OPT2025: 17th Annual Workshop on Optimization for Machine Learning

Implicit Bias of Polyak and Line-Search Step Sizes on Linear Classification with Separable Data

Chen Fan FANCHEN3@OUTLOOK.COM
University of British Columbia, Canada
Reza Babanezhad BABANEZHAD@GMAIL.COM
Samsung AI, Montreal
Christos Thrampoulidis CTHRAMPO@ECE.UBC.CA
University of British Columbia, Canada
Mark Schmidt SCHMIDTM@CS.UBC.CA
Canada CIFAR AI Chair (Amii)
University of British Columbia, Canada
Sharan Vaswani VASWANI.SHARAN@GMAIL.COM
Simon Fraser University, Canada

Abstract

Recent works have shown that Polyak and line-search step sizes are good for training deep neural networks. However, a theoretical understanding of their generalization performances is lacking. For overparameterized models, multiple solutions can generalize differently to unseen data despite all obtaining zero training loss. Given this, a natural question is whether an algorithm inherently prefers (without explicit regularization) certain simple solutions over others upon convergence-a phenomenon known as implicit bias/regularization. In this work, we characterize the implicit bias of gradient descent with Polyak and line-search step sizes in linear classification with the logistic or cross-entropy loss. Given these step sizes are adaptive to local smoothness of the loss, we prove that the margin of their iterates converges to the maximum l_2-norm margin at $\tilde{O}(\frac{1}{T})$ rate. In contrast to other adaptive step sizes that achieve the same rate [7] (also known as normalized gradient descent-NGD), line-search and Polyak step sizes do not depend on problem-specific constants that may not be accessible. Another subtle issue is that NGD can diverge on common losses with non-separable data, whereas line-search converges given it guarantees descent on the function value at every iteration. Finally, our analysis extends the analysis framework of Wang et al. [26] to the logistic/cross-entropy losses.

1. Introduction

Gradient descent (GD) of the form

$$\theta_{t+1} = \theta_t - \eta_t \nabla f(\theta_t), \tag{1}$$

is a standard optimization algorithm for training machine learning models. To obtain its optimal performance, the step size (η_t) needs to be carefully tuned [2]. From the optimization literature, one can set it to be $1/L$ to guarantee descent on L-smooth functions [12]. However, L is a global constant that may fail to capture local curvature information [5, 9]. To this end, adaptive step sizes such as Armijo line-search (`GD-LS`) and Polyak step size (`GD-Polyak`) have been proposed [1, 14]. These

adaptive step sizes do not require the knowledge of L and are adaptive to the local smoothness of the objective. Since the local smoothness might be much smaller, adapting to it allows GD to use larger step-sizes and converge faster. Concretely, for logistic loss with separable data, GD-LS and GD-Polyak have been shown to converge linearly [24], which is faster than GD with arbitrarily large constant step size [28]. In addition to GD-LS and GD-Polyak, adaptive step sizes of the form $\eta_t \approx \eta/f(\theta_t)$ (denoted as GD-AD[1]) can be even faster, achieving arbitrarily small loss after a constant number of iterations with a properly chosen (large) η [30].

Despite the fast rates of Polyak and line-search step sizes in minimizing (exponentially-tailed) losses, theoretical understanding of their generalization performance is lacking. Towards this end, understanding algorithmic-specific implicit bias is important given it informs the inherent (without explicit regularization) preference of an algorithm over solutions that are equally-favorable in terms of training loss minimization. Linear classification with separable data is a standard test bed for studying optimization implicit bias [6, 18]. Our work is situated in this setting and motivated by the followings:

1. Despite GD-AD (with $\eta \leq 1$) achieving the optimal margin convergence rate for the exponential loss [7], the adaptive step size on the logistic loss (also exponentially-tailed) that obtains the same margin convergence rate depends on problem-specific constants that may not be known a priori. For example, the step size in Ji and Telgarsky [7] switches from the form $\Theta(1/f(\theta))$ to $\Theta(\exp(f(\theta))/(\exp(f(\theta))-1))$ after $\Theta([\ln n]^2/\gamma^2)$ iterations, where n and γ denote the number of data points and data margin respectively. This raises questions regarding its true adaptivity and practical usage given that the information on the data margin may not be accessible.

2. The performance of GD-AD is highly sensitive to the choice of η. As observed in Figure 1 (experiments on logistic loss), GD-AD with large η's show significant loss non-monotonicity and abrupt decrease in loss after a certain point. However, it may fail to make any progress in improving its margin. As max-margin separators typically generalize well [10, 17], this raises questions on whether GD-AD (with large η's) is still favorable from a margin-maximization perspective.

3. The implicit bias of Polyak and line-search step sizes are not well understood. We note that GD-Polyak was studied in the (non-convex) self-attention models, limited to the (binary) exponential loss, with a suboptimal margin convergence rate compared to GD-AD [23]. It remains open whether the optimal rate can be achieved with GD-Polyak. In general, understanding the implicit bias of GD-LS and GD-Polyak can provide more guidance for their future practices given the close interplay between optimization bias and generalization [22].

Taking into account all these factors, we aim to answer the following question:

*What is the **implicit bias** of **line-search** and **Polyak step size** in linear classification with separable data and logistic/cross-entropy loss?*

We use the game framework in Wang et al. [26] to analyze the implicit bias of GD-LS and GD-Polyak. Along the way, we extend their framework to the logistic loss (this was left as an open question in their paper). Our contributions are:

1. More precisely, the step size is $\eta_t = \frac{\eta}{f(\theta_t)}$ and $\eta_t = \frac{\eta \exp(f(\theta_t))}{\exp(f(\theta_t))-1}$ for exponential and logistic loss respectively.

1. We show that `GD-LS` and `GD-Polyak` achieve the (optimal) margin convergence rate $\tilde{O}(1/T)$ for the linear setting considered.

2. We empirically verify that `GD-LS` is robust to any algorithmic-specific hyperparameters such as the search initialization of the step size. Given any (sufficiently large) search initialization, the algorithm finds (appropriate) step sizes to quickly converge the max-margin separator.

2. Preliminaries

Related Works The implicit bias of GD in the linear setting with separable data has been studied in several works [6, 7, 18, 27]. For constant step sizes, it was shown that l_2-norm margin convergence rate of GD is $\tilde{O}(1/\log T)$ [6, 18, 27], and the same rate holds for the multiclass setting [15]. For adaptive step-size of the form $\eta_t \approx 1/f(\theta_t)$, the rate was improved to $\tilde{O}(1/T)$ [7]. Besides GD, several works have studied other algorithms such as normalized steepest descent with/without momentum [4, 11], mirror descent [19], and Adam [4, 29]. Beyond linear settings, other works have focused on diagonal, homogeneous and non-homogeneous neural networks [3, 8, 13], and self-attention [21, 23] (refer to the survey Vardi [22] for additional details).

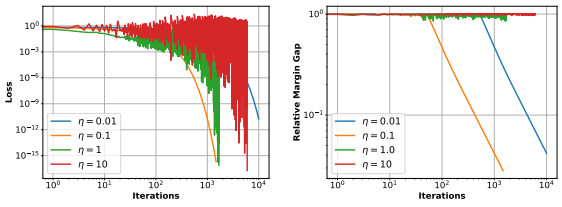


Figure 1: `GD-AD` on the logistic loss with separable data. Step size is of the form $\eta_t = \eta \exp(f(\theta_t))/(\exp(f(\theta_t))-1)$([7, 30]) for different values of η. Left: loss; Right: relative margin gap (see App. A for details).

Setup We focus on the linear classification setting with separable data (denoted as $\{(x_i, y_i)\}_{i=1}^n$). The exponential and logistic loss are defined as

$$f_{\exp}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \exp\big(-y_i\langle x_i, \theta\rangle\big), \quad f_{\log}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ln\big(1+\exp(-y_i\langle x_i, \theta\rangle)\big).$$

We use $f \in \{f_{\log}, f_{\exp}\}$ to denote either loss for simplicity. Denote the spectral norm or 2-norm by $\|\cdot\|$ when the argument is a matrix or vector respectively. The following self-boundedness properties satisfied by both the logistic and exponential loss ([16, 20, 24]) are central to our analysis

$$\left\|\nabla^2 f(\theta)\right\| \le L_1 f(\theta), \quad \text{and} \quad \|\nabla f(\theta)\| \le \nu f(\theta), \quad \forall\theta, \tag{2}$$

where $L_1 \ge 0$ and $\nu \ge 0$ are some constants. Intuitively, the (global) smoothness constant is replaced by $f(\theta)$ (approximately), which better captures the local behavior of a function. Furthermore, the max-margin of the data set is defined as $\gamma := \max_{\theta\in\mathbb{R}^d} \min_{i\in[n]} y_i\langle\theta, x_i\rangle/\|\theta\|$. We make the following standard assumptions that have appeared in various works that studied optimization implicit-bias [4, 6, 18, 26, 27]. The first one is on data separability: There exists $\theta \in \mathbb{R}^d$ such that $\min_{i\in[n]} y_i\langle\theta, x_i\rangle > 0$. The second one assumes that the data is properly scaled: It holds that $\|x_i\| \le 1$ for all $i \in [n]$.

Adaptive Step Sizes We consider the following adaptive step sizes

$$\texttt{GD-LS}: \quad f(\theta_t - \eta_t \nabla f(\theta_t)) \leq f(\theta_t) - c\eta_t \, \|\nabla f(\theta_t)\|_2^2 \quad \text{(Armijo Condition)},$$
$$\texttt{GD-Polyak}: \quad \eta_t = \min\left\{ \frac{f(\theta_t)}{c\,\|\nabla f(\theta_t)\|^2}, \eta_{\max}^t \right\},$$

where $c \in (0,1)$ and $c > 0$ for for `GD-LS` and `GD-Polyak` respectively. For `GD-LS`, a backtracking procedure starting from (sufficiently large) $\eta_{\max}^t$ is in place to find the (assumed) largest step size that satisfies the above Armijo condition [25]. Crucially, given both logistic and exponential losses satisfying (2) above, `GD-LS` and `GD-Polyak` return a step size at each iteration t s.t.

$$\eta_t \in \left[\min\left\{ \eta_{\max}^t, \frac{1}{C(L_1, \nu, c)\, f(\theta_t)} \right\}, \eta_{\max}^t \right], \tag{3}$$

where $C(L_1, \nu, c) := {}^{3L_1(\nu+1)}/_{(1-c)}$ (or $C(L_1, \nu, c) := c\nu^2$) for `GD-LS` (or `GD-Polyak`) [24]. Thus, we observe with a proper choice of $\eta_{\max}^t$, both step sizes are adaptive to the local smoothness.

