Implicit Bias of Spectral Descent and Muon on Multiclass Separable Data

Chen Fan FANCHEN3@OUTLOOK.COM
University of British Columbia, British Columbia , Canada

Mark Schmidt SCHMIDTM@CS.UBC.CA
University of British Columbia, British Columbia , Canada
Canada CIFAR AI Chair (Amii)

Christos Thrampoulidis CTHRAMPO@ECE.UBC.CA
University of British Columbia, British Columbia , Canada

Abstract

Different gradient-based methods for optimizing overparameterized models can all achieve zero training error yet converge to distinctly different solutions inducing different generalization properties. We provide the first complete characterization of implicit optimization bias for p-norm normalized steepest descent (NSD) and momentum steepest descent (NMD) algorithms in multi-class linear classification with cross-entropy loss. Our key theoretical contribution is proving that these algorithms converge to solutions maximizing the margin with respect to the classifier matrix's p-norm, with established convergence rates. These results encompass important special cases including Spectral Descent and Muon, which we show converge to max-margin solutions with respect to the spectral norm. A key insight of our contribution is that the analysis of general entry-wise and Schatten p-norms can be reduced to the analysis of NSD/NMD with max-norm by exploiting a natural ordering property between all p-norms relative to the max-norm and its dual sum-norm. Our results demonstrate that the multi-class linear setting, which is inherently richer than the binary counterpart, provides the most transparent framework for studying implicit biases of matrix-parameter optimization algorithms.

1. Introduction

The ever-increasing training cost of large language models (LLMs) has demanded better optimizer designs with improved performance and efficiency [1, 10, 20]. The de facto standard optimizers for deep learning training are Adam and AdamW [30, 34]. However, these algorithms that employ diagonal preconditioners to independently adjust the learning rate of each coordinate, may fail to capture their inter-dependencies and fully leverage the geometry of the loss landscape [67]. This has spurred a series of research efforts on improving Adam or AdamW's computational efficiency [17, 21, 43, 68], with LLM-training as the target application domain [28, 32, 39, 56].

A noticeable work by Jordan et al. [28] proposed the Muon optimizer, which was shown to have remarkable performances on NanoGPT benchmarks. More recently, it has been shown that Muon can be used for large-scale LLM training with the potential to replace AdamW as the standard choice [32]. The key step in Muon is to orthogonalize the updates via the Newton-Schulz iteration [6, 28]. More precisely, the update (denoted as $\boldsymbol{\Delta}$) is (approximately) replaced by the product of its singular-vector matrices $\boldsymbol{U}\boldsymbol{V}^T$ (where the (truncated) singular value decomposition (SVD) of $\boldsymbol{\Delta}$ is $\boldsymbol{\Delta} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$). Even though the benefits of orthogonalization are not fully understood, Jordan et al.

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 $\boldsymbol{\nabla} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ and $\boldsymbol{M} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{V}}^T$ respectively.

Method	Norm Constraint	Update $\boldsymbol{\Delta}$	Refernce	Rate [2]
NGD	Unit $\|\cdot\|_2$-ball	$\frac{\boldsymbol{\nabla}}{\|\boldsymbol{\nabla}\|_2}$	Hazan et al. [22]	-
NMD-GD		$\frac{\boldsymbol{M}}{\|\boldsymbol{M}\|_2}$	Cutkosky and Mehta [13]	-
SignGD	Unit $\|\cdot\|_{\max}$-ball	$\texttt{sign}(\boldsymbol{\nabla})$	Bernstein et al. [7]	-
Signum		$\texttt{sign}(\boldsymbol{M})$	Bernstein et al. [7]	-
Spectral-GD	Unit $\|\|\|\cdot\|\|\|_\infty$-ball	$\boldsymbol{U}\boldsymbol{V}^T$	Bernstein and Newhouse [6]	$O(\frac{\log t+n}{t^{1/2}})$
Muon [1]		$\tilde{\boldsymbol{U}}\tilde{\boldsymbol{V}}^T$	Jordan et al. [28]	$O(\frac{d\log t+dn}{t^{1/2}})$

[1] We consider EMA-style momentum of the form (4).
[2] NGD and SignGD rates are the same as Spectral-GD; Signum and NMD-GD rates are the same as Muon.

[28] pointed out that it could promote updates in directions of small magnitudes given the weight matrices are typically low-rank. Moreover, if the above SVD approximation is exact and gradient accumulations are turned off, then Muon becomes spectral descent [6, 12], which is the (normalized) steepest descent w.r.t the spectral norm [6]. As noted by Bernstein and Newhouse [6], spectral descent is also Shampoo (which won the AlgoPerf competition [14, 44]) 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:

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

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, 31, 39]). However, modern machine learning models are overparameterized, leading to multiple weight configurations that achieve identical training loss but exhibit markedly different generalization properties [5, 66]. 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) [7], normalized gradient descent (entry-wise 2-norm) [22], and spectral descent (Schatten ∞-norm) [6] as special cases.

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 [7], normalized momentum gradient descent [13], and Muon [28]. The margin convergence rates of various algorithms are summarized in Table 1, and numerical validations can be found in App. A.

2. Preliminaries

Notations We write $\|\boldsymbol{A}\|$ to refer to any entry-wise or Schatten p-norm with $p \geq 1$, and denote by $\|\boldsymbol{A}\|_*$ the dual-norm with respect to the standard matrix inner product $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B})$. We denote the gradient and its value at iteration t as $\boldsymbol{\nabla} := \nabla \mathcal{L}(\boldsymbol{W})$ and $\boldsymbol{\nabla}_t := \nabla \mathcal{L}(\boldsymbol{W}_t)$ respectively. Let $\mathbb{S} : \mathbb{R}^k \to \triangle^{k-1}$ the softmax map of k-dimensional vectors to the probability simplex $\triangle^{k-1}$ such that for any $\boldsymbol{a} \in \mathbb{R}^k$, it holds that $\mathbb{S}(\boldsymbol{a}) = \left[\frac{\exp(\boldsymbol{a}[c])}{\sum_{c \in [k]} \exp(\boldsymbol{a}[c])}\right]_{c=1}^k \in \triangle^{k-1}$. Let $\mathbb{S}_c(\boldsymbol{v})$ denote the c-th entry of $\mathbb{S}(\boldsymbol{v})$, and let $\{\boldsymbol{e}_c\}_{c=1}^k$ be the standard basis vectors of $\mathbb{R}^k$.

Setup Consider a multiclass classification problem with training data $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n$ and labels $y_1, \ldots, y_n$. Each datapoint $\boldsymbol{h}_i \in \mathbb{R}^d$ is a vector in a d-dimensional embedding space (denote data matrix $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n]^\top \in \mathbb{R}^{n \times d}$), and each label $y_i \in [k]$ represents one of k classes. We assume each class contains at least one datapoint. The classifier $f_{\boldsymbol{W}} : \mathbb{R}^d \to \mathbb{R}$ is a linear model with weight matrix $\boldsymbol{W} \in \mathbb{R}^{k \times d}$. The model outputs logits $\boldsymbol{\ell}_i = f_{\boldsymbol{W}}(\boldsymbol{h}_i) = \boldsymbol{W}\boldsymbol{h}_i$ for $i \in [n]$, which are passed through the softmax map to produce class probabilities $\hat{p}(c|\boldsymbol{h}_i) = \mathbb{S}_c(\boldsymbol{\ell}_i)$. We train using empirical risk minimization (ERM): $\mathcal{L}_{\text{ERM}}(\boldsymbol{W}) := -\frac{1}{n}\sum_{i \in [n]} \ell\left(\boldsymbol{W}\boldsymbol{h}_i; y_i\right)$, where the loss function ℓ takes as input the logits of a datapoint and its label. The predominant choice in classification is the CE loss: $\mathcal{L}(\boldsymbol{W}) := -\frac{1}{n}\sum_{i \in [n]} \log\left(\mathbb{S}_{y_i}(\boldsymbol{W}\boldsymbol{h}_i)\right)$ (see App. H for other losses). Define the maximum margin of the dataset w.r.t. any entry-wise or Schatten p-norm $\|\cdot\|$ as

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

Optimization Methods We study iterative algorithms that update the weight matrix via: $\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta_t \boldsymbol{\Delta}_t$. For the NSD family [9], the update direction[1] w.r.t. the norm $\|\cdot\|$ is:

$$\boldsymbol{\Delta}_t := \arg\max_{\|\boldsymbol{\Delta}\| \leq 1} \langle \boldsymbol{\nabla}_t, \boldsymbol{\Delta} \rangle \,. \tag{2}$$

Note that this reduces to SignGD, Coordinate Descent (e.g., Nutini et al. [37]), 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:

$$\text{SignGD:} \quad \boldsymbol{\Delta}_t = \mathtt{sign}(\boldsymbol{\nabla}_t), \quad \text{and} \quad \text{NGD:} \quad \boldsymbol{\Delta}_t = \boldsymbol{\nabla}_t / \|\boldsymbol{\nabla}_t\|_2,$$

[1] 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. (2). For $p = 1, \infty$ the maximizer is not necessarily unique and our results hold for any choice of $\boldsymbol{\Delta}_t$ in the set of maximizers; see e.g. Ziętak [69].

where the $\texttt{sign}(\cdot)$ and division $\frac{\cdot}{\cdot}$ operations are applied entry-wise. In the special case of spectral norm (i.e. $|||\cdot|||_\infty$), this becomes the Spectral-GD, for which $\boldsymbol{\Delta}_t = \boldsymbol{U}_t \boldsymbol{V}_t^T$, where $\boldsymbol{U}_t$ and $\boldsymbol{V}_t$ are the left/right singular matrices of $\boldsymbol{\nabla}_t$ respectively (i.e., $\boldsymbol{\nabla}_t = \boldsymbol{U}_t \boldsymbol{\Sigma}_t \boldsymbol{V}_t^T$ with singular values in $\boldsymbol{\Sigma}_t > 0$ 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 $\|\cdot\|$

$$\boldsymbol{\Delta}_t := \arg\max_{\|\boldsymbol{\Delta}\| \leq 1} \langle \boldsymbol{M}_t, \boldsymbol{\Delta} \rangle \,, \tag{3}$$

where the momentum $\boldsymbol{M}_t$ is computed as the the exponential moving averages of the gradient as

$$\mathbf{M}_t = \beta_1 \mathbf{M}_{t-1} + (1 - \beta_1) \boldsymbol{\nabla}_t. \tag{4}$$

This form of momentum is also known as the heavy-ball or the SGDM-style momentum [16, 33, 40]. 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:

$$\text{Signum:} \quad \boldsymbol{\Delta}_t = \texttt{sign}(\boldsymbol{M}_t), \quad \text{and} \quad \text{NMD-MD:} \quad \boldsymbol{\Delta}_t = \boldsymbol{M}_t / \|\boldsymbol{M}_t\|_2.$$

When spectral norm is used in (3), this becomes Muon [2] for which the SVD is on $\boldsymbol{M}_t$ (i.e. $\boldsymbol{M}_t = \tilde{\boldsymbol{U}}_t \tilde{\boldsymbol{\Sigma}}_t \tilde{\boldsymbol{V}}_t^T$) and the update direction is $\boldsymbol{\Delta}_t = \tilde{\boldsymbol{U}}_t \tilde{\boldsymbol{V}}_t^T$. Note that Muon reduces to Spectral-GD when the momentum parameter β_1 is set to 0 (similar reductions hold for Signum (to SignGD) and NMD-GD (to NGD) as well). Discussions on the related works can be found in App B.

