limited in  
1110 some way (e.g., to registered users), but it should be possible for other researchers  
1111 to have some path to reproducing or verifying the results.

## 1112 5. Open access to data and code

1113 Question: Does the paper provide open access to the data and code, with sufficient instruc-  
1114 tions to faithfully reproduce the main experimental results, as described in supplemental  
1115 material?

1116 Answer: [Yes]

1117 Justification: The code is provided in supplemental materials.

1118 Guidelines:

- 1119 • The answer NA means that paper does not include experiments requiring code.
- 1120 • Please see the NeurIPS code and data submission guidelines ([https://nips.cc/  
1121 public/guides/CodeSubmissionPolicy](https://nips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 1122 • While we encourage the release of code and data, we understand that this might not be  
1123 possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not  
1124 including code, unless this is central to the contribution (e.g., for a new open-source  
1125 benchmark).
- 1126 • The instructions should contain the exact command and environment needed to run to  
1127 reproduce the results. See the NeurIPS code and data submission guidelines ([https:  
1128 //nips.cc/public/guides/CodeSubmissionPolicy](https://nips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 1129 • The authors should provide instructions on data access and preparation, including how  
1130 to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- 1131 • The authors should provide scripts to reproduce all experimental results for the new  
1132 proposed method and baselines. If only a subset of experiments are reproducible, they  
1133 should state which ones are omitted from the script and why.
- 1134 • At submission time, to preserve anonymity, the authors should release anonymized  
1135 versions (if applicable).
- 1136 • Providing as much information as possible in supplemental material (appended to the  
1137 paper) is recommended, but including URLs to data and code is permitted.

## 1138 6. Experimental setting/details

1139 Question: Does the paper specify all the training and test details (e.g., data splits, hyper-  
1140 parameters, how they were chosen, type of optimizer, etc.) necessary to understand the  
1141 results?

1142 Answer: [Yes]

1143 Justification: See Sec. 6.

1144 Guidelines:

- 1145 • The answer NA means that the paper does not include experiments.
- 1146 • The experimental setting should be presented in the core of the paper to a level of detail  
1147 that is necessary to appreciate the results and make sense of them.
- 1148 • The full details can be provided either with the code, in appendix, or as supplemental  
1149 material.

## 1150 7. Experiment statistical significance

1151 Question: Does the paper report error bars suitably and correctly defined or other appropriate  
1152 information about the statistical significance of the experiments?

1153 Answer: [No]

1154 Justification: The algorithm is deterministic with no batching involved.



Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: See Sec. 6.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The paper follows the code of ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: The paper is theoretical.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: No existing assets used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets)

1263 has curated licenses for some datasets. Their licensing guide can help determine the  
 1264 license of a dataset.

- 1265 • For existing datasets that are re-packaged, both the original license and the license of  
 1266 the derived asset (if it has changed) should be provided.
- 1267 • If this information is not available online, the authors are encouraged to reach out to  
 1268 the asset’s creators.

1269 **13. New assets**

1270 Question: Are new assets introduced in the paper well documented and is the documentation  
 1271 provided alongside the assets?

1272 Answer: [NA]

1273 Justification: No new assets released.

1274 Guidelines:

- 1275 • The answer NA means that the paper does not release new assets.
- 1276 • Researchers should communicate the details of the dataset/code/model as part of their  
 1277 submissions via structured templates. This includes details about training, license,  
 1278 limitations, etc.
- 1279 • The paper should discuss whether and how consent was obtained from people whose  
 1280 asset is used.
- 1281 • At submission time, remember to anonymize your assets (if applicable). You can either  
 1282 create an anonymized URL or include an anonymized zip file.

1283 **14. Crowdsourcing and research with human subjects**

1284 Question: For crowdsourcing experiments and research with human subjects, does the paper  
 1285 include the full text of instructions given to participants and screenshots, if applicable, as  
 1286 well as details about compensation (if any)?

1287 Answer: [NA]

1288 Justification: No crowdsourcing and research with human subjects are involved.

1289 Guidelines:

- 1290 • The answer NA means that the paper does not involve crowdsourcing nor research with  
 1291 human subjects.
- 1292 • Including this information in the supplemental material is fine, but if the main contribu-  
 1293 tion of the paper involves human subjects, then as much detail as possible should be  
 1294 included in the main paper.
- 1295 • According to the NeurIPS Code of Ethics, workers involved in data collection, curation,  
 1296 or other labor should be paid at least the minimum wage in the country of the data  
 1297 collector.

1298 **15. Institutional review board (IRB) approvals or equivalent for research with human**  
 1299 **subjects**

1300 Question: Does the paper describe potential risks incurred by study participants, whether  
 1301 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)  
 1302 approvals (or an equivalent approval/review based on the requirements of your country or  
 1303 institution) were obtained?

1304 Answer: [NA]

1305 Justification: No crowdsourcing and research with human subjects are involved.

1306 Guidelines:

- 1307 • The answer NA means that the paper does not involve crowdsourcing nor research with  
 1308 human subjects.
- 1309 • Depending on the country in which research is conducted, IRB approval (or equivalent)  
 1310 may be required for any human subjects research. If you obtained IRB approval, you  
 1311 should clearly state this in the paper.
- 1312 • We recognize that the procedures for this may vary significantly between institutions  
 1313 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the  
 1314 guidelines for their institution.
- 1315 • For initial submissions, do not include any information that would break anonymity (if  
 1316 applicable), such as the institution conducting the review.

1317 **16. Declaration of LLM usage**  
1318 Question: Does the paper describe the usage of LLMs if it is an important, original, or  
1319 non-standard component of the core methods in this research? Note that if the LLM is used  
1320 only for writing, editing, or formatting purposes and does not impact the core methodology,  
1321 scientific rigorousness, or originality of the research, declaration is not required.  
1322 Answer: [NA]  
1323 Justification: No LLMs usage involved.  
1324 Guidelines:  
1325 • The answer NA means that the core method development in this research does not  
1326 involve LLMs as any important, original, or non-standard components.  
1327 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)  
1328 for what should or should not be described.