Game Framework for Margin Analysis Next, we introduce the framework proposed by Wang et al. [26], which relates margin maximization to solving a zero-sum game. Specifically, the objective of the game is defined via $\max_{w \in \mathbb{R}^d} \min_{p \in \triangle^n} g(p, w) := p^T X w - \frac{1}{2}\|w\|^2$. Note that while the choice of norm can be more general [26], here we limit our discussions to the l_2-norm given the underlying geometry of GD is Euclidean. We denote $\{\alpha_t\}_{t=1}^T$ as a sequence of positive weights. To solve this game, the w-player first tries to minimize the weighted loss $\alpha_{t-1} h_{t-1}(w) := -\alpha_{t-1} g(p_{t-1}, w)$ for a given p_{t-1}. After making a decision w_t, the weighted loss $\alpha_t l_t(p) := \alpha_t g(p, w_t)$ is passed to the p-player for minimization. This process alternates and uses the weighted-average decision $\tilde{w}_T = \sum_{t=1}^T \alpha_t w_t$ as the final output. The regret bounds on the weighted losses for the w-player and p-player are

$$\text{w-player:} \sum_{t=1}^T \alpha_t h_t(w_t) - \min_{w \in \mathbb{R}^d} \sum_{t=1}^T \alpha_t h_t(w) \leq \text{Reg}_T^w,$$
$$\text{p-player:} \sum_{t=1}^T \alpha_t l_t(p_t) - \min_{p \in \triangle^n} \sum_{t=1}^T \alpha_t l_t(p) \leq \text{Reg}_T^p,$$

where $\triangle^n$ is the n-dimensional probability simplex. Further denote the weighted-average regret as $C_T := {}^{(Reg_T^p + Reg_T^w)}/_{\sum_{t=1}^T \alpha_t}$ and the data matrix as $X = [-y_i x_i^T -] \in \mathbb{R}^{n \times d}$. Then, Wang et al. [26, Theorem 1] shows that the margin convergence rate depends on C_T via

$$\min_{i \in [n]} \frac{y_i \langle \tilde{w}_T, x_i \rangle}{\|\tilde{w}_T\|} = \min_{p \in \triangle^n} \frac{p^T X \tilde{w}_T}{\|\tilde{w}_T\|} \geq \gamma - \frac{4C_T}{\gamma^2}, \tag{4}$$

provided T is chosen such that $C_T \leq \frac{\gamma^2}{4}$. From this, we observe that the weights α_t play a crucial role in determining the margin convergence rate. Moreover, the above online learning protocol also suggests a general recipe for studying optimization implicit-bias: 1. Determine the moves of the w-player and p-player s.t. $\tilde{w}_t = \theta_t$ for all t; 2. Derive the regret bounds of the w-player and p-player (which in turn determines C_T) and translate them into a margin convergence rate via (4).

Considering GD on the exponential loss, Wang et al. [26] proposed the following online learning algorithms for the w-player and p-player

$$\text{w-player: } w_t = \arg\min_{w\in\mathbb{R}^d} \alpha_{t-1} h_{t-1}(w) = \arg\min_{w\in\mathbb{R}^d} -p_{t-1}^T X w + \frac{1}{2}\|w\|^2 \iff w_t = X^T p_{t-1},$$

$$\text{p-player: } p_t = \arg\min_{p\in\Delta_n} \left[\sum_{i=1}^{t} \alpha_i \ell_i(p) + \mathrm{KL}(p||\frac{\mathbf{1}}{n})\right], \text{ where } \mathrm{KL}(p||\frac{\mathbf{1}}{n}) := \sum_{i\in[n]} p_i \log(\frac{p_i}{1/n}). \quad (5)$$

With these choices, the overall output is

$$\tilde{w}_t = \tilde{w}_{t-1} - \frac{\alpha_t}{f_{\exp}(\tilde{w}_{t-1})}\nabla f_{\exp}(\tilde{w}_{t-1}),$$

which is equivalent to GD with a step size $\eta_{t-1} = \frac{\alpha_t}{f_{\exp}(\tilde{w}_{t-1})}$.

3. Margin Convergence Results

Exponential Loss To further match the iterates of `GD-LS` and `GD-Polyak` to the output of the players, we apply the bounds on the step size (3) with $\eta_{\max}^t = 1/f_{\exp}(\theta_t)$ to determine the range of α_t, i.e. $\alpha_t \in [\min\{C(L_1,\nu,c),1\},1]$. Importantly, this range is time-independent ensured via the specific form of $\eta_{\max}^t$, which implies that $\sum_{t=1}^T \alpha_t = \Theta(T)$. Hence, as long as $\mathrm{Reg}_T^p + \mathrm{Reg}_T^w \le \Theta(1)$, we obtain the margin convergence rate via (4) stated in Theorem 1.

Theorem 1 *Set $\eta_{\max}^t = \frac{1}{f_{\exp}(\theta_t)}$. `GD-LS` and `GD-Polyak` achieve the following rate for the exponential loss:* $\min_{i\in[n]} \frac{y_i\langle\theta_T,x_i\rangle}{\|\theta_T\|} \ge \gamma - \Theta\big(\frac{\log n}{T}\big)$.

The rate in Theorem 1 matches the rate of NGD (i.e. GD with the step size $\eta_t = 1/f_{\exp(\theta_t)}$). However, NGD can diverge when the data is non-separable, whereas line-search still converges given it guarantees loss monotonicity.

Logistic Loss The current framework nicely integrates with the exponential loss due to the relationship $\nabla f_{\exp}(w)/f_{\exp}(w) = -X^T p$ (recall that X is the data matrix and $p \in \triangle^n$). However, it no longer holds for the logistic loss. To overcome this technical challenge, we introduce a time-dependent error term (ϵ_t) that captures the difference of the gradient-to-loss ratio between the logistic and exponential losses (defined in Lemma 2 below). For the logistic loss, we note that the p-player can still perform the same update as (5). However, for the w-player, instead of minimizing $\alpha_{t-1}h_{t-1}(w)$ at round t, it additionally adds ϵ_{t-1} to the output of this minimization problem. This gives rise to the form of w_t for $f_{\log}$ (in comparison to the form for $f_{\exp}$)

$$f_{\log}\text{: } w_t = X^T p_{t-1} - \epsilon_{t-1} \quad \text{vs.} \quad f_{\exp}\text{: } w_t = X^T p_{t-1}.$$

Finally, we set $\alpha_t = \eta_{t-1}/f_{\log}(\tilde{w}_{t-1})$ to match the players' output $\tilde{w}_t$ to GD's iterate θ_t. The following lemma formalizes these discussions and its proof can be found in App. D.

Lemma 2 *Considering the iterates of `GD` in (1). Set $\tilde{w}_1 = \theta_1$ and $\eta_{t-1} = \frac{\alpha_t}{f_{\log}(\tilde{w}_{t-1})}$. Suppose that the p-player performs the update in (5). If the w-player performs*

$$\tilde{w}_t = \tilde{w}_{t-1} + \alpha_t\big(X^T p_{t-1} - \epsilon_{t-1}\big), \quad \textit{where} \quad \epsilon_{t-1} := \frac{\nabla f_{\log}(\tilde{w}_{t-1})}{f_{\log}(\tilde{w}_{t-1})} - \frac{\nabla f_{\exp}(\tilde{w}_{t-1})}{f_{\exp}(\tilde{w}_{t-1})},$$

then it holds that $\tilde{w}_t = \theta_t$ for all $t \ge 1$.

After relating GD to the online learning protocol, we bound the regret of the w-player and p-player to obtain the results in Lemma 3. For the logistic loss, we choose $\eta_{\max}^t$ to be $\eta_{\max}^t = 1/f_{\log}(t)$ for the same reason as the exponential loss. The range of α_t is the same as that of the exponential loss.

Lemma 3 *Consider the updates of* `GD-LS` *and* `GD-Polyak`*. Set* $\eta_{\max}^t = \frac{1}{f_{\log}(t)}$*. Then it holds*

$$\frac{\textit{Regret}_T^p + \textit{Regret}_T^w}{\sum_{t=1}^T \alpha_t} \leq \Theta\left(\frac{\ln(n) + \sum_{t=1}^T [\|\epsilon_t\|_2^2 + \|\epsilon_t\|_2]}{T}\right).$$

Note that we can substitute θ_t for $\tilde{w}_t$ in the definition of ϵ_t above given $\tilde{w}_t = \theta_t, \forall t \geq 1$ (established in Lemma 2). To control the term $\sum_{t=1}^T \|\epsilon_t\|_2^2 + \|\epsilon_t\|_2$, we rely on Lemma 4 to bound the gradient-to-loss difference ϵ_t using the loss $f_{\log}(\theta_t)$ provided it is sufficiently small. Given the linear convergence rates of `GD-LS` and `GD-Polyak` on the logistic loss [24], we can show that $\sum_{t=1}^T \|\epsilon_t\|_2^2 + \|\epsilon_t\|_2 \leq \Theta([\ln n]^2)$ (details in App. D).

Lemma 4 *Let* w *be s.t.* $f_{\log}(w) \leq \frac{1}{2n}$*, then it holds that*

$$\left\| \frac{\nabla f_{\log}(w)}{f_{\log}(w)} - \frac{\nabla f_{\exp}(w)}{f_{\exp}(w)} \right\| \leq 3n\, f_{\log}(w).$$

Putting Lemma 2, 3, and 4 together, we arrive at the following theorem for `GD-LS` and `GD-Polyak`, which states that the l_2-norm margin of their (normalized) iterates converges to the max-margin (γ) at a $\tilde{O}(1/T)$ rate matching that of `GD-AD` [7]. In App. E, we discuss the extension of the game-framework to the multiclass setting by deriving Lemma 15 in analogous to Lemma 4. The experimental evaluations of `GD-LS` can be found in App A.

Theorem 5 *Suppose that* $T \geq \frac{4\Phi}{\gamma^2}$ *where* $\Phi = \Theta([\log n]^2)$*. Set* $\eta_{\max}^t = \frac{1}{f_{\log}(\theta_t)}$*.* `GD-LS` *and* `GD-Polyak` *achieve the following rate for the logistic loss:*

$$\min_{i\in[n]} \frac{y_i \langle \theta_T, x_i \rangle}{\|\theta_T\|} \geq \gamma - \Theta\left(\frac{[\log n]^2}{T}\right).$$

Remark 6 *This rate matches the rate of exponential loss up to a factor of* $\log n$*. Unlike the step size schedule in Ji and Telgarsky [7] that requires the knowledge of the data margin,* `GD-LS` *and* `GD-Polyak` *achieve the same rate (up to a* $\log n$ *factor) without requiring the access to any problem-specific constants. Finally, the rate in Theorem 5 also holds for the cross-entropy loss shown in App. E.*

4. Conclusion

In this paper, we have characterized the implicit bias of gradient descent with Polyak or line-search step sizes on linear separable data with the logistic/cross-entropy loss, and empirically verified their performances. Future works involve extending these adaptive step sizes to steepest descent or Nesterov momentum algorithms.