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 γ is strictly positive, an assumption routinely used in previous works [19, 36, 41, 45, 63].

Assumption 1 *There exists* $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ *such that* $\min_{c \neq y_i} (\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^T \boldsymbol{W} \boldsymbol{h}_i > 0$ *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. [8], Nacson et al. [36], and Sun et al. [47]) including Adam [65].

Assumption 2 *The learning rate schedule* $\{\eta_t\}$ *is decreasing with respect to* t *and satisfies the following conditions:* $\lim_{t \to \infty} \eta_t = 0$ *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. [65, Lemma C.1]. It is used in our analysis of NMD.

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 [23], Nacson et al. [36], Wu et al. [63], and Zhang et al. [65].

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

[2] The implementation in Jordan et al. [28] uses Nesterov-type momentum: Newton-Schulz iteration applied to $\beta_1 \boldsymbol{M}_t + \boldsymbol{\nabla}_t$ instead of $\beta_1 \boldsymbol{M}_{t-1} + \boldsymbol{\nabla}_t$ [32].

3. Implicit Bias of NSD and NMD

In this section, we show the implicit bias results of NSD and NMD algorithms. To do the analysis, we introduce a unified 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 (details in App. C). We first state the convergence result of NSD (proof details in App. F).

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

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

Remark 2 *For margin convergence rates of NSD, Nacson et al. [36] showed a rate of $\mathcal{O}(\frac{\log t}{t^{1/2}})$ in the binary setting, limited to the entry-wise p-norms and the exponential loss. Compared to this, our results hold for the more practical setting of multiclass data and CE loss. To the best of our knowledge, this is the first non-asymptotic result on the implicit bias of spectral-GD for linear multiclass separable data, and it holds for other p-norms as well. Upon completion of this work, we became aware of an update on the arXiv version of Tsilivis et al. [53], which includes an extension of their previous results to steepest descent w.r.t. the spectral norm. In comparison to ours, their gradient-flow analysis applies to homogeneous neural networks with the restriction of infinitesimal step-sizes. Moreover, it does not include normalization nor momentum (like Muon, which we analyze), and the convergence is (asymptotic) to a KKT point of a spectral-norm margin maximization problem.*

For the analysis of NMD, we additionally use the same proxy function to bound the sum-norm difference between the gradient and the momentum, which translates to a bound on the dual norm through the fundamental norm-relationships used in the study of NSD (see Lemma 25 in App. G). Then, we obtain the following rate for NMD (proof details in App. G).

Theorem 3 *Suppose that Ass. 1, 2, 3, and 4 hold, the margin gap of NMD with $\eta_t = \Theta(\frac{1}{t^{1/2}})$ is $O(\frac{d\log t + dn}{t^{1/2}})$.*

Remark 4 *Wang et al. [58] studied implicit bias of un-normalized GD with momentum, and showed its iterates converge asymptotically to the max 2-norm margin solution. In contrast, our rates are non-asymptotic and cover a much wider family of algorithms converging to non-Euclidean geometric margins (w.r.t. entry-wise/Schatten norms). Note the convergence rate of NMD matches that of NSD (Thm. 1) up to a factor of d. It could be interesting to remove this dependence in a future work.*

4. Conclusion

We have characterized the margin convergence rates of Spectral-GD and Muon for multilcass 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. Future directions include removing the factor-d from the bound of NMD and studying other related algorithms such as Shampoo that involves non-diagonal preconditioners. It is also important to extend our results to (multiclass) non-separable settings [51] and nonlinear models such as diagonal neural nets [38], self-attention mechanisms [3, 29, 48, 49, 55] and homogeneous neural nets [11, 35, 53], 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] Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. *Advances in neural information processing systems*, 36:48314–48362, 2023.

[4] 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.

[5] 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.

[6] Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

[7] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd: Compressed optimisation for non-convex problems. *arXiv preprint arXiv:1802.04434*, 2018.

[8] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

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

[10] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[11] 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.

[12] 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.

[13] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized sgd. In *International conference on machine learning*, pages 2260–2268. PMLR, 2020.

[14] 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.

[15] Steven Diamond and Stephen Boyd. Cvxpy: A python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[21] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018.

[22] Elad Hazan, Kfir Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. *Advances in neural information processing systems*, 28, 2015.

[23] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on learning theory*, pages 1772–1798. PMLR, 2019.

[24] Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

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

[26] 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.

[27] Ziwei Ji, Nathan Srebro, and Matus Telgarsky. Fast margin maximization via dual acceleration. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2021.

[28] 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/.

[29] Addison Kristanto Julistiono, Davoud Ataee Tarzanagh, and Navid Azizan. Optimizing attention with mirror descent: Generalized max-margin token selection. *arXiv preprint arXiv:2410.14581*, 2024.

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[31] Jiaxiang Li and Mingyi Hong. A note on the convergence of muon and further. *arXiv preprint arXiv:2502.02900*, 2025.

[32] 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.

[33] 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.

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[41] 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.

[42] Saharon Rosset, Ji Zhu, and Trevor J. Hastie. Margin maximizing loss functions. In *NIPS*, 2003.

[43] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines, 2023.

[49] Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism, 2023.

[50] Matus Telgarsky. Margins, shrinkage, and boosting. In *International Conference on Machine Learning*, pages 307–315. PMLR, 2013.

[51] Christos Thrampoulidis. Implicit optimization bias of next-token prediction in linear models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[52] Christos Thrampoulidis, Ganesh R Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *arXiv preprint arXiv:2208.05512*, 2022.

[53] 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.

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

[55] Bhavya Vasudeva, Puneesh Deora, and Christos Thrampoulidis. Implicit bias and fast convergence rates for self-attention. *arXiv preprint arXiv:2402.05738*, 2024.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] Nan Wang, Zhen Qin, Le Yan, Honglei Zhuang, Xuanhui Wang, Michael Bendersky, and Marc Najork. Rank4class: a ranking formulation for multiclass classification. *arXiv preprint arXiv:2112.09727*, 2021.

[61] Yutong Wang and Clayton Scott. Unified binary and multiclass margin-based classification. *Journal of Machine Learning Research*, 25(143):1–51, 2024.

[62] Jingfeng Wu, Peter L Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency. *arXiv preprint arXiv:2402.15926*, 2024.

[63] Jingfeng Wu, Vladimir Braverman, and Jason D Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. *Advances in Neural Information Processing Systems*, 36, 2024.

[64] Shuo Xie and Zhiyuan Li. Implicit bias of adamw: L-infinity norm constrained optimization. *arXiv preprint arXiv:2404.04454*, 2024.

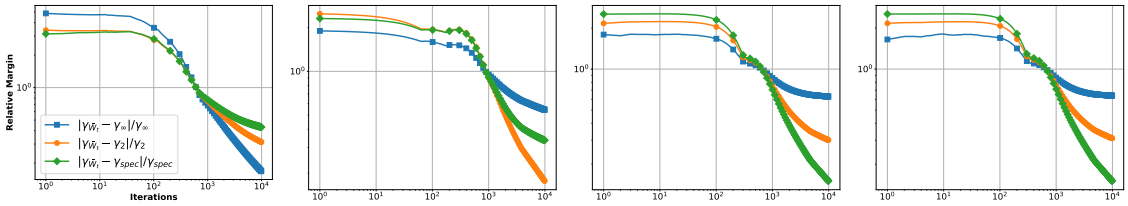
[65] Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of adam on separable data. *arXiv preprint arXiv:2406.10650*, 2024.

[66] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2017.

[67] Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhiquan Luo. Why transformers need adam: A hessian perspective. *Advances in Neural Information Processing Systems*, 37:131786–131823, 2024.

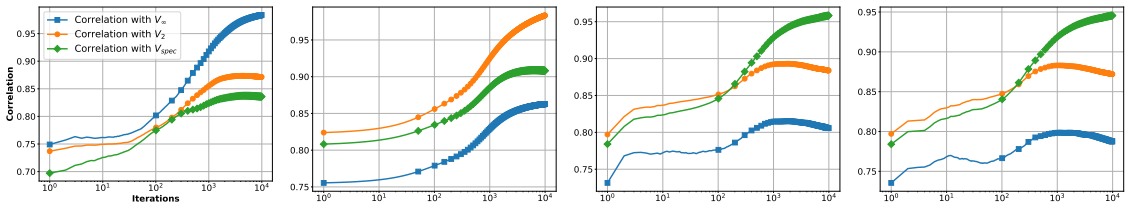
[68] Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik P Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. *arXiv preprint arXiv:2406.16793*, 2024.

[69] Krystyna Zięłak. On the characterization of the extremal points of the unit sphere of matrices. *Linear Algebra and its Applications*, 106:57–75, 1988.



(*a*) SignGD (*b*) NGD (*c*) Spectral-GD (*d*) Muon

Figure 1: **(a)** We normalize the iterates of SignGD w.r.t. the max-norm (denoted as $\bar{\boldsymbol{W}}_t$), compute the margin (denoted as $\gamma_{\bar{\boldsymbol{W}}_t}$), then plot its difference to data margins $\gamma_{\|\cdot\|_\infty}$, $\gamma_{\|\cdot\|_2}$, and $\gamma_{|||\cdot|||_\infty}$ (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.