Acknowledgement

This work is funded partially by NSERC Discovery Grants RGPIN-2021-03677 and RGPIN-2022-03669, Alliance GrantALLRP 581098-22, a CIFAR AI Catalyst grant, and the Canada CIFAR AI Chair Program.

References

[1] Larry Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*, 1966.

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

[3] 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.

[4] Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. *arXiv preprint arXiv:2502.04664*, 2025.

[5] Curtis Fox, Aaron Mishkin, Sharan Vaswani, and Mark Schmidt. Glocal smoothness: Line search can really help! *arXiv preprint arXiv:2506.12648*, 2025.

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

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

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

[9] Aaron Mishkin, Ahmed Khaled, Yuanhao Wang, Aaron Defazio, and Robert Gower. Directional smoothness and gradient methods: Convergence and adaptivity. *Advances in Neural Information Processing Systems*, 37:14810–14848, 2024.

[10] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[11] 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.

[12] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[13] 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.

[14] Boris T Polyak. Introduction to optimization. 1987.

[15] Hrithik Ravi, Clay Scott, Daniel Soudry, and Yutong Wang. The implicit bias of gradient descent on separable multiclass data. *Advances in Neural Information Processing Systems*, 37:81324–81359, 2024.

[16] Matan Schliserman and Tomer Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In *Conference on Learning Theory*, pages 3380–3394. PMLR, 2022.

[17] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[18] 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.

[19] 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.

[20] Hossein Taheri and Christos Thrampoulidis. Generalization and stability of interpolating neural networks with minimal width. *Journal of Machine Learning Research*, 25(156):1–41, 2024.

[21] Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023.

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

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

[24] Sharan Vaswani and Reza Babanezhad. Armijo line-search makes (stochastic) gradient descent go fast. *arXiv preprint arXiv:2503.00229*, 2025.

[25] Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. *Advances in neural information processing systems*, 32:3732–3745, 2019.

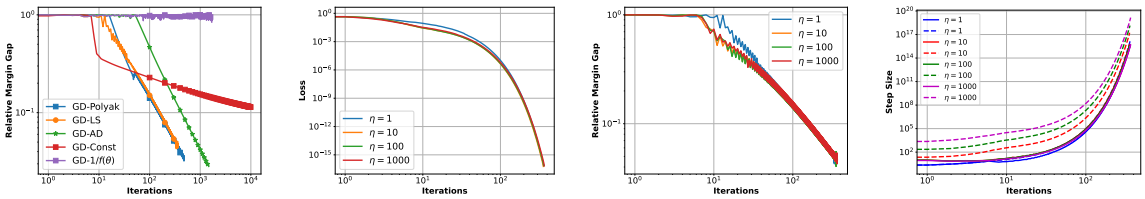
[26] Guanghui Wang, Zihao Hu, Vidya Muthukumar, and Jacob D Abernethy. Faster margin maximization rates for generic optimization methods. *Advances in Neural Information Processing Systems*, 36:62488–62518, 2023.

[27] 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*, 36:74229–74256, 2023.

[28] 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.

[29] Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of adam on separable data. *Advances in Neural Information Processing Systems*, 37:23988–24021, 2024.

[30] Ruiqi Zhang, Jingfeng Wu, Licong Lin, and Peter L Bartlett. Minimax optimal convergence of gradient descent in logistic regression via large and adaptive stepsizes. *arXiv preprint arXiv:2504.04105*, 2025.



(*a*) Margin (*b*) `GD-LS` Loss (*c*) `GD-LS` Margin (*d*) Step Size

Figure 2: (a) Relative margin gap defined via $|\gamma - \tilde{\gamma}(\theta_t)|/\gamma$ against iterations. Legend indicates different algorithms. (b) Loss of `GD-LS` against iterations. Legend indicates different η's for search initialization of the form $\eta^t_{\max} =: \eta/f(\theta_t)$. (c) Same plot as (b) with loss replaced by relative margin gap. (d) Step size of `GD-LS` against iterations. Dash-line: search initialization of different η's; Solid-line: step size return from backtracking line-search with the corresponding search initialization.

Appendix A. Experiments

Experiments We perform experiments on synthetic (binary) data generated from a standard multivariate normal distribution with $n = 500$ and $d = 500$. We ensure the data is separable by checking the margin being positive. We denote $\tilde{\gamma}(\theta) := \min_{i\in[n]} y_i\langle\theta_T, x_i\rangle/\|\theta_T\|$ as the margin of GD's (normalized) iterates. The results are shown in Figure 2 in Appendix, from which we conclude the followings: (a) The iterates of `GD-LS` converge to l_2-norm max-margin fast compared against others (Figure 2(*a*)); (b) `GD-LS` is robust to search initialization for loss minimization (Figure 2(*b*)) and margin maximization (Figure 2(*c*)); (c) For different search initializations, step sizes of `GD-LS` follow closely to each other and to $1/f(\theta_t)$ (Figure 2(*d*)). Note that direct use of $\eta_t = 1/f(\theta_t)$ does not decrease the relative margin gap (Figure 2(*a*)). This suggests that the fine adjustments of `GD-LS` made to the step sizes helps with margin convergence.

Appendix B. Auxiliary Lemmas

From Vaswani and Babanezhad [24, Proposition 5], we know that a function f is (L_0, L_1) non-uniform smooth and satisfies the following inequalities:

(a) For all x, y such that $\|x - y\| \le \frac{q}{L_1}$ where $q \ge 1$ is a constant, if $A := 1 + e^q - \frac{e^q - 1}{q}$ and $B := \frac{e^q - 1}{q}$,

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{(A\,L_0 + B\,L_1\,f(x))}{2}\,\|y - x\|_2^2\,, \tag{6}$$

(b) For all θ, $\left\|\nabla^2 f(\theta)\right\| \le L_0 + L_1\,f(\theta)$,

(c) $\|\nabla f(\theta)\| \le \nu\,f(\theta) + \omega$,

Lemma 7 *(Vaswani and Babanezhad [24, Lemma 1]) Let $f \in \{f_{\exp}, f_{\log}\}$. At iteration t, the step size of `GD-LS` and `GD-Polyak` satisfy*

$$\eta_t \in \left[\min\left\{\frac{1}{C(L_1, \nu, c)\,f(\theta_t)}, \eta^t_{\max}\right\}, \eta^t_{\max}\right]$$

where $C(L_1, \nu, c) := 3\,\frac{L_1(\nu+1)}{(1-c)}$ *for* `GD-LS` *and* $C(L_1, \nu, c) := c\,\nu^2$ *for* `GD-Polyak`.

Lemma 8 *(Vaswani and Babanezhad [24, Lemma 4]) For* $\epsilon \in (0, M)$ *and a comparator* u *s.t.* $f(u) \leq \epsilon$*, if* f *satisfies (6) with* $L_0 = 0$ *and* $\omega = 0$*, then, for all* θ *s.t.* $\|\theta - u\| \leq \frac{q}{L_1}$*,*

$$f(\theta) - f(u) \leq \frac{\epsilon}{2} + \left[\nu^2\, M + B\, L_1\, M\right] \frac{\|\theta - u\|_2^2}{2}\,,$$

where $B := \frac{e^q - 1}{q}$*. Furthermore, if* f *is also* L *uniform smooth, then, for all* θ*,*

$$f(\theta) - f(u) \leq \frac{\epsilon}{2} + \left[\nu^2\, M + L\right] \frac{\|\theta - u\|_2^2}{2}\,,$$

Appendix C. Exponential Loss

In order to use the framework in Wang et al. [26], we need to choose the updates for the w and p players. For the w player at iteration $t-1$, let us consider the following GD update on the weighted loss. For a step-size $\delta_{t-1} = \frac{1}{\alpha_{t-1}}$,

$$w_t = w_{t-1} - \delta_{t-1}\left[\alpha_{t-1}\,\nabla h_{t-1}(w_{t-1})\right] = w_{t-1} + \delta_{t-1}\,\alpha_{t-1}\left[p_{t-1}^T\, X - w_{t-1}\right] \tag{7}$$

Since $\delta_{t-1}\,\alpha_{t-1} = 1$ for all t,

$$\implies w_t = X^T\, p_{t-1} \tag{8}$$

Recall that $\tilde{w}_t := \sum_{i=1}^t \alpha_i\, w_i$. For the p player, we use the following FTRL update – at iteration t,

$$p_t = \operatorname*{arg\,min}_{p \in \Delta_n} \left[\sum_{i=1}^t \alpha_i\, \ell_i(p) + \mathrm{KL}(p||1/n)\right] \tag{9}$$

$$\implies (p_t)_j \propto \exp\left(-\sum_{i=1}^t \alpha_i\, [\nabla \ell_i(p_t)]_j\right) \implies (p_t)_j \propto \exp\left(-y_j \left\langle x_j, \sum_{i=1}^t \alpha_i\, w_i \right\rangle\right)$$

(Since $[\nabla \ell_t(p_t)]_j = y_j \langle x_j, w_t \rangle$)

$$\implies (p_t)_j \propto \exp\left(-y_j\, \langle x_j, \tilde{w}_t \rangle\right)$$ (By definition of $\tilde{w}_t$)

Next, we use the above inequalities and the properties of the exponential loss to prove that,

$$\frac{\nabla f_{\exp}(\tilde{w}_t)}{f_{\exp}(\tilde{w}_t)} = -X^T\, p_t \implies \nabla f_{\exp}(\tilde{w}_t) = -f_{\exp}(\tilde{w}_t)\, X^T\, p_t \tag{10}$$

Using the definition of $\tilde{w}_t$,

$$\tilde{w}_t = \tilde{w}_{t-1} + \alpha_t\, w_t = \tilde{w}_{t-1} + \alpha_t\, [X^T p_{t-1}]$$ (Using eq. (8))

$$\implies \tilde{w}_t = \tilde{w}_{t-1} - \alpha_t\, \frac{\nabla f_{\exp}(\tilde{w}_{t-1})}{f_{\exp}(\tilde{w}_{t-1})}$$ (Using eq. (10))

Comparing this to the GD-LS update in eq. (1), if (i) $\tilde{w}_1 = \theta_1$ and (ii) $\eta_{t-1} = \frac{\alpha_t}{f_{\exp}(\tilde{w}_{t-1})}$ for all t, then, $\tilde{w}_t = \theta_t$ for all t. From Lemma 7, we know that, for all t,

$$\eta_t \geq \min\left\{\eta_{\max}^t, \frac{1-c}{3L_1\,(\nu+1)}\,\frac{1}{f_{\exp}(\theta_t)}\right\} \implies \alpha_t \geq \min\left\{\eta_{\max}^{t-1}\,f_{\exp}(\theta_{t-1}), \frac{1-c}{3L_1\,(\nu+1)}\right\}$$

(Using the above relation)

$$\implies \delta_t \leq \max\left\{\frac{1}{\eta_{\max}^{t-1}\,f_{\exp}(\theta_{t-1})}, \frac{3L_1\,(\nu+1)}{1-c}\right\} \qquad \text{(Since } \alpha_{t-1}\,\delta_{t-1} = 1 \text{ for all } t\text{)}$$

On the other hand, we know that

$$\eta_t \leq \eta_{\max}^t \implies \alpha_t \leq \eta_{\max}^{t-1}\,f_{\exp}(\theta_{t-1}) \implies \delta_t \geq \frac{1}{\eta_{\max}^{t-1}\,f_{\exp}(\theta_{t-1})}$$

Now, we will bound the regret for the w player and the p player, and use Wang et al. [26, Theorem 1]. For the w player, recall the update,

$$w_{t+1} = w_t - \delta_t\,[\alpha_t\,\nabla h_t(w_t)] = w_t - \nabla h_t(w_t) \qquad \text{(Since } \delta_t\,\alpha_t = 1\text{)}$$

For a comparator u,

$$\begin{aligned}
\implies \|w_{t+1} - u\|_2^2 &= \|w_t - \nabla h_t(w_t) - u\|_2^2 = \|w_t - u\|_2^2 - 2\langle w_t - u, \nabla h_t(w_t)\rangle + \|\nabla h_t(w_t)\|_2^2 \\
&\leq \|w_t - u\|_2^2 - 2\left[h_t(w_t) - h_t(u) + \frac{1}{2}\,\|w_t - u\|_2^2\right] + \|\nabla h_t(w_t)\|_2^2 \\
&\qquad \text{(Since } h \text{ is 1 strongly-convex)} \\
\implies \alpha_t\,[h_t(w_t) - h_t(u)] &\leq \frac{\alpha_t\,\|\nabla h_t(w_t)\|_2^2}{2} \\
&\qquad \text{(Rearranging and multiplying throughout by } \alpha_t > 0\text{)}
\end{aligned}$$

Bounding $\|\nabla h_t(w_t)\|_2^2$ similar to the proof of Wang et al. [26, Theorem 9],

$$\begin{aligned}
\|\nabla h_t(w_t)\|_2^2 &= \left\|w_t - p_t^T\,X\right\|_2^2 = \left\|p_{t-1}^T X - p_t^T\,X\right\|_2^2 \quad \text{(Using the definition of } h_t(w) \text{ and eq. (8))} \\
&= \left(\left\|\sum_{i=1}^n y_i\,x_i\,(p_t(i) - p_{t-1}(i))\right\|\right)^2 \qquad \text{(By definition of } X\text{)} \\
&\leq \left(\sum_{i=1}^n |p_t(i) - p_{t-1}(i)|\right)^2 \qquad \text{(Triangle inequality and since } \|y_i\,x_i\| \leq 1\text{)} \\
&= \|p_t - p_{t-1}\|_1^2
\end{aligned}$$

Combining the above relations and summing from $t = 1$ to T,

$$\text{Regret}_T^w \leq \sum_{t=1}^T \frac{\alpha_t}{2}\,\|p_t - p_{t-1}\|_1^2$$

Setting $\eta_{\max}^{t-1} = \frac{1}{f_{\exp}(\theta_{t-1})}$ ensures that,

$$\alpha_t \in \left[\min\left\{1, \frac{1-c}{3L_1\,(\nu+1)}\right\}, 1\right] \implies \alpha_t \in [\min\{C,1\},1] \qquad (\text{Define } C := \tfrac{1-c}{3L_1\,(\nu+1)})$$

Using this relation to simplify Regret_T^w, since $\alpha_t \le 1$,

$$\text{Regret}_T^w \le \sum_{t=1}^{T} \frac{1}{2}\,\|p_t - p_{t-1}\|_1^2$$

For the p player, the regret for FTRL can be directly bounded using Wang et al. [26, Lemma 6],

$$\text{Regret}_T^p \le \ln(n) - \frac{1}{2}\sum_{t=1}^{T}\|p_t - p_{t-1}\|_1^2$$

Using the definition of C_T,

$$C_T \le \frac{\ln(n)}{\sum_{t=1}^{T}\alpha_t} \le \frac{\ln(n)}{\min\{C,1\}\,T} \qquad (\text{Since } \alpha_t \ge \min\{C,1\})$$

Using Wang et al. [26, Theorem 1], for $T \ge \frac{4\ln(n)}{\min\{C,1\}\,\gamma^2}$,

$$\tilde{\gamma}(\tilde{w}_T) = \tilde{\gamma}(\theta_T) \ge \gamma - \frac{4\,C_T}{\gamma^2} = \gamma - \frac{4}{\gamma^2}\,\frac{\ln(n)}{\min\{C,1\}\,T}$$

Appendix D. Main Proofs

Lemma 9 *Considering the update of* GD *in* (1). *Set* $\tilde{w}_1 = \theta_1$ *and* $\eta_t = \frac{\alpha_{t+1}}{f_{\log}(\tilde{w}_t)}$. *Suppose that* w*-player and* p*-player perform the following updates*

$$\textit{p-player:} \quad p_t = \arg\min_{p\in\Delta_n}\left[\sum_{i=1}^{t}\alpha_i\,\ell_i(p) + \mathit{KL}(p||\frac{\mathbf{1}}{n})\right],$$

$$\textit{w-player:} \quad \tilde{w}_{t+1} = \tilde{w}_t + \alpha_{t+1}\big(X^T p_t - \epsilon_t\big), \quad \textit{where} \quad \epsilon_t := \frac{\nabla f_{\log}(\tilde{w}_t)}{f_{\log}(\tilde{w}_t)} - \frac{\nabla f_{\exp}(\tilde{w}_t)}{f_{\exp}(\tilde{w}_t)}.$$

Then, it holds that $\tilde{w}_t = \theta_t$ *for all* $t \ge 1$.

Proof For the p player, we use the following FTRL update

$$p_t = \arg\min_{p\in\Delta_n}\left[\sum_{i=1}^{t}\alpha_i\,\ell_i(p) + \text{KL}(p||\frac{\mathbf{1}}{n})\right],$$

where $\frac{\mathbf{1}}{n} \in \mathbb{R}^n$ is a vector with all entries being $\frac{1}{n}$. The result of this minimization problem gives

$$p_t[i] = \frac{\exp(-y_i\,\langle\tilde{w}_t, x_i\rangle)}{\sum_{j=1}^{n}\exp(-y_j\,\langle\tilde{w}_t, x_j\rangle)}.$$

For the w player, we let $\epsilon_t := \frac{\nabla f_{\log}(\tilde{w}_t)}{f_{\log}(\tilde{w}_t)} - \frac{\nabla f_{\exp}(\tilde{w}_t)}{f_{\exp}(\tilde{w}_t)}$ and perform the following

$$w_{t+1} = w_t - [\nabla h_t(w_t) + \epsilon_t] \overset{\text{By def of } h_t(w)}{=} X^T p_t - \epsilon_t \quad \text{and} \quad \tilde{w}_{t+1} = \tilde{w}_t + \alpha_{t+1} w_{t+1}.$$

Note that for exponential loss, it holds

$$\frac{\nabla f_{\exp}(w)}{f_{\exp}(w)} = \sum_{i=1}^{n} -y_i\, x_i \frac{\exp(-y_i \langle w, x_i \rangle)}{\sum_{j=1}^{n} \exp(-y_j \langle w, x_j \rangle)} = -X^T p_t.$$

On the other hand, by $\nabla f_{\log}(w) = -\frac{1}{n} \sum_{i=1}^{n} y_i\, x_i \frac{\exp(-y_i \langle w, x_i \rangle)}{1+\exp(-y_i \langle w, x_i \rangle)}$, it holds for logistic loss that

$$\frac{\nabla f_{\log}(\tilde{w}_t)}{f_{\log}(\tilde{w}_t)} = -\frac{X^T q_t}{f_{\log}(\tilde{w}_t)} \quad \text{where} \quad q_t \in \mathbb{R}^n \ \text{ s.t } \ [q_t]_i = \frac{1}{n} \frac{\exp(-y_i \langle \tilde{w}_t, x_j \rangle)}{1 + \exp(-y_i \langle \tilde{w}_t, x_i \rangle)}.$$

First, we prove that $\tilde{w}_t = \theta_t$ for all t by induction where θ_t are the iterates of `GD-LS` on the logistic loss. The base case can be satisfied by initializing $\tilde{w}_1 = \theta_1$. Assuming that $\tilde{w}_t = \theta_t$, then,

$$\begin{aligned}
\tilde{w}_{t+1} &= \tilde{w}_t + \alpha_{t+1}\, w_{t+1} = \theta_t + \alpha_{t+1}\, X^T p_t - \alpha_{t+1}\, \epsilon_t \\
&= \theta_t - \alpha_{t+1} \frac{\nabla f_{\exp}(\tilde{w}_t)}{f_{\exp}(\tilde{w}_t)} - \alpha_{t+1} \left[\frac{\nabla f_{\log}(\tilde{w}_t)}{f_{\log}(\tilde{w}_t)} - \frac{\nabla f_{\exp}(\tilde{w}_t)}{f_{\exp}(\tilde{w}_t)} \right] = \theta_t - \alpha_{t+1} \frac{\nabla f_{\log}(\tilde{w}_t)}{f_{\log}(\tilde{w}_t)}.
\end{aligned}$$

Recall that the update for GD is $\theta_{t+1} = \theta_t - \eta_t \nabla f(\theta_t)$. Comparing those two equations, we conclude that if (i) $\tilde{w}_1 = \theta_1$ and (ii) $\eta_t = \frac{\alpha_{t+1}}{f_{\log}(\tilde{w}_t)}$, then it holds that $\tilde{w}_t = \theta_t$ for all t.