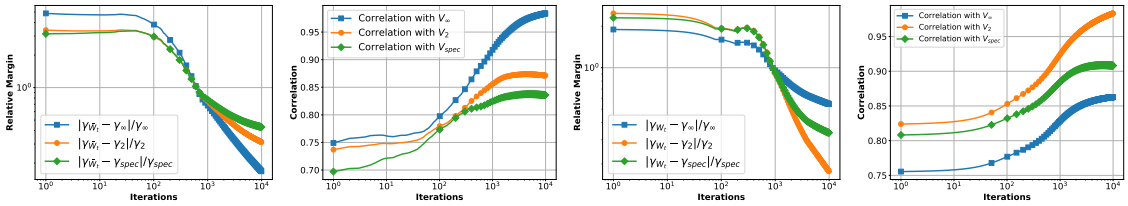


(*a*) SignGD (*b*) NGD (*c*) Spectral-GD (*d*) Muon

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

Appendix A. Experiments

We generate snythetic multiclass separable data as follows: $k = 10$ class centers are sampled from a standard normal distribution; within each class, data is sampled from normal distribution $\mathcal{N}(0, \sigma^2 I), \sigma = 0.1$. We set $d = 25$, sample 50 data points for each class, and ensure that margin is positive (thus data is separable). We run different algorithms to minimize CE loss using $\eta_t = \frac{\eta_0}{t^a}$ ($\eta_0 = 0.1$ for SignGD and NGD; $\eta_0 = 0.05$ for Spectral-GD and Muon), where (based on our theorems) a is set to $1/2$. We apply truncated SVD on the gradient and momentum for Spectral-GD and Muon respectively. Data margins w.r.t. different norms are found via CVXPY [15]. We denote max-margin classifiers defined w.r.t. the 2-norm, the max-norm, and the spectral-norm as $\boldsymbol{V}_2$, $\boldsymbol{V}_\infty$, and $\boldsymbol{V}_{\text{spec}}$ respectively. Based on the margin-gap results in Figure 1, we observe that SignGD, NGD, and Spectral-GD favor max-norm, 2-norm, and spectral-norm margin respectively. Besides this, the behavior of Muon is very similar to that of Spectral-GD (in agreement with our theories). Figure 2 further confirms that the iterates of these algorithms correlate well with the corresponding max margin separators. Furthermore, based on the experiment results of Signum and NMD-GD shown in Figure 3, we conclude that their margin convergence properties are the same as SignGD and NGD respectively.



(*a*) Signum (*b*) Signum (*c*) NMD-GD (*d*) NMD-GD

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 $\boldsymbol{V}_\infty$, $\boldsymbol{V}_2$, and $\boldsymbol{V}_{\text{spec}}$ 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.

Appendix B. Related Works

Starting with GD, the foundational result by Soudry et al. [45] showed that gradient descent optimization of logistic loss on linearly separable data converges in direction to the L_2 max-margin classifier at a rate $O(1/\log(t))$. Contemporaneous work by Ji and Telgarsky [23] generalized this by relaxing the data separability requirement. Ji et al. [26] later connected these findings to earlier work on regularization paths of logistic loss minimization [42], which enabled extensions to other loss functions (e.g., those with polynomial tail decay). More recently, Wu et al. [63] 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. [36] 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 [25] using constant step-sizes, and further to $O(1/t^2)$ through a specific momentum formulation [27]. Besides linear classifications, implicit bias of GD has been studied for least squares [4, 18, 19], homogeneous [24, 35, 62] or non-homogeneous neural networks [11], and matrix factorization [18]; see Vardi [54] for a survey.

All the above mentioned works focus almost exclusively on binary classification. The noticeable gap in analysis of multiclass classification in most existing literature is highlighted by Thrampoulidis et al. [52], and more recently emphasized by Ravi et al. [41], who extended the implicit bias result of Soudry et al. [45] to multiclass classification for losses with exponential tails, including CE, multiclass exponential, and PairLogLoss. Their approach leverages a framework of Wang and Scott [61] that allows multiclass losses and separability conditions to be written in margin-based forms similar to binary cases.

Beyond GD, Gunasekar et al. [19] and Nacson et al. [36] 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. [46, 47] 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. [59] 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. [53] 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. [57] 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. [65] 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_∞-norm. The study of excluding the stability constant is also the focus of another recent work on the implicit bias of AdamW [64], where the authors again establish that convergence aligns with the L_∞ geometry.

Appendix C. 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 $\mathbb{S}'(\boldsymbol{v}) = \operatorname{diag}(\boldsymbol{v}) - \boldsymbol{v}\boldsymbol{v}^\top$):

$$\mathcal{L}(\boldsymbol{W}+\boldsymbol{\Delta}) = \mathcal{L}(\boldsymbol{W}) + \langle \nabla\mathcal{L}(\boldsymbol{W}), \boldsymbol{\Delta}\rangle + \frac{1}{2n}\sum\nolimits_{i\in[n]} \boldsymbol{h}_i^\top \boldsymbol{\Delta}^\top \mathbb{S}'(\boldsymbol{W}\boldsymbol{h}_i)\boldsymbol{\Delta}\boldsymbol{h}_i + o(\|\boldsymbol{\Delta}\|_F^3), \quad (5)$$

To bound the loss at $\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta_t\boldsymbol{\Delta}_t$, we must bound both terms in (5). For NSD updates in Eq. (2), the first term evaluates to $-\eta_t\|\nabla\mathcal{L}(\boldsymbol{W})\|_*$ (recall that $\|\cdot\|_*$ is the dual norm). This leads to two key tasks: (1) Lower-bounding the dual gradient norm; (2) Upper-bounding the second-order term.

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

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

$$\|\boldsymbol{A}\|_{\max} \leq |||\boldsymbol{A}|||_p \leq \|\boldsymbol{A}\|_{\text{sum}}, \quad \text{and} \quad \|\boldsymbol{A}\|_{\max} \leq \|\boldsymbol{A}\|_p \leq \|\boldsymbol{A}\|_{\text{sum}}. \quad (6)$$

These relationships (proved in Lemma 10 in App. E) are crucial for unifying the analysis of NSD and NMD algorithms w.r.t. either the entry-wise or the Schatten norms (details below).

Construction of $\mathcal{G}(\boldsymbol{W})$ Before showing our construction for the CE loss, it is insightful to discuss how previous works do this in the binary case with labels $y_{b,i} \in \{\pm 1\}$, classifier vector $\boldsymbol{w} \in \mathbb{R}^d$ and binary margin $\gamma_b \coloneqq \max_{\|\boldsymbol{w}\|\leq 1} \min_{i\in[n]} y_{b,i}\boldsymbol{w}^\top\boldsymbol{h}_i$. For exponential loss, Gunasekar et al. [19] showed that $\|\nabla\mathcal{L}(\boldsymbol{w})\| \geq \gamma_b\mathcal{L}(\boldsymbol{w})$. For logistic loss $\ell(t) = \log(1+\exp(-t))$, Zhang et al. [65] proved $\|\nabla\mathcal{L}(\boldsymbol{w})\|_1 \geq \gamma_b\mathcal{G}(\boldsymbol{w})$, where $\mathcal{G}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n |\ell'(y_{b,i}\boldsymbol{w}^\top\boldsymbol{h}_i)|$ and ℓ' is the first-order derivative. In both cases, one can take the common form $\mathcal{G}_b(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n |\ell'(y_{b,i}\boldsymbol{w}^\top\boldsymbol{h}_i)|$. The proof relies on showing $\gamma \leq \min_{\boldsymbol{r}\in\triangle^{n-1}}\|\boldsymbol{H}^T\boldsymbol{r}\|$ via Fenchel Duality [19, 50] and appropriately choosing $\boldsymbol{r}$.

In the multiclass setting, where the loss function is vector-valued, it is unclear how to extend the binary proof or definition of $\mathcal{G}(\boldsymbol{W})$. To this end, we realize that the key is in the proper manipulation of the gradient inner product $\langle \boldsymbol{A}, -\nabla\mathcal{L}(\boldsymbol{W})\rangle$ (for arbitrary matrix $\boldsymbol{A} \in \mathbb{R}^{k\times d}$). The CE gradient evaluates to $\nabla\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^n(\boldsymbol{e}_{y_i} - \mathbb{S}(\boldsymbol{W}\boldsymbol{h}_i))\boldsymbol{h}_i^\top$ and using the fact that $\mathbb{S}(\boldsymbol{W}\boldsymbol{h}_i) \in \triangle^{k-1}$, it turns out that we can express (details in Lemma 8): $\langle \boldsymbol{A}, -\nabla\mathcal{L}(\boldsymbol{W})\rangle = \frac{1}{n}\sum_{i\in[n]}\sum_{c\neq y_i}\mathbb{S}_c(\boldsymbol{W}\boldsymbol{h}_i)(\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^\top\boldsymbol{A}\boldsymbol{h}_i$.
This motivates defining $\mathcal{G}(\boldsymbol{W})$ as:

$$\mathcal{G}(\boldsymbol{W}) \coloneqq \frac{1}{n}\sum\nolimits_{i\in[n]}(1 - \mathbb{S}_{y_i}(\boldsymbol{W}\boldsymbol{h}_i)). \quad (7)$$

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

Lemma 5 (Lower bounding the gradient dual-norm) *For any $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ and any entry-wise or Schatten p-norm $\|\cdot\|$ with $p \geq 1$, it holds that $\|\nabla\mathcal{L}(\boldsymbol{W})\|_* \geq \gamma \cdot \mathcal{G}(\boldsymbol{W})$, where $\|\cdot\|_*$ is the dual-norm.*

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

$\mathcal{G}(\boldsymbol{W})$ and second-order term We now show how to bound the second-order term in (5). For this, we establish the following essential lemma whose proof relies on the relationships in (6).

Lemma 6 *For any entry-wise or Schatten p-norm $\|\cdot\|$ with $p \geq 1$, any $\boldsymbol{s} \in \Delta^{k-1}$ in the k-dimensional simplex, any index $c \in [k]$, and $\boldsymbol{v} \in \mathbb{R}^k$, it holds that*

$$\boldsymbol{v}^\top \left(\operatorname{diag}(\boldsymbol{s}) - \boldsymbol{s}\boldsymbol{s}^\top\right)\boldsymbol{v} \leq 4(1-s_c)\|\boldsymbol{v}\boldsymbol{v}^T\|.$$

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

$$\boldsymbol{v}^\top\boldsymbol{S}\boldsymbol{v} = \operatorname{tr}\left(\boldsymbol{S}\boldsymbol{v}\boldsymbol{v}^\top\right) \leq \|\boldsymbol{S}\|_q\|\boldsymbol{v}\boldsymbol{v}^\top\| \leq \|\boldsymbol{S}\|_{\text{sum}}\|\boldsymbol{v}\boldsymbol{v}^\top\|,$$

where $\|\cdot\|_q$ is the dual of $\|\cdot\|$ and the second inequality is by (6). Direct calculation yields $\|\boldsymbol{S}\|_{\text{sum}} = 2\sum_{c\in[k]} s_c(1-s_c)$. The advertised bound then follows by noting the following $\sum_{c\in[k]} s_c(1-s_c) \leq 2(1-s_{c'})$ for any $c' \in [k]$ (verified in Lemma 12 in App. E). ■

Next, we apply the above lemma with $\boldsymbol{v} \leftarrow \boldsymbol{\Delta}\boldsymbol{h}_i$ and $c \leftarrow y_i$, and further use the inequalities: $\|\boldsymbol{v}\boldsymbol{v}^T\|_p = \|\boldsymbol{v}\|_p^2 \leq \|\boldsymbol{\Delta}\|_p^2\|\boldsymbol{h}\|_q^2$ for entry-wise norms and $\left|\left|\left|\boldsymbol{v}\boldsymbol{v}^\top\right|\right|\right|_p = \|\boldsymbol{v}\|_2^2 \leq |||\boldsymbol{\Delta}|||_\infty^2\|\boldsymbol{h}\|_2^2 \leq |||\boldsymbol{\Delta}|||_p^2\|\boldsymbol{h}\|_2^2$ 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:

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

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

Lemma 7 (Properties of $\mathcal{G}(\boldsymbol{W})$ and $\mathcal{L}(\boldsymbol{W})$) *Let $\boldsymbol{W} \in \mathbb{R}^{k\times d}$. The followings hold: (i) Under Ass. 4, $2B\cdot\mathcal{G}(\boldsymbol{W}) \geq \|\nabla\mathcal{L}(\boldsymbol{W})\|_*$; (ii) $1 \geq \frac{\mathcal{G}(\boldsymbol{W})}{\mathcal{L}(\boldsymbol{W})} \geq 1 - \frac{n\mathcal{L}(\boldsymbol{W})}{2}$; (iii) If $\boldsymbol{W}$ satisfies $\mathcal{L}(\boldsymbol{W}) \leq \frac{\log 2}{n}$ or $\mathcal{G}(\boldsymbol{W}) \leq \frac{1}{2n}$, then $\mathcal{L}(\boldsymbol{W}) \leq 2\mathcal{G}(\boldsymbol{W})$.*

Lemma 7 (i) extends Lemma 5 by establishing a sandwich relationship between $\mathcal{G}(\boldsymbol{W})$ and the gradient's dual norm. The lemma's statements (ii) and (iii) show that $\mathcal{G}(\boldsymbol{W})$ can substitute for the loss - it lower bounds $\mathcal{L}(\boldsymbol{W})$ and serves as an upper bound when either $\mathcal{L}(\boldsymbol{W})$ or $\mathcal{G}(\boldsymbol{W})$ is sufficiently small. Specifically, the ratio $\mathcal{G}(\boldsymbol{W})/\mathcal{L}(\boldsymbol{W})$ 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 17 in App. E) that both sides of the sandwich relationship follow from the elementary fact that $\forall x > 0 : 1 - x \leq e^{-x} \leq 1 - x + x^2/2$.

Appendix D. Facts about CE loss and Softmax

Lemma 8 is on the gradient of the cross-entropy loss. It will be used for showing the form of $\mathcal{G}(\boldsymbol{W})$ in (7) lower bounds $\|\nabla\mathcal{L}(\boldsymbol{W})\|$ in Lemma 15.

Lemma 8 (Gradient) *Let CE loss*

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

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

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

$$\begin{aligned}\langle \boldsymbol{A}, -\nabla\mathcal{L}(\boldsymbol{W})\rangle &= \frac{1}{n}\sum_i (1-s_{iy_i})\left(\boldsymbol{e}_{y_i}^\top\boldsymbol{A}\boldsymbol{h}_i - \frac{\sum_{c\neq y_i} s_{ic}\,\boldsymbol{e}_c^\top\boldsymbol{A}\boldsymbol{h}_i}{(1-s_{iy_i})}\right)\\ &= \frac{1}{n}\sum_{i\in[n]}\sum_{c\neq y_i} s_{ic}\,(\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)^\top\boldsymbol{A}\boldsymbol{h}_i \end{aligned} \tag{8}$$

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

$$\langle \boldsymbol{A}, -\nabla\mathcal{L}(\boldsymbol{W})\rangle \geq \frac{1}{n}\sum_{i\in[n]}(1-s_{iy_i})\cdot\min_{c\neq y_i}(\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)^\top\boldsymbol{A}\boldsymbol{h}_i. \tag{9}$$

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

Lemma 9 is on the Taylor expansion of the loss. It will be used in showing the descent properties of NSD and NMD.

Lemma 9 (Hessian) *Let perturbation $\boldsymbol{\Delta}\in\mathbb{R}^{k\times d}$ and denote $\boldsymbol{W}' = \boldsymbol{W}+\boldsymbol{\Delta}$. Then,*

$$\begin{aligned}\mathcal{L}(\boldsymbol{W}') = \mathcal{L}(\boldsymbol{W}) &- \frac{1}{n}\sum_{i\in[n]}\langle(\boldsymbol{e}_{y_i}-\mathbb{S}(\boldsymbol{W}\boldsymbol{h}_i))\boldsymbol{h}_i^\top,\boldsymbol{\Delta}\rangle\\ &+\frac{1}{2n}\sum_{i\in[n]}\boldsymbol{h}_i^\top\boldsymbol{\Delta}^\top\left(\operatorname{diag}(\mathbb{S}(\boldsymbol{W}\boldsymbol{h}_i)) - \mathbb{S}(\boldsymbol{W}\boldsymbol{h}_i)\mathbb{S}(\boldsymbol{W}\boldsymbol{h}_i)^\top\right)\boldsymbol{\Delta}\,\boldsymbol{h}_i + o(\|\boldsymbol{\Delta}\|^3)\,.\end{aligned} \tag{10}$$

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

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

From Lemma 8,

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

Thus,

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

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

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

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

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

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

$$\ell_y(\boldsymbol{W}') = \ell_y(\boldsymbol{W}) - \langle (\boldsymbol{e}_y - \mathbb{S}(\boldsymbol{l}))\boldsymbol{h}^\top, \boldsymbol{\Delta}\rangle + \frac{1}{2}\boldsymbol{h}^\top\boldsymbol{\Delta}^\top \left(\operatorname{diag}(\mathbb{S}(\boldsymbol{l}')) - \mathbb{S}(\boldsymbol{l}')\mathbb{S}(\boldsymbol{l}')^\top\right) \boldsymbol{\Delta}\boldsymbol{h}, \quad (11)$$

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

We prove the relationships in (6), which are useful for unifying the analysis of entry-wise and Schatten norms.

Lemma 10 *For any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and any entry-wise or Schatten p-norm $\|\cdot\|$ with $p \geq 1$, it holds that*

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

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

$$|||\boldsymbol{A}|||_\infty = \sigma_1 \leq |||\boldsymbol{A}|||_p = \left(\sum_{i=1}^r \sigma_i^p\right)^{1/p} \leq \sum_{i=1}^r \sigma_i = |||\boldsymbol{A}|||_1\,. \quad (12)$$

It is also well-known that

$$|||\boldsymbol{A}|||_\infty = \max_{\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1} \boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{v} \geq \max_{i,j} |\boldsymbol{A}[i,j]| = \|\boldsymbol{A}\|_{\max} \quad (13)$$

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

Using this together with duality, it also holds that

$$|||\boldsymbol{A}|||_1 \leq \|\boldsymbol{A}\|_{\text{sum}}\,. \quad (14)$$

This follows from the following sequnece of inequalities

$$|||\boldsymbol{A}|||_1 = \max_{|||\boldsymbol{B}|||_\infty \leq 1} \langle \boldsymbol{A}, \boldsymbol{B}\rangle \leq \|\boldsymbol{A}\|_{\text{sum}} \cdot \max_{|||\boldsymbol{B}|||_\infty \leq 1} \|\boldsymbol{B}\|_{\max} \leq \|\boldsymbol{A}\|_{\text{sum}} \cdot \max_{|||\boldsymbol{B}|||_\infty \leq 1} |||\boldsymbol{B}|||_\infty \leq \|\boldsymbol{A}\|_{\text{sum}}\,, \quad (15)$$

where the first inequality follows from generalized Cauchy-Scwhartz and the second inequality by (13). ■

Lemma 11 is used in bounding the second order term in the Taylor expansion of $\mathcal{L}(\boldsymbol{W})$.

Lemma 11 *For any entry-wise or Schatten p-norm $\|\cdot\|$ with $p \geq 1$, any $\boldsymbol{s} \in \Delta^{k-1}$ in the k-dimensional simplex, any index $c \in [k]$, and $\boldsymbol{v} \in \mathbb{R}^k$, it holds that*

$$\boldsymbol{v}^\top \left(\operatorname{diag}(\boldsymbol{s}) - \boldsymbol{s}\boldsymbol{s}^\top\right) \boldsymbol{v} \leq 4(1 - s_c)\|\boldsymbol{v}\boldsymbol{v}^T\|.$$

Proof See main text. ■

Lemma 12 is used in the proof of Lemma 11.

Lemma 12 *For any $\boldsymbol{s} \in \Delta^{k-1}$ in the k-dimensional simplex and any index $c \in [k]$ it holds that*

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

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

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

Since this holds true, the lemma holds. ■

Lemma 13 *For any $\boldsymbol{s} \in \Delta^{k-1}$ in the k-dimensional simplex, any index $c \in [k]$, any $\boldsymbol{\Delta} \in \mathbb{R}^{k \times d}$, and any $\boldsymbol{h} \in \mathbb{R}^d$, it holds:*

$$\boldsymbol{h}^\top \boldsymbol{\Delta}^\top \left(\operatorname{diag}(\boldsymbol{s}) - \boldsymbol{s}\boldsymbol{s}^\top\right) \boldsymbol{\Delta}\boldsymbol{h} \leq 4B^2 \|\boldsymbol{\Delta}\|^2 (1 - s_c)\,.$$

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

$$\left|\left|\left|\boldsymbol{v}\boldsymbol{v}^\top\right|\right|\right| = \|\boldsymbol{v}\|_2^2 \leq |||\boldsymbol{\Delta}|||_\infty^2 \|\boldsymbol{h}\|_2^2 \leq |||\boldsymbol{\Delta}|||^2 \|\boldsymbol{h}\|_2^2 \leq B^2 |||\boldsymbol{\Delta}|||^2\,.$$

For any entry-wise p-norm, we have

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

This implies

$$\|\boldsymbol{v}\boldsymbol{v}^T\| = \|\boldsymbol{v}\|^2 = \|\boldsymbol{\Delta}\boldsymbol{h}\|^2 \leq \|\boldsymbol{\Delta}\|^2 \|\boldsymbol{h}\|_\star^2 \leq B^2 \|\boldsymbol{\Delta}\|^2.$$

Combine these results and apply Lemma 11, we obtain the desired. ■

The following lemma summarizes the properties of the softmax map that will be used in the proof of Lemma 26 and **??**.