■

Lemma 10 *Consider the updates of `GD-LS` and `GD-Polyak`. Set $\eta_{\max}^t = \frac{1}{f_{\log}(t)}$. Then it holds*

$$\frac{\textit{Regret}_T^p + \textit{Regret}_T^w}{\sum_{t=1}^{T} \alpha_t} \leq \Theta\Big(\frac{\ln(n) + \sum_{t=1}^{T} [\|\epsilon_t\|_2^2 + \|\epsilon_t\|]}{T}\Big),$$

recall that $\epsilon_t = \frac{\nabla f_{\log}(\tilde{w}_t)}{f_{\log}(\tilde{w}_t)} - \frac{\nabla f_{\exp}(\tilde{w}_t)}{f_{\exp}(\tilde{w}_t)}$.

Proof For the p player, we can directly use the result from Wang et al. [26, Theorem 5] to get

$$\text{Regret}_T^p \leq \ln(n) - \frac{1}{2} \sum_{t=1}^{T} \|p_t - p_{t-1}\|_1^2.$$

For the w player, we have that for an arbitrary comparator u

$$\begin{aligned}
\|w_{t+1} - u\|_2^2 &= \|w_t - u - \nabla h_t(w_t) - \epsilon_t\|_2^2 \\
&= \|w_t - u\|_2^2 - 2\langle w_t - u, \nabla h_t(w_t) \rangle - 2\langle w_t - u, \epsilon_t \rangle + \|\nabla h_t(w_t) + \epsilon_t\|_2^2 \\
&\overset{(a)}{\leq} \|w_t - u\|_2^2 - 2 \left[h_t(w_t) - h_t(u) + \frac{1}{2} \|w_t - u\|_2^2 \right] - 2\langle w_t - u, \epsilon_t \rangle + \|\nabla h_t(w_t) + \epsilon_t\|_2^2,
\end{aligned}$$

where (a) is by h_t being 1-strongly convex. This leads to

$$\begin{aligned}
2\left[h_t(w_t) - h_t(u)\right] &\leq - \left\|w_{t+1} - u\right\|_2^2 - 2\langle w_t - u, \epsilon_t\rangle + \left\|\nabla h_t(w_t) + \epsilon_t\right\|_2^2 \\
&= - \left\|w_{t+1} - u\right\|_2^2 - 2\langle w_t - u, \epsilon_t\rangle + \left\|\nabla h_t(w_t)\right\|_2^2 + \left\|\epsilon_t\right\|_2^2 + 2\langle \epsilon_t, \nabla h_t(w_t)\rangle \\
&= - \left\|w_{t+1} - u\right\|_2^2 + \left\|\nabla h_t(w_t)\right\|_2^2 + \left\|\epsilon_t\right\|_2^2 - 2\langle \epsilon_t, w_t - u - \nabla h_t(w_t)\rangle \\
&= - \left\|w_{t+1} - u\right\|_2^2 + \left\|\nabla h_t(w_t)\right\|_2^2 + \left\|\epsilon_t\right\|_2^2 - 2\langle \epsilon_t, w_t - u - \nabla h_t(w_t) - \epsilon_t\rangle - 2\langle \epsilon_t, \epsilon_t\rangle \\
&= - \left\|w_{t+1} - u\right\|_2^2 + \left\|\nabla h_t(w_t)\right\|_2^2 - \left\|\epsilon_t\right\|_2^2 - 2\langle \epsilon_t, w_{t+1} - u\rangle \\
&\overset{(b)}{\leq} - \left\|w_{t+1} - u\right\|_2^2 + \left\|\nabla h_t(w_t)\right\|_2^2 - \left\|\epsilon_t\right\|_2^2 + \left\|\epsilon_t\right\|_2^2 + \left\|w_{t+1} - u\right\|_2^2 \\
&= \left\|\nabla h_t(w_t)\right\|_2^2,
\end{aligned}$$

where (b) is by Young's inequality. This implies that

$$\alpha_t \left[h_t(w_t) - h_t(u)\right] \leq \frac{\alpha_t \left\|\nabla h_t(w_t)\right\|_2^2}{2}.$$

Next, we bound $\left\|\nabla h_t(w_t)\right\|_2^2$ as

$$\begin{aligned}
\left\|\nabla h_t(w_t)\right\|_2^2 &= \left\|w_t - p_t^T X\right\|_2^2 = \left\|X^T p_{t-1} - \epsilon_{t-1} - X^T p_t\right\|_2^2 \\
&= \left\|X^T p_{t-1} - X^T p_t\right\|_2^2 + \left\|\epsilon_{t-1}\right\|_2^2 + 2\langle \epsilon_{t-1}, X^T[p_t - p_{t-1}]\rangle \\
&\leq \left\|X^T p_{t-1} - X^T p_t\right\|_2^2 + \left\|\epsilon_{t-1}\right\|_2^2 + 2\left\|\epsilon_{t-1}\right\|_2 \left\|X^T p_{t-1} - X^T p_t\right\|_2
\end{aligned}$$

Note the following

$$\begin{aligned}
\left\|p_{t-1}^T X - p_t^T X\right\|_2^2 &= \left(\left\|\sum_{i=1}^n y_i\, x_i \left(p_t(i) - p_{t-1}(i)\right)\right\|_2\right)^2 \\
&\overset{(c)}{\leq} \left(\sum_{i=1}^n \left|p_t(i) - p_{t-1}(i)\right|\right)^2 \\
&= \left\|p_t - p_{t-1}\right\|_1^2 \leq 2,
\end{aligned}$$

where (c) is by triangle's inequality the the assumption $\|x_i\| \leq 1, \forall i \in [n]$. Putting things together,

$$\left\|\nabla h_t(w_t)\right\|_2^2 \leq \left\|p_t - p_{t-1}\right\|_1^2 + \left\|\epsilon_{t-1}\right\|_2^2 + 4\left\|\epsilon_{t-1}\right\|_2.$$

Combining the above relations and summing from $t = 1$ to T,

$$\begin{aligned}
\text{Regret}_T^w &\leq \sum_{t=1}^T \frac{\alpha_t}{2} \left\|p_t - p_{t-1}\right\|_1^2 + \sum_{t=1}^T \frac{\alpha_t}{2} \left[\left\|\epsilon_{t-1}\right\|_2^2 + 4\left\|\epsilon_{t-1}\right\|_2\right] \\
&\overset{(d)}{\leq} \sum_{t=1}^T \frac{1}{2} \left\|p_t - p_{t-1}\right\|_1^2 + \sum_{t=1}^T \frac{1}{2} \left[\left\|\epsilon_{t-1}\right\|_2^2 + 4\left\|\epsilon_{t-1}\right\|_2\right].
\end{aligned}$$

To justify (d): Let $f \in \{f_{\log}, f_{\exp}\}$. Given $\eta_{t-1} = \frac{\alpha_t}{f(\tilde{w}_{t-1})} = \frac{\alpha_t}{f(\theta_{t-1})}$ as $\tilde{w}_t = \theta_t$ for all t, we have from Lemma 7

$$\eta_t \in \left[\min\{\eta_{\max}^t, \frac{1}{C(L_1,\nu,c) f(\theta_t)}\}, \eta_{\max}^t\right] \implies \alpha_t \left[\min\{\eta_{\max}^t f(\theta_t), \frac{1}{C(L_1,\nu,c)}\}, \eta_{\max}^t f(\theta_t)\right]$$
$$\implies \alpha_t \in \left[\min\{1, \frac{1}{C(L_1,\nu,c)}\}, 1\right],$$

where $\eta_{\max}^t = \frac{1}{f(\theta_t)}$, $C(L_1,\nu,c) = \frac{3L_1(\nu+1)}{1-c}$ for `GD-LS`, and $C(L_1,\nu,c) = c\nu^2$ for `GD-Polyak`. Thus, we have that $\sum_{t=1}^T \alpha_t \in \left[\min\{1, \frac{1}{C(L_1,\nu,c)}\}T, T\right] = \Theta(T)$. Summing the regrets of the w-player and p-player leads to the desired. ■

Lemma 11 *Let $f \in \{f_{\exp}, f_{\log}\}$. For any initialization θ_0, choose an $\epsilon \in (0, f(\theta_0))$. Then, `GD-LS` with $\eta_{\max}^t = \frac{1}{f(\theta_t)}$ requires*

$$T \geq \frac{1}{C'\gamma^2} \left[\ln\left(\frac{1}{\epsilon}\right)\right]^2$$

iterations to ensure that $f(\theta_T) \leq 2\,\epsilon$, where $C' := \frac{2c-1}{c} \min\left\{\frac{1}{\lambda_1}, 1\right\}$ and $\lambda_1 := 3\,\frac{L_1(\nu+1)}{(1-c)}$.

Proof The proof follows the same steps as (Vaswani and Babanezhad [24, Theorem 2 and Corollary 2]). In their case, the constant $\eta_{\max}^t$ is set to ∞, resulting in $C' = \frac{2c-1}{c\lambda_1}$. ■

Lemma 12 *Let $f \in \{f_{\exp}, f_{\log}\}$. For any initialization θ_0, choose an $\epsilon \in (0, f(\theta_0))$. Then, GD with Polyak step-size $\eta_t = \min\left\{\frac{f(\theta_t)}{c\,\|\nabla f(\theta_t)\|_2^2}, \frac{1}{f(\theta_t)}\right\}$ for some $c > 1$ requires*

$$T \geq \frac{1}{C'\gamma^2} \left[\ln\left(\frac{c\nu^2}{\epsilon}\right)\right]^2$$

iterations to ensure that $f(\theta_T) \leq 2\,\epsilon$, where $C' := \frac{c-1}{c^2\nu^2}$.