Lemma 14 *For any $\boldsymbol{v}, \boldsymbol{v}', \boldsymbol{q}, \boldsymbol{q}' \in \mathbb{R}^k$ and $c \in [k]$, the following inequalities hold:*

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

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

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

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

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

Proof We prove each inequality:

(i) First, observe that

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

For any $i \in [k]$, we have $\frac{|e^{v_c'+v_i} - e^{v_c+v_i'}|}{e^{v_c+v_i'}} = |e^{v_c'-v_c+v_i-v_i'} - 1| \leq e^{|v_c'-v_c+v_i-v_i'|} - 1 \leq e^{2\|\boldsymbol{v}-\boldsymbol{v}'\|_\infty} - 1$. This implies $\sum_{i\in[k]} |e^{v_c'+v_i} - e^{v_c+v_i'}| \leq \left(e^{2\|\boldsymbol{v}-\boldsymbol{v}'\|_\infty} - 1\right) \sum_{i\in[k]} e^{v_c+v_i'}$, from which we obtain the desired inequality.

(ii) For the second inequality:

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

For any $j \in [k]$, $j \neq c$, and $i \in [k]$, we have $\frac{|e^{v_j'+v_i} - e^{v_j+v_i'}|}{e^{v_j+v_i'}} \leq e^{2\|\boldsymbol{v}-\boldsymbol{v}'\|_\infty} - 1$. This implies that $\sum_{j\in[k],j\neq c} \sum_{i\in[k]} |e^{v_j'+v_i} - e^{v_j+v_i'}| \leq (e^{2\|\boldsymbol{v}-\boldsymbol{v}'\|_\infty} - 1) \sum_{j\in[k],j\neq c} \sum_{i\in[k]} e^{v_j+v_i'}$, from which the result follows.

(iii) For the third inequality:

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

For any $i \in [k]$ and $j \in [k]$, $\frac{|e^{v_c'+v_i+q_c'+q_j} - e^{v_c+v_i'+q_c+q_j'}|}{e^{v_c+v_i'+q_c+q_j'}} = |e^{v_c'-v_c+v_i-v_i'+q_c'-q_c+q_j-q_j'} - 1| \le e^{|v_c'-v_c|+|v_i-v_i'|+|q_c'-q_c|+|q_j-q_j'|} - 1 \le e^{2(\|\boldsymbol{v}'-\boldsymbol{v}\|_\infty+\|\boldsymbol{q}'-\boldsymbol{q}\|_\infty)} - 1$. Then, rearranging and summing over i and j leads to the result.

(iv) For the fourth inequality:

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

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

(v) Finally, for the fifth inequality:

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

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

$$
\begin{aligned}
\frac{\left|e^{v_j'+q_i'+v_s+q_t}-e^{v_j+q_i+v_s'+q_t'}\right|}{e^{v_j+q_i+v_s'+q_t'}} &= \left|e^{v_j'-v_j+q_i'-q_i+v_s-v_s'+q_t-q_t'}-1\right| \\
&\leq e^{|v_j'-v_j|+|q_i'-q_i|+|v_s-v_s'|+|q_t-q_t'|}-1 \\
&\leq e^{2(\|\boldsymbol{v}'-\boldsymbol{v}\|_\infty+\|\boldsymbol{q}'-\boldsymbol{q}\|_\infty)}-1
\end{aligned}
$$

Then, rearranging and summing over j, i, s, and t leads to the result. ■

Appendix E. Lemmas on Loss and Proxy Function

Lemma 15 shows that $\mathcal{G}(\boldsymbol{W})$ upper and lower bound the dual norm of the loss gradient.

Lemma 15 ($\mathcal{G}(\boldsymbol{W})$ as proxy to the loss-gradient norm) *Under Assumption 4. For any $\boldsymbol{W}\in\mathbb{R}^{k\times d}$, it holds that*

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

Proof First, we prove the lower bound. By duality and direct application of (9)

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

Second, for the upper bound, it holds by triangle inequality and relationships (6) that

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

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

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

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

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

together with using the assumption $\|\boldsymbol{h}_i\| \le B$ yields the advertised upper bound. ■

Built upon Lemma 15, we obtain a simple bound on the loss difference at two points.

Lemma 16 *For any $\boldsymbol{W}, \boldsymbol{W}_0 \in \mathbb{R}^{k\times d}$, suppose that $\mathcal{L}(\boldsymbol{W})$ is convex, we have*

$$|\mathcal{L}(\boldsymbol{W}) - \mathcal{L}(\boldsymbol{W}_0)| \le 2B\|\boldsymbol{W} - \boldsymbol{W}_0\|.$$

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

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

where the last inequality is by Lemma 15. Similarly, we can also show that $\mathcal{L}(\boldsymbol{W}) - \mathcal{L}(\boldsymbol{W}_0) \le 2B\|\boldsymbol{W}_0 - \boldsymbol{W}\|$. ■

Lemma 17 shows the close relationships between $\mathcal{G}(\boldsymbol{W})$ and $\mathcal{L}(\boldsymbol{W})$. The proxy $\mathcal{G}(\boldsymbol{W})$ not only lower bounds $\mathcal{L}(\boldsymbol{W})$, but also upper bounds $\mathcal{L}(\boldsymbol{W})$ up to a factor depending on $\mathcal{L}(\boldsymbol{W})$. Moreover, the rate of convergence $\frac{\mathcal{G}(\boldsymbol{W})}{\mathcal{L}(\boldsymbol{W})}$ depends on the rate of decrease in the loss.

Lemma 17 ($\mathcal{G}(\boldsymbol{W})$ as proxy to the loss) *Let $\boldsymbol{W} \in \mathbb{R}^{k\times d}$, we have*

(i) $1 \ge \frac{\mathcal{G}(\boldsymbol{W})}{\mathcal{L}(\boldsymbol{W})} \ge 1 - \frac{n\mathcal{L}(\boldsymbol{W})}{2}$

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

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

The lower bound can be proved using the exact same arguments in the proof of Zhang et al. [65, Lemma C.7] for the binary case. For completeness, we provide an alternative elementary proof. It suffices to prove for $n = 1$ that for $s \in (0, 1)$:

$$1 - s \ge \log(1/s) - \frac{1}{2}\log^2(1/s). \tag{16}$$

The general case follows by summing over $s = s_i$ and using $\sum_{i\in[n]}\log^2(1/s_i) \le \left(\sum_{i\in[n]}\log(1/s_i)\right)^2$ since $\log(1/s_i) > 0$. For (16), let $x = \log(1/s) > 0$. The inequality becomes $e^{-x} \le 1 - x + x^2/2$, which holds for $x > 0$ by the second-order Taylor expansion of e^{-x} around 0.

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

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

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

Lemma 18 shows that the data becomes separable when the loss is small. It is used in deriving the lower bound on the un-normalized margin.

Lemma 18 (Low $\mathcal{L}(\boldsymbol{W})$ implies separability) *Suppose that there exists $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ such that $\mathcal{L}(\boldsymbol{W}) \leq \frac{\log 2}{n}$, then we have*

$$(\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^T \boldsymbol{W}\boldsymbol{h}_i \geq 0, \quad \textit{for all } i \in [n] \textit{ and for all } c \in [k] \textit{ such that } c \neq y_i. \tag{17}$$

Proof We rewrite the loss into the form:

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

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

$$\log(1+\sum_{c\neq y_i} e^{-(\boldsymbol{\ell}_i[y_i]-\boldsymbol{\ell}_i[c])}) \leq n\mathcal{L}(\boldsymbol{W}) \leq \log(2).$$

This implies:

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

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

Lemma 19 shows that the ratio of $\mathcal{G}(\boldsymbol{W})$ at two points can be bounded by exponentiating the max-norm of their differences. It is used in handling the second order term in the Taylor expansion of the loss.

Lemma 19 (Ratio of $\mathcal{G}(\boldsymbol{W})$) *For any $\psi \in [0,1]$, we have the following:*

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

Proof Note that the second inequality is by relationships (6). Here, we only prove the first inequality. By the definition of $\mathcal{G}(\boldsymbol{W})$, we have:

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

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

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

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

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

Then, we rearrange and sum over $i \in [n]$ to obtain: $\sum_{i\in[n]}\big(1-\mathbb{S}_{y_i}((\boldsymbol{W}-\psi\eta\boldsymbol{\Delta})\boldsymbol{h}_i)\big) \leq e^{2B\eta\psi\|\boldsymbol{\Delta}\|_{\max}}\sum_{i\in[n]}\big(1-\mathbb{S}_{y_i}(\boldsymbol{W}\boldsymbol{h}_i)\big)$, from which the desired inequality follows. The second inequality in the lemma statement follows from the relationship (6). ■

Appendix F. Implicit Bias of Normalized Steepest Descent

Proof Overview We consider a decay learning rate schedule of the form $\eta_t = \Theta(\frac{1}{t^a})$ where $a \in (0, 1]$. The first step is to show that the **loss monotonically decreases** after certain time and the rate depends on $\mathcal{G}(\boldsymbol{W})$. To obtain this, we apply Lemma 15 and Lemma 11 to upper bound the first-order and second-order terms in the Taylor expansion of the loss (18), respectively. Next, we use the decrease in loss to derive a lower bound on the unnormalized margin which involves the ratio $\frac{\mathcal{G}(\boldsymbol{W})}{\mathcal{L}(\boldsymbol{W})}$. A crucial step involved is to find a time $\bar{t}_2$ such that separability (29) holds for all $t \geq \bar{t}_2$, and the existence of $\bar{t}_2$ is guaranteed by loss monotonicity such that the condition $\mathcal{L}(\boldsymbol{W}_t) \leq \frac{\log 2}{n}$ will be satisfied for sufficitently large t's.