Proof The logistic loss on linearly separable data is convex, satisfies (6) with $L_0 = 0$, $\omega = 0$, $\nu = 8$, and $f^* = 0$. We also know that $\|\nabla f(\theta)\| \leq \nu\, f(\theta)$ and we can bound the Polyak step-size as:

$$\eta_t \in \left[\min\left\{\frac{1}{c\,\nu^2\, f(\theta_t)}, \frac{1}{f(\theta_t)}\right\}, \frac{1}{f(\theta_t)}\right]. \tag{11}$$

Using the GD update: $\theta_{t+1} = \theta_t - \eta_t\, \nabla f(\theta_t)$, consider a comparator u s.t. $f(u) \leq \frac{\epsilon}{2\max\{c\,\nu^2, 1\}}$ and $f(u) \leq f(\theta_t)$ for all $t \in [T]$. Assuming that T is the first iteration such that $f(\theta_T) - f(u) \leq \epsilon$,

we have that,

$$\begin{aligned}
\|\theta_{t+1} - u\|_2^2 &= \|\theta_t - u\|_2^2 - 2\,\eta_t\,\langle \nabla f(\theta_t), \theta_t - u\rangle + {\eta_t}^2\,\|\nabla f(\theta_t)\|_2^2 \\
&\leq \|\theta_t - u\|_2^2 - 2\,\eta_t\,[f(\theta_t) - f(u)] + {\eta_t}^2\,\|\nabla f(\theta_t)\|_2^2 && \text{(Convexity)} \\
&\leq \|\theta_t - u\|_2^2 - 2\,\eta_t\,[f(\theta_t) - f(u)] + \frac{\eta_t}{c}\,[f(\theta_t)] \\
&= \|\theta_t - u\|_2^2 - \left(2 - \frac{1}{c}\right)\,\eta_t\,f(\theta_t) + 2\,\eta_t\,f(u) \\
&\leq \|\theta_t - u\|_2^2 - \left(2 - \frac{1}{c}\right)\,\min\left\{\frac{1}{c\,\nu^2}, 1\right\} + 2\,\eta_t\,f(u) \\
&\leq \|\theta_t - u\|_2^2 - \left(2 - \frac{1}{c}\right)\,\frac{1}{\max\{c\,\nu^2, 1\}} + \frac{2f(u)}{f(\theta_t)} \\
&\leq \|\theta_t - u\|_2^2 - \left(2 - \frac{1}{c}\right)\,\frac{1}{\max\{c\,\nu^2, 1\}} + \frac{2f(u)}{\epsilon} && \text{(Since } f(\theta_t) \geq \epsilon \text{ for all } t \in [T]\text{)} \\
&\leq \|\theta_t - u\|_2^2 - \left(2 - \frac{1}{c}\right)\,\frac{1}{\max\{c\,\nu^2, 1\}} + \frac{1}{\max\{c\,\nu^2, 1\}} \\
&= \|\theta_t - u\|_2^2 - \underbrace{\left(1 - \frac{1}{c}\right)\,\frac{1}{\max\{c\,\nu^2, 1\}}}_{:=C}
\end{aligned}$$

Summing up from $t = 0$ to $t = T - 1$,

$$\|\theta_T - u\|_2^2 \leq \|\theta_0 - u\|_2^2 - CT = \|u\|_2^2 - C\,Tu \qquad \text{(Since } \theta_0 = 0\text{)}$$

Since f is 1 uniformly smooth, using Lemma 8 with $M = f(\theta_0)$, we get

$$f(\theta_T) - f(u) \leq \frac{\epsilon}{2} + \underbrace{[\nu^2\,f(\theta_0) + 1]}_{:=L}\,\frac{\|\theta_T - u\|_2^2}{2} \leq \frac{\epsilon}{2} + L\,\left[\|u\|_2^2 - C\,T\right]$$

To ensure that $f(\theta_T) - f(u) \leq \epsilon$, it is sufficient to set $T \geq \frac{\|u\|_2^2}{C}$. In order to bound $\|u\|$, we define u^* to be the max-margin solution i.e. $\|u^*\| = 1$ and γ to be the corresponding margin, i.e. $\gamma := \min_i\,y_i\langle x_i, u^*\rangle$. Consider $u = \beta\,u^*$, for a scalar $\beta = \frac{1}{\gamma}\,\ln\left(\frac{\max\{c\nu^2,1\}}{\epsilon}\right)$, we have that

$$f(u) = \frac{1}{n}\,\sum_{i=1}^{n}\ln(1 + \exp(-y_i\langle x_i, \beta u^*\rangle)) \leq \frac{1}{n}\,\sum_{i=1}^{n}\exp(-y_i\langle x_i, \beta u^*\rangle) \leq \exp(-\beta\gamma) = \frac{\epsilon}{\max\{c\nu^2, 1\}}$$

This satisfies the requirement on $f(u)$. Note that we have $\max\{c\,\nu^2, 1\} = c\nu^2$ since $\nu = 8$ and $\nu = 1$ for $f_{\log}$ and $f_{\exp}$ respectively, and $c > 1$. Using this to bound T, we obtain

$$T \geq \frac{\beta^2}{C} = \frac{c^2\,\nu^2}{(c-1)\,\gamma^2}\,\left[\ln\left(\frac{c\,\nu^2}{\epsilon}\right)\right]^2$$

Finally, we conclude that after $T = \frac{\beta^2}{C} = \frac{c^2\,\nu^2}{(c-1)\,\gamma^2}\,\left[\ln\left(\frac{c\,\nu^2}{\epsilon}\right)\right]^2$ iterations it holds $f(\theta_T) - f(u) \leq \epsilon$. Given $f(u) \leq \epsilon$, we obtain the desired. ■

Lemma 13 *Let w be s.t. $f_{\log}(w) \leq \frac{1}{2n}$, then it holds that*

$$\left\| \frac{\nabla f_{\log}(w)}{f_{\log}(w)} - \frac{\nabla f_{\exp}(w)}{f_{\exp}(w)} \right\| \leq 3n\, f_{\log}(w).$$

Proof Define $u_i := \exp(-y_i \langle w, x_i \rangle)$. Note that for all finite w, $u_i > 0$. Using the expressions for the exponential and logistic losses,

$$\begin{aligned}
\frac{\nabla f_{\exp}(w)}{f_{\exp}(w)} &= -X^T p \quad \text{where } p \in \Delta_n \text{ s.t } p_i = \frac{u_i}{\sum_{j=1}^n u_j} \\
\frac{\nabla f_{\log}(w)}{f_{\log}(w)} &= -\frac{X^T q}{f_{\log}(w)} \quad \text{where } q \in \mathbb{R}^n \text{ s.t } q_i = \frac{1}{n}\frac{u_i}{1+u_i} \\
&= -X^T C p \quad \text{where } C \in \mathbb{R}^{n \times n} \text{ s.t } C_{i,i} = \frac{f_{\exp}(w)}{f_{\log}(w)}\frac{1}{1+u_i}
\end{aligned}$$

Using these relations,

$$\begin{aligned}
&\frac{\nabla f_{\log}(w)}{f_{\log}(w)} - \frac{\nabla f_{\exp}(w)}{f_{\exp}(w)} = X^T p - X^T C\, p = X^T \underbrace{(I_n - C)}_{:=D} p \\
\implies &\left\| \frac{\nabla f_{\log}(w)}{f_{\log}(w)} - \frac{\nabla f_{\exp}(w)}{f_{\exp}(w)} \right\| = \left\| X^T D p \right\| = \left\| \sum_i D_{i,i}\, p_i\, x_i \right\| \leq \sum_i |D_{i,i}|\, p_i\, \|x_i\| \\
&\leq \sum_i |D_{i,i}|\, p_i \leq \max_j |D_{j,j}| \sum_i p_i = \max_j |D_{j,j}|
\end{aligned}$$

Hence, we have reduced the problem to bounding $\max_j |D_{j,j}| = \max_j \left| 1 - \frac{f_{\exp}(w)}{f_{\log}(w)} \frac{1}{1+u_j} \right|$.

Next, we derive a uniform upper-bound on $\left| 1 - \frac{f_{\exp}(w)}{f_{\log}(w)} \frac{1}{1+u_i} \right|$. Since $u_i \geq 0$, $\ln(1+u_i) \leq u_i \implies f_{\log}(w) \leq f_{\exp}(w) \implies \frac{f_{\exp}(w)}{f_{\log}(w)} \geq 1$. This establishes a lower-bound on $\frac{f_{\exp}(w)}{f_{\log}(w)}$. Furthermore, we know that, $\ln(1+u_i) \geq \frac{u_i}{1+u_i}$. Using this relation,

$$f_{\log}(w) \geq \frac{1}{n} \sum_{i=1}^n \frac{u_i}{1+u_i} \implies n f_{\log}(w) \geq \sum_{i=1}^n \frac{u_i}{1+u_i} \implies \forall i, \quad \frac{u_i}{1+u_i} \leq n\, f_{\log}(w)$$

Now, we will use the fact that $f_{\log}(w) \leq \frac{1}{2n}$, $n f_{\log}(w) < 1 \implies 1 - n\, f_{\log}(w) > 0$. Manipulating the above expression, we get that, $\forall i$,

$$u_i \leq \frac{n\, f_{\log}(w)}{1 - n\, f_{\log}(w)} \tag{12}$$

This gives an upper-bound on u_i. Next, we manipulate the above expression to get an upper-bound on $\frac{f_{\exp}(w)}{f_{\log}(w)}$. Recall that,

$$nf_{\log}(w) \geq \sum_{i=1}^{n} \frac{u_i}{1+u_i} \geq \sum_{i=1}^{n} \frac{u_i}{1+\max_j u_j} = \frac{1}{1+\max_j u_j} \sum_{i=1}^{n} u_i = \frac{n\, f_{\exp}(w)}{1+\max_j u_j} \tag{13}$$

$$\implies \frac{f_{\exp}(w)}{f_{\log}(w)} \leq 1 + \max_j u_j \leq 1 + \frac{n\, f_{\log}(w)}{1 - n\, f_{\log}(w)} \qquad \text{(Using the upper-bound in (12))}$$

$$\implies \frac{f_{\exp}(w)}{f_{\log}(w)} \leq \frac{1}{1 - n\, f_{\log}(w)} \tag{14}$$

This establishes an upper-bound on $\frac{f_{\exp}(w)}{f_{\log}(w)}$. We will now use the above expressions to bound $\left|1 - \frac{f_{\exp}(w)}{f_{\log}(w)} \frac{1}{1+u_i}\right|$:

$$\begin{aligned}
\left|1 - \frac{f_{\exp}(w)}{f_{\log}(w)} \frac{1}{1+u_i}\right| &= \left|\left(\frac{f_{\exp}(w)}{f_{\log}(w)} - 1\right) \frac{1}{1+u_i} + \frac{1}{1+u_i} - 1\right| \\
&\leq \left|\left(\frac{f_{\exp}(w)}{f_{\log}(w)} - 1\right) \frac{1}{1+u_i}\right| + \left|\frac{-u_i}{1+u_i}\right| && \text{(Triangle inequality)} \\
&= \left|\left(\frac{f_{\exp}(w)}{f_{\log}(w)} - 1\right)\right| \frac{1}{1+u_i} + \frac{u_i}{1+u_i} && \text{(Since } u_i > 0\text{)} \\
&\leq \left|\left(\frac{f_{\exp}(w)}{f_{\log}(w)} - 1\right)\right| + \frac{u_i}{1+u_i} && \text{(Since } 1+u_i > 1\text{)} \\
&= \left(\frac{f_{\exp}(w)}{f_{\log}(w)} - 1\right) + \frac{u_i}{1+u_i} && \text{(Since } f_{\exp}(w) \geq f_{\log}(w)\text{)} \\
&\leq \frac{n\, f_{\log}(w)}{1 - n\, f_{\log}(w)} + n\, f_{\log}(w) && \text{(Using (14) and (12))} \\
\implies \left|1 - \frac{f_{\exp}(w)}{f_{\log}(w)} \frac{1}{1+u_i}\right| &\leq 3n\, f_{\log}(w) && \text{(Since } f_{\log}(w) \leq \tfrac{1}{2n}\text{)}
\end{aligned}$$

$$\implies \max_j |D_{j,j}| \leq 3n\, f_{\log}(w) \implies \left\| \frac{\nabla f_{\log}(w)}{f_{\log}(w)} - \frac{\nabla f_{\exp}(w)}{f_{\exp}(w)} \right\| \leq 3n\, f_{\log}(w).$$