Then, we argue that the ratio $\frac{\mathcal{G}(\boldsymbol{W}_t)}{\mathcal{L}(\boldsymbol{W}_t)}$ converges to 1 exponentially fast (recalling that $1 \geq \frac{\mathcal{G}(\boldsymbol{W}_t)}{\mathcal{L}(\boldsymbol{W}_t)} \geq 1 - \frac{n\mathcal{L}(\boldsymbol{W}_t)}{2}$) by showing the loss $\mathcal{L}(\boldsymbol{W}_t)$ decreases exponentially fast. We first choose a time t_1 after t_0 (recall that t_0 is the time that satisfies Assumption 3) such that $\mathcal{L}(\boldsymbol{W}_{t+1}) \leq \mathcal{L}(\boldsymbol{W}_t) - \frac{\eta_t\gamma}{2}\mathcal{G}(\boldsymbol{W}_t)$ for all $t \geq t_1$. Next, we lower bound $G(\boldsymbol{W}_t)$ using $\mathcal{L}(\boldsymbol{W}_t)$. By Lemma 17, there are two sufficient conditions (namely, $\mathcal{L}(\boldsymbol{W}_t) \leq \frac{\log 2}{n} =: \tilde{\mathcal{L}}$ or $\boldsymbol{G}(\boldsymbol{W}_t) \leq \frac{1}{2n}$) that guarantee $\mathcal{L}(\boldsymbol{W}_t) \leq 2\mathcal{G}(\boldsymbol{W}_t)$. We 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 have $\mathcal{G}(\boldsymbol{W}_{t^*}) \leq \frac{\tilde{\mathcal{L}}}{2} \leq \frac{1}{2n}$. This not only guarantees that $\mathcal{L}(\boldsymbol{W}_{t^*}) \leq 2\mathcal{G}(\boldsymbol{W}_{t^*})$ at time t^*, but also

(crucially due to monotonicity) implies that $\mathcal{L}(\boldsymbol{W}_t) \le \mathcal{L}(\boldsymbol{W}_{t^*}) \le 2\mathcal{G}(\boldsymbol{W}_{t^*}) \le \frac{\log 2}{n}$ for all $t \ge t_2$. Thus, we observe that the other sufficient condition $\mathcal{L}(\boldsymbol{W}_t) \le \frac{\log 2}{n}$ is satisfied, from which we conclude that $\mathcal{L}(\boldsymbol{W}_t) \le 2\mathcal{G}(\boldsymbol{W}_t)$ for all $t \ge t_2$. We remark that the choice of t_2 depends on $\mathcal{L}(\boldsymbol{W}_{t_1})$ (whose magnitude is bounded using Lemma 16), and t_2 can be used as $\bar{t}_2$ above. To recap, t_1 is the time (after t_0) after which the successive loss decrease is lower bounded by the product $\eta_t \gamma \mathcal{G}(\boldsymbol{W}_t)$; t_2 (after t_1) is the time after which $\mathcal{L}(\boldsymbol{W}_t) \le \frac{\log 2}{n}$ (thus, both $\mathcal{L}(\boldsymbol{W}_t) \le 2\mathcal{G}(\boldsymbol{W}_t)$ and separability condition (29) hold for all $t \ge t_2$).

In this following, we break the proof of implicit bias of NSD into several parts following previous arguments. Lemma 20 shows the descent properties of NSD. It is used in Lemma 21 to lower bound the un-normalized margin, and in the proof of Theorem 23 to show the convergence of $\frac{\mathcal{G}(\boldsymbol{W}_t)}{\mathcal{L}(\boldsymbol{W}_t)}$.

Lemma 20 (NSD Descent) *Under the same setting as Theorem 23, it holds for all $t \ge 0$,*

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

where α_{s_1} is some constant that depends on B and γ.

Proof By Lemma 9, we let $\boldsymbol{W}' = \boldsymbol{W}_{t+1}$, $\boldsymbol{W} = \boldsymbol{W}_t$, $\tilde{\boldsymbol{\Delta}}_t = \boldsymbol{W}_{t+1} - \boldsymbol{W}_t$, and define $\boldsymbol{W}_{t,t+1,\zeta} := \boldsymbol{W}_t + \zeta(\boldsymbol{W}_{t+1} - \boldsymbol{W}_t)$. We choose ζ^* such that $\boldsymbol{W}_{t,t+1,\zeta^*}$ satisfies (11), we have:

$$\mathcal{L}(\boldsymbol{W}_{t+1}) = \mathcal{L}(\boldsymbol{W}_t) + \underbrace{\langle \nabla\mathcal{L}(\boldsymbol{W}_t), \tilde{\boldsymbol{\Delta}}_t \rangle}_{\spadesuit_t} + \frac{1}{2n}\sum_{i\in[n]} \underbrace{\boldsymbol{h}_i^\top \tilde{\boldsymbol{\Delta}}_t^\top \left(\operatorname{diag}(\mathbb{S}(\boldsymbol{W}_{t,t+1,\gamma}\boldsymbol{h}_i)) - \mathbb{S}(\boldsymbol{W}_{t,t+1,\zeta^*}\boldsymbol{h}_i)\mathbb{S}(\boldsymbol{W}_{t,t+1,\zeta^*}\boldsymbol{h}_i)^\top\right)\tilde{\boldsymbol{\Delta}}_t \boldsymbol{h}_i}_{\clubsuit_t}. \tag{18}$$

For the $\spadesuit_t$ term, we have by Lemma 15:

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

For the $\clubsuit_t$ term, we let $\boldsymbol{v} = \tilde{\boldsymbol{\Delta}}_t\boldsymbol{h}_i$ and $\boldsymbol{s} = \mathbb{S}(\boldsymbol{W}_{t,t+1,\zeta^*}\boldsymbol{h}_i)$, and apply Lemma 13 to obtain

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

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

$$\begin{aligned}
\mathcal{L}(\boldsymbol{W}_{t+1}) &\le \mathcal{L}(\boldsymbol{W}_t) - \gamma\eta_t\mathcal{G}(\boldsymbol{W}_t) + 2\eta_t^2B^2\frac{1}{n}\sum_{i\in[n]}(1 - \mathbb{S}_{y_i}(\boldsymbol{W}_{t,t+1,\zeta^*}\boldsymbol{h}_i)) \\
&= \mathcal{L}(\boldsymbol{W}_t) - \gamma\eta_t\mathcal{G}(\boldsymbol{W}_t) + 2\eta_t^2B^2\mathcal{G}(\boldsymbol{W}_{t,t+1,\zeta^*}) \\
&\le \mathcal{L}(\boldsymbol{W}_t) - \gamma\eta_t\mathcal{G}(\boldsymbol{W}_t) + 2\eta_t^2B^2\sup_{\zeta\in[0,1]}\mathcal{G}(\boldsymbol{W}_{t,t+1,\zeta}) \\
&= \mathcal{L}(\boldsymbol{W}_t) - \gamma\eta_t\mathcal{G}(\boldsymbol{W}_t) + 2\eta_t^2B^2\mathcal{G}(\boldsymbol{W}_t)\sup_{\zeta\in[0,1]}\frac{\mathcal{G}(\boldsymbol{W}_t + \zeta\tilde{\boldsymbol{\Delta}}_t)}{\mathcal{G}(\boldsymbol{W}_t)} \\
&\overset{(a)}{\le} \mathcal{L}(\boldsymbol{W}_t) - \gamma\eta_t\mathcal{G}(\boldsymbol{W}_t) + 2\eta_t^2B^2\mathcal{G}(\boldsymbol{W}_t)\sup_{\zeta\in[0,1]}e^{2B\zeta\|\tilde{\boldsymbol{\Delta}}_t\|} \\
&\overset{(b)}{\le} \mathcal{L}(\boldsymbol{W}_t) - \gamma\eta_t\mathcal{G}(\boldsymbol{W}_t) + 2\eta_t^2B^2e^{2B\eta_0}\mathcal{G}(\boldsymbol{W}_t),
\end{aligned} \tag{19}$$

where (a) is by Lemma 19 and (b) is by $\|\tilde{\boldsymbol{\Delta}}_t\| \leq \eta_t$. Letting $\alpha_{s_1} = \frac{2B^2 e^{2B\eta_0}}{\gamma}$, Eq. (19) simplifies to:

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

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

For a decaying learning rate schedule, Lemma 20 implies that the loss monotonically decreases after a certain time. Thus, we know that the assumption of Lemma 21 can be satisfied. In the proof of Theorem 23, we will specify a concrete form of $\tilde{t}$ in Lemma 21.

Lemma 21 (NSD Unnormalized Margin) *Suppose that there exist $\tilde{t}$ such that $\mathcal{L}(\boldsymbol{W}_t) \leq \frac{\log 2}{n}$ for all $t > \tilde{t}$, then we have*

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

where α_{s_2} is some constant that depends on B.

Proof We let $\alpha_{s_2} = 2Be^{2B\eta_0}$, then from (19), we have for $t > \tilde{t}$:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{W}_{t+1}) &\leq \mathcal{L}(\boldsymbol{W}_t) - \gamma\eta_t\mathcal{G}(\boldsymbol{W}_t) + \alpha_{s_2}\eta_t^2\mathcal{G}(\boldsymbol{W}_t) \\
&= \mathcal{L}(\boldsymbol{W}_t)\big(1 - \gamma\eta_t \frac{\mathcal{G}(\boldsymbol{W}_t)}{\mathcal{L}(\boldsymbol{W}_t)} + \alpha_{s_2}\eta_t^2 \frac{\mathcal{G}(\boldsymbol{W}_t)}{\mathcal{L}(\boldsymbol{W}_t)}\big) \\
&\leq \mathcal{L}(\boldsymbol{W}_t)\exp\big(-\gamma\eta_t \frac{\mathcal{G}(\boldsymbol{W}_t)}{\mathcal{L}(\boldsymbol{W}_t)} + \alpha_{s_2}\eta_t^2 \frac{\mathcal{G}(\boldsymbol{W}_t)}{\mathcal{L}(\boldsymbol{W}_t)}\big) \\
&\leq \mathcal{L}(\boldsymbol{W}_{\tilde{t}})\exp\big(-\gamma \sum_{s=\tilde{t}}^{t} \eta_s \frac{\mathcal{G}(\boldsymbol{W}_s)}{\mathcal{L}(\boldsymbol{W}_s)} + \alpha_{s_2} \sum_{s=\tilde{t}}^{t} \eta_s^2\big). \\
&\leq \frac{\log 2}{n}\exp\big(-\gamma \sum_{s=\tilde{t}}^{t} \eta_s \frac{\mathcal{G}(\boldsymbol{W}_s)}{\mathcal{L}(\boldsymbol{W}_s)} + \alpha_{s_2} \sum_{s=\tilde{t}}^{t} \eta_s^2\big), \qquad (20)
\end{aligned}$$

where the penultimate inequality uses Lemma 17, and the last inequality uses the assumption that $\mathcal{L}(\boldsymbol{W}_t) \leq \frac{\log 2}{n}$ for all $t \geq \tilde{t}$. Then, we have for all $t > \tilde{t}$:

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

(a) is by the following: the assumption $\mathcal{L}(\boldsymbol{W}_t) \leq \frac{\log 2}{n}$ implies that $\min_{c \neq y_i} (\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^T \boldsymbol{W}_t \boldsymbol{h}_i \geq 0$ for all $i \in [n]$ by Lemma 18. 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 $z = \min_{c \neq y_i} (\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^T \boldsymbol{W}_t \boldsymbol{h}_i$ to obtain the desired inequality; and (b) is by (20). Finally, taking log on both sides leads to the result. ■

Next Lemma upper bounds the p-norm of NSD's iterates using learning rates. It is used in the proof of Theorem 23.

Lemma 22 (NSD $\|\boldsymbol{W}_t\|$) *For NSD, we have for any $t > 0$ that*

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

Proof By the NSD update rule (2), we have

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_0 - \sum_{s=0}^{t} \eta_s \boldsymbol{\Delta}_s.$$

This leads to $\|\boldsymbol{W}_t\| \leq \|\boldsymbol{W}_0\| + \sum_{s=0}^{t-1} \eta_s$ given $\boldsymbol{\Delta}_s \leq 1$ for all $s \geq 0$. ■

The main step in the proof of Theorem 23 is to determine the time that satisfies the assumption in Lemma 21 and show the convergence of $\frac{\mathcal{G}(\boldsymbol{W}_t)}{\mathcal{L}(\boldsymbol{W}_t)}$. Then, Lemma 21 and Lemma 22 will be combined to obtain the final result.

Theorem 23 *Suppose that Assumption 1, 2, and 4 hold, then there exists $t_{s_2} = t_{s_2}(n, \gamma, B, \boldsymbol{W}_0)$ such that NSD achieves the following for all $t > t_{s_2}$*

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

Proof **Determination of t_{s_1}.** In Lemma 20 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 $t \geq t_{s_1}$

$$\mathcal{L}(\boldsymbol{W}_{t+1}) \leq \mathcal{L}(\boldsymbol{W}_t) - \frac{\eta_t \gamma}{2} \mathcal{G}(\boldsymbol{W}_t). \tag{21}$$

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

Determination of t_{s_2}. By Lemma 16, we can bound $\mathcal{L}(\boldsymbol{W}_{t_{s_1}})$ as follows

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

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

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

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

Convergence of $\frac{\mathcal{G}(\boldsymbol{W}_t)}{\mathcal{L}(\boldsymbol{W}_t)}$ Given $\mathcal{G}(\boldsymbol{W}_{t^*}) \leq \frac{\tilde{\mathcal{L}}}{2} \leq \frac{1}{2n}$, we obtain that $\mathcal{L}(\boldsymbol{W}_t) \leq \mathcal{L}(\boldsymbol{W}_{t^*}) \leq 2\mathcal{G}(\boldsymbol{W}_{t^*}) \leq \tilde{L}$ for all $t \geq t_{s_2}$, where the first and second inequalities are due to monotonicity in the risk and Lemma 17, respectively. Thus, the other sufficient condition $\mathcal{L}(\boldsymbol{W}_t) \leq \frac{\log 2}{n}$ in Lemma 17 is satisfied, from which we conclude that $\mathcal{L}(\boldsymbol{W}_t) \leq 2\mathcal{G}(\boldsymbol{W}_t)$ for all $t \geq t_{s_2}$. Substituting this into (21), we obtain for all $t > t_{s_2}$

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

Then, by Lemma 17, we obtain

$$\frac{\mathcal{G}(\boldsymbol{W}_t)}{\mathcal{L}(\boldsymbol{W}_t)} \geq 1 - \frac{n\mathcal{L}(\boldsymbol{W}_t)}{2} \geq 1 - \frac{n\tilde{\mathcal{L}}e^{-\frac{\gamma}{4}\sum_{s=t_{s_2}}^{t-1}\eta_s}}{2} \geq 1 - e^{-\frac{\gamma}{4}\sum_{s=t_{s_2}}^{t-1}\eta_s}. \tag{22}$$

Margin Convergence Finally, we combine Lemma 21, Lemma 22, and (22) to obtain

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

■

Next, we explicitly upper bound t_{s_2} in Theorem 23 to derive the margin convergence rates of NSD.

Corollary 24 *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 23, then we have for SignGD*

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

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

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

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

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

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. [65, Corollary 4.7]. ■

Appendix G. Implicit Bias of Normalized Momentum Steepest Descent

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

Lemma 25 *Consider the following $\boldsymbol{W}^\dagger := \boldsymbol{W} - \eta\boldsymbol{\Delta}$, where $\boldsymbol{\Delta} \in \mathbb{R}^{k\times d}$ is defined in (3). Let $\boldsymbol{M} \in \mathbb{R}^{k\times d}$ be any matrix. It holds:*

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

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

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

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

where (a) is by Cauchy Schwarz inequality and $\langle \boldsymbol{M}, \boldsymbol{\Delta}\rangle = \|\boldsymbol{M}\|_*$, (b) is by $\|\boldsymbol{\Delta}\| \le 1$, (c) is via Lemma 10, (d) is by reverse triangle inequality, and (e) is via Lemma 15. ■

The following Lemma bounds the entries of the momentum ($\boldsymbol{M}_t$) of NMD in terms of the product of η_t with the sume of $\mathcal{G}_c(\boldsymbol{W}_t)$ and $\mathcal{Q}_c(\boldsymbol{W}_t)$.

Lemma 26 *Suppose that Ass. 1, 2, 3, and 4 hold. Let $c \in [k]$ and $j \in [d]$. There exists time t_0 such that for all $t \geq t_0$:*

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

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

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

$$\begin{aligned}|\mathbf{M}_t[c,j] - (1-\beta_1^{t+1})\nabla\mathcal{L}(\boldsymbol{W}_t)[c,j]| &= |\sum_{\tau=0}^{t}(1-\beta_1)\beta_1^\tau\big(\nabla\mathcal{L}(\boldsymbol{W}_{t-\tau})[c,j] - \nabla\mathcal{L}(\boldsymbol{W}_t)[c,j]\big)| \\ &\leq \sum_{\tau=0}^{t}(1-\beta_1)\beta_1^\tau \underbrace{|\nabla\mathcal{L}(\boldsymbol{W}_{t-\tau})[c,j] - \nabla\mathcal{L}(\boldsymbol{W}_t)[c,j]|}_{\clubsuit}.\end{aligned} \tag{23}$$

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

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

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

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

where (a) is by Lemma 14, (b) is by $\|\boldsymbol{h}_i\|_1 \leq B$ for all $i \in [n]$, (c) is by (2) and triangle inequality, and (d) is by $\|\boldsymbol{\Delta}_{t-s}\|_{\max} \leq \|\boldsymbol{\Delta}_{t-s}\| \leq 1$ (for any entry-wise or Schatten p-norm) and the definition of $\mathcal{G}(\boldsymbol{W}_t)$. For the second term, we obtain:

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

where (e) is by Lemma 14, and (f) is by the same approach taken for $\clubsuit_1$. Based on the upper bounds for $\clubsuit_1$ and $\clubsuit_2$, we obtain the following: $\clubsuit \leq 2B\big(e^{2\alpha B\sum_{s=1}^{\tau}\eta_{t-s}} - 1\big)(\mathcal{G}_c(\boldsymbol{W}_t) + \mathcal{Q}_c(\boldsymbol{W}_t))$. Then, we substitute this into (23) to obtain:

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

where (g) is by the Assumption 3. ■

Lemma 27 *Let $\boldsymbol{\Omega}_t = \boldsymbol{M}_t - \nabla\mathcal{L}(\boldsymbol{W}_t)$, where $\boldsymbol{M}_t$ is defined in (4). Then, it holds*

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

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

Proof For simplicity, we drop the subscripts t. Denote $\mathcal{T}_c(\boldsymbol{W}) := \mathcal{G}_c(\boldsymbol{W}) + \mathcal{Q}_c(\boldsymbol{W})$. Then, by Lemma 26, we have for any $c \in [k]$ and $j \in [d]$:

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

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 $\boldsymbol{\Omega} := \boldsymbol{M} - \nabla\mathcal{L}(\boldsymbol{W})$, then we have

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

This implies the following:

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

where in the last inequality we have used Lemma 15. ■

Lemma 28 *Suppose that there exist $\tilde{t}$ such that $\mathcal{L}(\boldsymbol{W}_t) \leq \frac{\log 2}{n}$ for all $t > \tilde{t}$, then we have*

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

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

Proof We follow a similar approach as Lemma 20 to show the descent of NMD. Specifically, we apply Lemma 25 to bound the first-order term. For the Hessian term, we apply Lemma 13 and Lemma 19 similar to NSD. Then, we can obtain the following

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

where (a) is by Lemma 25. In (b), we have defined $a_1 := 4B$ and $a_2 = (4\alpha_M + 2B^2 e^{2B\eta_0})d$. In (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 proof follows the same steps as Lemma 21. ■

Theorem 29 *Suppose that Ass. 1, 2, 3, and 4 hold. Set learning rate $\eta_t = \Theta(\frac{1}{t^{1/2}})$. The margin gap of NMD's iterates satisfy*

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

Proof Given the updates of NMD are normalized (i.e., $\|\boldsymbol{\Delta}\| \leq 1$), we can obtain the following via Lemma 28:

$$\gamma - \frac{\min_{i\in[n],c\neq y_i}(\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^T \boldsymbol{W}_t \boldsymbol{h}_i}{\|\boldsymbol{W}_t\|} \leq \frac{\gamma(\|W_0\| + \sum_{s=0}^{t_2-1}\eta_s + \sum_{s=t_2}^{t-1}\eta_s e^{\frac{\gamma}{4}\sum_{\tau=t_2}^{s-1}\eta_\tau}) + a_2 d\sum_{s=t_2}^{t-1}\eta_s^2 + Q}{\|\boldsymbol{W}_0\| + \sum_{s=0}^{t-1}\eta_s}$$
$$\leq O(\frac{\sum_{s=t_2}^{t-1}\eta_s e^{-\frac{\gamma}{4}\sum_{\tau=t_2}^{s-1}\eta_\tau} + \sum_{s=0}^{t_2-1}\eta_s + d\sum_{s=t_2}^{t-1}\eta_s^2}{\sum_{s=0}^{t-1}\eta_s}).$$

Then, we follow the same approach as Corollary 24 for a decreasing learning rate of the form $\eta_t = \Theta(\frac{1}{t^a})$. Specifically, we have $t_1 = \Theta(d^{1/a})$ and $t_2 \leq C_1 n^{\frac{1}{1-a}} t_1 + C_2 n^{\frac{1}{1-a}} L(\boldsymbol{W}_0)^{\frac{1}{1-a}}$. This leads to

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

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

Appendix H. Other multiclass loss functions

H.1. Exponential Loss

The multiclass exponential loss is given as

$$\mathcal{L}_{\exp}(\boldsymbol{W}) := \frac{1}{n}\sum_{i\in[n]}\sum_{c\neq y_i}\exp\left(-(\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^\top \boldsymbol{W}\boldsymbol{h}_i\right).$$

The gradient of $\mathcal{L}_{\exp}(\boldsymbol{W})$ is

$$\nabla\mathcal{L}_{\exp}(\boldsymbol{W}) = \frac{1}{n}\sum_{i\in[n]}\sum_{c\neq y_i} -\exp(-(\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^T \boldsymbol{W}\boldsymbol{h}_i)(\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)\boldsymbol{h}_i^T.$$