■

Theorem 14 *Suppose that* $T \geq \frac{4\Phi}{\gamma^2}$ *where* $\Phi = \Theta\big([\log n]^2\big)$. *Set* $\eta_{\max}^t = \frac{1}{f_{\log}(t)}$. `GD-LS` *and* `GD-Polyak` *achieve the following rate for logistic loss:*

$$\min_{i\in[n]} \frac{y_i \langle \theta_T, x_i \rangle}{\|\theta_T\|} \geq \gamma - \Theta\big(\frac{[\log n]^2}{T}\big).$$

Proof From Lemma 11 and 12, if we set $C := \min\left\{\frac{2c-1}{c} \min\{\frac{1}{\lambda_1}, 1\}, \frac{c-1}{c^2\nu^2}\right\}$ where $\lambda_1 = 3\,\frac{L_1(\nu+1)}{(1-c)}$, then it takes $T \geq \frac{1}{C\gamma^2}[\log(\frac{c\nu^2}{\epsilon})]^2$ iterations to ensure that $f(\theta_T) \leq 2\epsilon$ for both `GD-LS` and `GD-Polyak`. Note that we have used $c\nu^2 > 1$ given $c > 1$ and $\nu \geq 1$. Then, we set $\epsilon = \frac{1}{6n}$ to

conclude that after $T_0 := \frac{1}{\gamma^2 C'} [\log(6nc\nu^2)]^2$ iterations, $f_{\log}(\theta_t) \leq \frac{1}{3n}$ for all $t \geq T_0$. From Lemma 13, we obtain

$$\|\epsilon_t\| = \left\| \frac{\nabla f_{\log}(\theta_t)}{f_{\log}(\theta_t)} - \frac{\nabla f_{\exp}(\theta_t)}{f_{\exp}(\theta_t)} \right\| \leq 3n\, f_{\log}(\theta_t) \leq 1,$$

which implies that $\|\epsilon_t\|_2^2 + 4\,\|\epsilon_t\| \leq 5\,\|\epsilon_t\| \leq 15n\,[f_{\log}(\theta_t)]$. With this choice of T_0, we have that for any $t \geq T_0$

$$c\nu^2 \exp(-\sqrt{\gamma^2 t C'}) \leq c\nu^2 \exp(-\sqrt{\gamma^2 T_0 C'}) = \frac{1}{6n} < f(\theta_0).$$

Hence, for any $t \geq T_0$, we can set $\epsilon = c\nu^2 \exp(-\sqrt{\gamma^2 t C'})$, which implies that $f(\theta_t) \leq 2\epsilon = 2c\nu^2 \exp(-\sqrt{\gamma^2 t C'})$. Therefore, we conclude that

$$\|\epsilon_t\|_2^2 + 4\,\|\epsilon_t\| \leq 15n\,[f_{\log}(\theta_t)] \leq 30cn\nu^2 \exp(-\sqrt{\gamma^2 t C'}). \tag{15}$$

For $t < T_0$, we use (2) to obtain

$$\begin{aligned} \left\| \frac{\nabla f_{\log}(w)}{f_{\log}(w)} - \frac{\nabla f_{\exp}(w)}{f_{\exp}(w)} \right\| \leq \left\| \frac{\nabla f_{\log}(w)}{f_{\log}(w)} \right\| + \left\| \frac{\nabla f_{\exp}(w)}{f_{\exp}(w)} \right\| \leq 2\,\nu \\ \implies \|\epsilon_t\|_2^2 + 4\,\|\epsilon_t\| \leq 4\nu^2 + 8\nu \end{aligned} \tag{16}$$

Putting together the bounds in (15) and (16),

$$\begin{aligned} \frac{1}{2}\sum_{t=1}^{T} \left[\|\epsilon_t\|_2^2 + 4\,\|\epsilon_t\|\right] &\leq \sum_{t=1}^{T_0} \left[\|\epsilon_t\|_2^2 + 4\,\|\epsilon_t\|\right] + \sum_{t=T_0}^{T} \left[\|\epsilon_t\|_2^2 + 4\,\|\epsilon_t\|\right] \\ &\leq (4\nu^2 + 8\nu)\, T_0 + 30cn\nu^2 \sum_{t=T_0}^{T} \exp(-\gamma\sqrt{C'}\sqrt{t}) \\ &\leq \frac{4\nu^2 + 8\nu}{\gamma^2\, C'} [\log(6nc\nu^2)]^2 + \frac{30cn\nu^2}{1 - \exp(-\gamma\sqrt{C'})} \exp(-\gamma\sqrt{C'}\,\sqrt{T_0}) \\ &= \frac{4\nu^2 + 8\nu}{2\gamma^2\, C'} [\log(6nc\nu^2)]^2 + \frac{5}{2(1 - \exp(-\gamma\sqrt{C'}))} \end{aligned}$$

From Lemma 10, we obtain that (recall $\sum_{t=1}^{T} \alpha_t \geq \min\{1, \frac{1}{C(L_1,\nu,c)}\}T = \frac{1}{\max\{1, C(L_1,\nu,c)\}}T$)

$$\begin{aligned} \frac{\text{Regret}_T^p + \text{Regret}_T^w}{\sum_{t=1}^{T} \alpha_t} &\leq \frac{\ln(n) + \frac{1}{2}\sum_{t=1}^{T}[\|\epsilon_t\|_2^2 + \|\epsilon_t\|]}{\sum_{t=1}^{T} \alpha_t} \\ &\leq \underbrace{\max\{C(L_1,\nu,c), 1\}\big(\ln(n) + \frac{4\nu^2 + 8\nu}{\gamma^2\, C'}[\log(6nc\nu^2)]^2 + \frac{5}{1 - \exp(-\gamma\sqrt{C'})}\big)}_{:=\Phi} \frac{1}{T} = \frac{\Phi}{T}, \end{aligned}$$

where $C(L_1,\nu,c) = \frac{3L_1(\nu+1)}{1-c}$ for `GD-LS` and $C(L_1,\nu,c) = c\nu^2$ for `GD-Polyak`. Note that $\Phi = \Theta(\frac{[\log(n)]^2}{T})$. Using Wang et al. [26, Theorem 1], we have for $T \geq \frac{4\Phi}{\gamma^2}$

$$\tilde{\gamma}(\tilde{w}_T) = \tilde{\gamma}(\theta_T) \geq \gamma - \frac{4\Phi}{\gamma^2} = \gamma - \Theta(\frac{[\log(n)]^2}{T}).$$

■

Appendix E. Extensions to Multiclass Setting

We start by extending the game framework in Wang et al. [26] to the multiclass setting. First, note that cross-entropy satisfies the self-boundedness properties in (2) (see Vaswani and Babanezhad [24, Proposition 5.]). The multiclass margin of weight $W \in \mathbb{R}^{k\times d}$ is defined as (k is number of classes)

$$\gamma := \max_{\|W\|_F \le 1} \min_{i\in[n], c\neq y_i} (e_{y_i} - e_c)^T W x_i. \tag{17}$$

To extend the above framework, we let $X = [-\operatorname{vec}((e_{y_i} - e_c)x_i^T)-] \in \mathbb{R}^{n(k-1)\times kd}$. Then the following holds

$$\begin{aligned}\min_{i\in[n],c\neq y_i} (e_{y_i} - e_c)^T W x_i &= \min_{i\in[n],c\neq y_i} \operatorname{Tr}(W x_i (e_{y_i} - e_c)^T) \\ &= \min_{i\in[n],c\neq y_i} \operatorname{vec}((e_{y_i} - e_c)x_i^T)\operatorname{vec}(W) \\ &= \min_{p\in\triangle^{n(k-1)}} p^T X \operatorname{vec}(W).\end{aligned}$$

To get a lower bound on $\|\tilde{W}_T\|$ (as above), we let $(x, y) \in \{(x^{(i)}, y^{(i)})\}_{i=1}^n$, then it holds that

$$\min_{p\in\triangle^{n(k-1)}} p^T X \operatorname{vec}(\tilde{W}_T) \le \min_{c\neq y} (e_y - e_c)^T \tilde{W}_T x \le \|e_y - e_{c'}\|_2 \left\|\tilde{W}_T x\right\|_2 \le \sqrt{2}\left\|\tilde{W}_T\right\|_F.$$

Following the same approach as in the binary case, we obtain that

$$\left\|\tilde{W}_T\right\|_F \ge \frac{1}{\sqrt{2}} \min_{p\in\triangle^{n(k-1)}} p^T A \operatorname{vec}(\tilde{W}_T) \ge \frac{\gamma^2}{4\sqrt{2}} \sum_{t=1}^T \alpha_t$$

when $\frac{\gamma^2}{2}\sum_{t=1}^T \alpha_t \ge 2\left(\mathrm{Reg}_T^p + \mathrm{Reg}_T^W\right)$. This leads to

$$\frac{\min_{p\in\Delta^{n(k-1)}} p^\top A \operatorname{vec}(\widetilde{W}_T)}{\left\|\widetilde{W}_T\right\|_F} \ge \gamma - \frac{4\sqrt{2}\left(\mathrm{Reg}_T^p + \mathrm{Reg}_T^W\right)}{\gamma^2 \sum_{t=1}^T \alpha_t} = \gamma - \frac{4\sqrt{2}C_T}{\gamma^2}.$$

Hence, as in the binary case, we need to upper bound C_T to obtain a margin convergence rate. Below, we give a multiclass extension of Lemma 4. The proof essentially follows the same steps as the binary case. We start by recalling the definitions of multiclass exponential loss and cross-entropy loss:

$$\text{Exponential:} \qquad L_{exp}(\operatorname{vec}(W)) := \frac{1}{n}\sum_{i\in[n]}\sum_{c\neq y_i} \exp\big(-\operatorname{vec}((e_{y_i} - e_c)x_i^T)\operatorname{vec}(W)\big);$$

$$\text{Cross-entropy:} \qquad L_{\text{cross}}(\operatorname{vec}(W)) := \frac{1}{n}\sum_{i\in[n]} \log\Big(1 + \sum_{c\neq y_i} \exp\big(-\operatorname{vec}((e_{y_i} - e_c)x_i^T)\operatorname{vec}(W)\big)\Big).$$

Note through the above vectorization, for multiclass exponential loss, it still holds that

$$\frac{\nabla f_{\exp}(\tilde{W}_{t-1})}{f_{\exp}(\tilde{W}_{t-1})} = -X^T p_{t-1},$$

where $p_{t-1} \in \triangle^{n(k-1)}$. Given this, w-player and p-player can perform the same updates as the (binary) exponential loss case. For cross-entropy loss, we define ϵ_t as

$$\epsilon_t := \left\| \frac{\nabla L_{\text{cross}}(\tilde{W}_t)}{L_{\text{cross}}(\tilde{W}_t))} - \frac{\nabla L_{\exp}(\tilde{W}_t)}{L_{\exp}(\tilde{W}_t)} \right\|_F,$$

and analogous result to Lemma 4 can be proved. Note that the moves of w-player and p-player are the same as Lemma 2, with simplex $\triangle^n$ replaced by $\triangle^{n(k-1)}$ and data matrix replaced by the one above. The conclusion of Lemma 2 still hold following the same proof.