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

$$\langle \boldsymbol{A}, -\nabla\mathcal{L}_{\exp}(\boldsymbol{W})\rangle = \frac{1}{n}\sum\nolimits_{i\in[n]}\sum\nolimits_{c\neq y_i}\exp\left(-(\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^\top \boldsymbol{W}\boldsymbol{h}_i\right)\cdot(\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^\top \boldsymbol{A}\boldsymbol{h}_i.$$

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

$$\mathcal{G}_{\exp}(\boldsymbol{W}) = \frac{1}{n}\sum_{i\in[n]}\sum_{c\neq y_i}\exp\left(-(\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^\top \boldsymbol{W}\boldsymbol{h}_i\right),$$

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

H.2. PairLogLoss

The PairLogLoss loss [60] is given as

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

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

$$\begin{aligned}\mathcal{L}_{\text{pll}}(\boldsymbol{W}+\boldsymbol{\Delta}) = \mathcal{L}(\boldsymbol{W}) &+ \frac{1}{n}\sum_{i\in[n]}\sum_{c\neq y_i} f'\left((\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)^\top \boldsymbol{W}\boldsymbol{h}_i\right)\cdot(\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)^\top\boldsymbol{\Delta}\boldsymbol{h}_i \\ &+ \frac{1}{n}\sum_{i\in[n]}\sum_{c\neq y_i} f''\left((\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)^\top \boldsymbol{W}\boldsymbol{h}_i\right)\cdot \boldsymbol{h}_i^\top\boldsymbol{\Delta}^\top(\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)(\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)^\top\boldsymbol{\Delta}\boldsymbol{h}_i + o\left(\|\boldsymbol{\Delta}\|^3\right).\end{aligned} \tag{24}$$

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

$$\begin{aligned}\nabla\mathcal{L}_{\text{pll}}(\boldsymbol{W}) &= \frac{1}{n}\sum_{i\in[n]}\sum_{c\neq y_i} f'\left((\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)^\top \boldsymbol{W}\boldsymbol{h}_i\right)\cdot(\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)\,\boldsymbol{h}_i^\top \\ &= \frac{1}{n}\sum_{i\in[n]}\sum_{c\neq y_i}\frac{-\exp\left(-(\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)^\top \boldsymbol{W}\boldsymbol{h}_i\right)}{1+\exp\left(-(\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)^\top \boldsymbol{W}\boldsymbol{h}_i\right)}(\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)\,\boldsymbol{h}_i^\top\end{aligned} \tag{25}$$

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

$$\langle \boldsymbol{A}, -\nabla\mathcal{L}_{\text{pll}}(\boldsymbol{W})\rangle = \frac{1}{n}\sum_{i\in[n]}\sum_{c\neq y_i}|f'\left((\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)^\top \boldsymbol{W}\boldsymbol{h}_i\right)|\cdot(\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)^\top\boldsymbol{A}\boldsymbol{h}_i\,. \tag{26}$$

This motivates us to define

$$\mathcal{G}_{\text{pll}}(\boldsymbol{W}) = \frac{1}{n}\sum_{i\in[n]}\sum_{c\neq y_i}\left|f'\left(-(\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)^\top \boldsymbol{W}\boldsymbol{h}_i\right)\right| = \frac{1}{n}\sum_{i\in[n]}\sum_{c\neq y_i}\frac{\exp\left(-(\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)^\top \boldsymbol{W}\boldsymbol{h}_i\right)}{1+\exp\left(-(\boldsymbol{e}_{y_i}-\boldsymbol{e}_c)^\top \boldsymbol{W}\boldsymbol{h}_i\right)} \tag{27}$$

Lemma 30 (Analogue of Lemma 15 for PLL) *For any $\boldsymbol{W}$, the PairLogLoss (PLL) satisfies:*

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

Proof The lower bound follows immediately from (26) and expressing $\|\nabla\mathcal{L}_{\text{pll}}(\boldsymbol{W})\|_* = \max_{\|\boldsymbol{A}\|\leq 1}\langle \boldsymbol{A}, -\nabla\mathcal{L}_{\text{pll}}(\boldsymbol{W})\rangle$. The lower bound follows from triangle inequality applied to (25):

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

and use the relationships in (6), i.e. $\|\nabla\mathcal{L}_{\text{pll}}(\boldsymbol{W})\| \leq \|\nabla\mathcal{L}_{\text{pll}}(\boldsymbol{W})\|_{\text{sum}}$ for any entry-wise or Schatten p-norm with $p\geq 1$. ■

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

$$\begin{aligned}\boldsymbol{h}_i^\top \boldsymbol{\Delta}^\top (\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)(\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^\top \boldsymbol{\Delta} \boldsymbol{h}_i &= \langle (\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)(\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^\top, \boldsymbol{\Delta}\boldsymbol{h}_i \boldsymbol{h}_i^\top \boldsymbol{\Delta}^T \rangle \\ &\leq \left\| (\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)(\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^\top \right\|_{\text{sum}} \left\| \boldsymbol{\Delta}\boldsymbol{h}_i \boldsymbol{h}_i^\top \boldsymbol{\Delta}^T \right\|_{\max} \\ &\leq \|\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)\|_1^2 \cdot (\|\boldsymbol{\Delta}\boldsymbol{h}_i\|_\infty)^2 \\ &\leq 4 \cdot (\|\boldsymbol{\Delta}\|_{\max})^2 \cdot \|\boldsymbol{h}_i\|_1^2 \leq 4B^2 (\|\boldsymbol{\Delta}\|_{\max})^2 \\ &\leq 4B^2 \|\boldsymbol{\Delta}\|^2.\end{aligned}$$

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

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

Finally, we verify PLL satisfies Lemma 17.

Lemma 31 (Analogue of Lemma 17 for PLL) *Let $\boldsymbol{W} \in \mathbb{R}^{k \times d}$, we have*

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

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

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

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

$$\frac{x}{1+x} \geq \log(1+x) - \frac{1}{2}\log^2(1+x). \tag{28}$$

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

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

where the last line used $\log(1+x_{ic}) \geq 0$. For (16), let $a = \log(1+x) > 0$. The inequality becomes $e^{-a} \leq 1 - a + a^2/2$, which holds for $a > 0$ by the second-order Taylor expansion of e^{-a} around 0.

(ii) Denote $\mathcal{L} \coloneqq \mathcal{L}_{pll}$ and $\mathcal{G} \coloneqq \mathcal{G}_{pll}$. 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 follows from (i). For the second part, denote $l_{ic} \coloneqq (\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^\top \boldsymbol{W}\boldsymbol{h}_i, i \in [n], c \neq y_i$. For $\mathcal{L} \leq 2\mathcal{G}$ 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 when $l_{ic} \geq -1.366$, which is clearly satisfied given the assumption $\mathcal{G} \leq \frac{1}{2n}$ implying $l_{ic} \geq 0$. ■

Lemma 32 (Analogue of Lemma 19 for PLL) *For any $\psi \in [0,1]$, we have the following:*

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

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

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

Denote $x_{ic}^{\boldsymbol{W}} := (\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^T \boldsymbol{W}\boldsymbol{h}_i$ and $x_{ic}^{\boldsymbol{W}'} := (\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^T (\boldsymbol{W} - \psi\eta\boldsymbol{\Delta})\boldsymbol{h}_i$, then we have for $i \in [n]$, $c \neq y_i$

$$\begin{aligned}
\frac{f'(x_{ic}^{\boldsymbol{W}'})}{f'(x_{ic}^{\boldsymbol{W}})} = \Big| \frac{f'(x_{ic}^{\boldsymbol{W}'})}{f'(x_{ic}^{\boldsymbol{W}})} \Big| \leq e^{|x_{ic}^{\boldsymbol{W}} - x_{ic}^{\boldsymbol{W}'}|} + 2 &= e^{\psi\eta|(\boldsymbol{e}_c - \boldsymbol{e}_{y_i})^T \boldsymbol{\Delta}\boldsymbol{h}_i|} + 2 = e^{\psi\eta|\langle \boldsymbol{\Delta}, (\boldsymbol{e}_c - \boldsymbol{e}_{y_i})\boldsymbol{h}_i^T \rangle|} + 2 \\
&\leq e^{\psi\eta\|\boldsymbol{\Delta}\|_{\max} \|(\boldsymbol{e}_c - \boldsymbol{e}_{y_i})\boldsymbol{h}_i^T\|_{\text{sum}}} + 2 \\
&= e^{\psi\eta\|\boldsymbol{\Delta}\|_{\max} \|\boldsymbol{e}_c - \boldsymbol{e}_{y_i}\|_{\text{sum}} \|\boldsymbol{h}_i\|_{\text{sum}}} + 2 \\
&\leq e^{2B\psi\eta\|\boldsymbol{\Delta}\|_{\max}} + 2.
\end{aligned}$$

This leads to $\sum_{i\in[n]} \sum_{c\neq y_i} f'(x_{ic}^{\boldsymbol{W}'}) \leq (e^{2B\psi\|\boldsymbol{\Delta}\boldsymbol{W}\|_{\max}} + 2) \sum_{i\in[n]} \sum_{c\neq y_i} f'(x_{ic}^{\boldsymbol{W}})$. Rearrange and using the definition of $\mathcal{G}_{pll}(\boldsymbol{W})$ and relationships in (6), we obtain the desired. ■

Lemma 33 (Analogue of Lemma 18 for PLL) *Suppose that there exists $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ such that $\mathcal{L}_{pll}(\boldsymbol{W}) \leq \frac{\log 2}{n}$, then we have*

$$(\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^T \boldsymbol{W}\boldsymbol{h}_i \geq 0, \quad \textit{for all } i \in [n] \textit{ and for all } c \in [k] \textit{ such that } c \neq y_i. \tag{29}$$

Proof Denote $x_{ic} = (\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^T \boldsymbol{W}\boldsymbol{h}_i$. Then, by the assumption, we have for any $i \in [n], c \neq y_i$

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

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

Lemma 34 (Analogue of Lemma 16 for PLL) *For any $\boldsymbol{W}, \boldsymbol{W}_0 \in \mathbb{R}^{k\times d}$, suppose that $\mathcal{L}(\boldsymbol{W})$ is convex, we have*

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

Proof This lemma is a direct consequence of Lemma 30 and can be proved in the same way as Lemma 16. ■

Thus, we have proved all the Lemmas for $\mathcal{G}_{pll}(\boldsymbol{W})$ and its relationships to $\mathcal{L}_{pll}(\boldsymbol{W})$ in analogous to those in section E. The proof of NSD ((2)) with PairLogLoss follow the same steps as with cross-entropy loss given in section F.