Lemma 15 *For all W s.t. $L_{cross}(W) \le \frac{1}{2n}$, it holds*

$$\left\| \frac{\nabla L_{cross}(W)}{L_{cross}(W))} - \frac{\nabla L_{\exp}(W)}{L_{\exp}(W)} \right\|_F \le 3\sqrt{2} n L_{cross}(W).$$

Proof Define $u_{ic} := \exp\big(-\operatorname{vec}((e_{y_i} - e_c)x_i^T)\operatorname{vec}(W)\big)$ and $p_{ic} := \frac{u_{ic}}{\sum_{i\in[n]}\sum_{c\neq y_i} u_{ic}}$ (where $c \neq y_i$), and further let $P \in \mathbb{R}^{n\times(k-1)}$ be s.t. $P[i,c] = p_{ic}$. Then, we have that

$$\begin{aligned}\frac{\nabla L_{\exp}(\operatorname{vec}(W))}{L_{\exp}(\operatorname{vec}(W))} &= -\sum_{i\in[n],c\neq y_i} \frac{u_{ic}}{\sum_{i\in[n],c\neq y_i} u_{ic}} \operatorname{vec}((e_{y_i} - e_c)h_i^T) \\ &= -\sum_{i\in[n],c\neq y_i} p_{ic} \operatorname{vec}((e_{y_i} - e_c)x_i^T) = -X^T \operatorname{vec}(P^T),\end{aligned}$$

where recall that $X = [-\operatorname{vec}((e_{y_i} - e_c)x_i^T)-] \in \mathbb{R}^{n(k-1)\times kd}$. Next, for the cross-entropy loss, we have that

$$\nabla L_{\text{cross}}(\operatorname{vec}(W)) = -\frac{1}{n}\sum_{i\in[n]}\sum_{c\neq y_i} \frac{u_{ic}}{1+u_{ic}} \operatorname{vec}((e_{y_i} - e_c)h_i^T) = -X^T \operatorname{vec}(Q^T),$$

where we have defined $Q \in \mathbb{R}^{n\times(k-1)}$ s.t. $Q[i,c] = q_{ic} := \frac{1}{n}\frac{u_{ic}}{1+u_{ic}}$. For a fixed $(j \in [n], p \in [k]\backslash y_j)$ pair. Note the following:

$$\sum_{i\in[n],c\neq y_i} \frac{u_{ic}}{\sum_{i\in[n],c\neq y_i} u_{ic}} \frac{(\frac{1}{n}\sum_{i\in[n],c\neq y_i} u_{ic})\mathbb{1}(i=j, c=p)}{1+u_{jp}} = \frac{1}{n}\frac{u_{jp}}{1+u_{jp}}.$$

Next, we define the matrix $V^{j,p} \in \mathbb{R}^{n\times(k-1)}$ (that corresponds to the (j,p) pair) to be $V^{j,p}[i,c] = \frac{L_{\exp}(W)}{1+u_{jp}}$ when $i = j, c = p$ and 0 otherwise. Then, we have that for all $(j \in [n], c \neq y_j)$

$$\operatorname{vec}((V^{j,p})^T)^T \operatorname{vec}(P^T) = Q[j,p].$$

If we define the matrix $C \in \mathbb{R}^{n(k-1)\times n(k-1)}$ to be $C = [-\operatorname{vec}((V^{i,c})^T)-]_{i\in[n],c\neq y_i}$, then we have that $C\operatorname{vec}(P^T) = \operatorname{vec}(Q^T)$. This leads to

$$\frac{\nabla L_{\text{cross}}(\operatorname{vec}(W))}{L_{\text{cross}}(\operatorname{vec}(W))} = -A^T \frac{C}{L_{\text{cross}}(\operatorname{vec}(W))} \operatorname{vec}(P^T) = -X^T\tilde{C}\operatorname{vec}(P^T).$$

where we let $\tilde{C} := \frac{C}{L_{\text{cross}}(\text{vec}(W))}$. Thus, we have the following

$$\begin{aligned}\left\|\frac{\nabla L_{\text{cross}}(\text{vec}(W))}{L_{\text{cross}}(\text{vec}(W))} - \frac{\nabla L_{\exp}(\text{vec}(W))}{L_{\exp}(\text{vec}(W))}\right\|_F &= \left\|A^T(I_{n(k-1)} - \tilde{C})\,\text{vec}(P^T)\right\|_F \\ &\leq \sum_{i\in[n],c\neq y_i} \left\|\text{vec}((e_{y_i} - e_c)x_i^T))\right\|_2 |1 - \frac{L_{\exp}(W)}{L_{\text{cross}}(W)}\frac{1}{1+u_{ic}}|p_{ic} \\ &\leq \sqrt{2}\max_{i\in[n],c\neq y_i} |1 - \frac{L_{\exp}(W)}{L_{\text{cross}}(W)}\frac{1}{1+u_{ic}}|,\end{aligned}$$

where we have used that $\|x_i\| \leq 1, \forall i \in [n]$ and $\sum_{i\in[n],c\neq y_i} p_{ic} = 1$. For any $i \in [n], c \neq y_i$, following the same steps as the binary case to obtain

$$|1 - \frac{L_{\exp}(W)}{L_{\text{cross}}(W)}\frac{1}{1+u_{ic}}| \leq |\frac{L_{\exp}(W)}{L_{\text{cross}}(W)} - 1| + \frac{u_{ic}}{1+u_{ic}}.$$

We denote $u_i := \sum_{c\neq y_i} u_{ic}$. Similar to the binary case, we have that for all $i \in [n]$

$$L_{\text{cross}}(\text{vec}(W)) \geq \frac{1}{n}\sum_{i=1}^{n}\frac{u_i}{1+u_i} \implies \frac{u_i}{1+u_i} \leq nL_{\text{cross}}(\text{vec}(W)) \implies u_i \leq \frac{nL_{\text{cross}}(\text{vec}(W))}{1 - nL_{\text{cross}}(\text{vec}(W))}, \tag{18}$$

where the first inequality is by $\log(1+x) \geq \frac{x}{1+x}$, and the second implication is by $1 - nL_{\text{cross}}(W) > 0$ (holds because of the assumption). Thus, we conclude that for all $i \in [n], c \neq y_i$, it holds that $u_{ic} \leq u_i \leq \frac{nL_{\text{cross}}(\text{vec}(W))}{1-nL_{\text{cross}}(\text{vec}(W))}$. From this, we can show that $\frac{L_{\exp}(W)}{L_{\text{cross}}(W)} \leq \frac{1}{1-nL_{\text{cross}}(W)}$ (following the same steps as the binary case). Moreover, it also holds that $\frac{L_{\exp}(W)}{L_{\text{cross}}(W)} \geq 1$ because of the inequality $\log(1+u_i) \leq u_i$. Putting these pieces together, we conclude (as in the binary case) that

$$\left\|\frac{\nabla L_{\text{cross}}(\text{vec}(W))}{L_{\text{cross}}(\text{vec}(W))} - \frac{\nabla L_{\exp}(\text{vec}(W))}{L_{\exp}(\text{vec}(W))}\right\|_F \leq 3\sqrt{2}nL_{\text{cross}}(W),$$

for all W s.t. $L_{\text{cross}}(W) \leq \frac{1}{2n}$. Finally, note that

$$\left\|\frac{\nabla L_{\text{cross}}(W)}{L_{\text{cross}}(W))} - \frac{\nabla L_{\exp}(W)}{L_{\exp}(W)}\right\|_F = \left\|\frac{\nabla L_{\text{cross}}(\text{vec}(W))}{L_{\text{cross}}(\text{vec}(W))} - \frac{\nabla L_{\exp}(\text{vec}(W))}{L_{\exp}(\text{vec}(W))}\right\|_F.$$

■

The next step is to obtain results analogous to Lemma 11 and 12. We note that Lemma 7 still holds for cross-entropy loss [24]. To do this, for `GD-Polyak` as an example, we can define u^* as the solution to the max-margin problem (17). Similarly, let $u = \beta u^*$ where $\beta = \frac{1}{\gamma}\ln\frac{\max\{c\nu^2,1\}}{\epsilon}$.

Then we have that

$$
\begin{aligned}
L_{cross}(u) &= \frac{1}{n}\sum_{i\in[n]} \log\big(1+\sum_{c\neq y_i} \exp\big(-\operatorname{vec}((e_{y_i}-e_c)x_i^T)\operatorname{vec}(u)\big)\big) \\
&\leq \frac{1}{n}\sum_{i\in[n]} \exp\big(-\operatorname{vec}((e_{y_i}-e_c)x_i^T)\operatorname{vec}(u)\big)\big) \\
&= \frac{1}{n}\sum_{i\in[n]} \exp\big(-\beta\operatorname{vec}((e_{y_i}-e_c)x_i^T)\operatorname{vec}(u^*)\big)\big) \\
&\leq \exp(-\beta\gamma) \\
&= \frac{\epsilon}{\max\{c\nu^2,1\}}.
\end{aligned}
$$

The rest can follow the same steps as Lemma 11 and 12. With all these ingredients, we can conclude that the rate in Theorem 5 also applies to the cross-entropy loss.