THE IMPLICIT BIAS OF MIRROR DESCENT ON MULTICLASS SEPARABLE DATA: STOCHASTICITY AND MOMENTUM

Anonymous authors
Paper under double-blind review

ABSTRACT

The success of deep neural networks has spurred growing interest in understanding optimization dynamics and the solutions selected by different algorithms. In this paper, we study the implicit bias of normalized mirror descent (NMD) for multiclass linear classification with cross entropy loss, elucidating how stochasticity and momentum shape the selected direction. Our key insight is that mini-batch sampling can alter NMD's implicit bias, while momentum counteracts this effect. More specifically, under homogeneous mirror potentials, we show that full-batch NMD converges to the corresponding maximum-margin direction, irrespective of momentum. In contrast, under mini-batch sampling without momentum, NMD can lose sight of the full-batch maximum-margin direction. Crucially, momentum restores the full-batch implicit bias, enabling NMD to recover the same convergence direction. Experiments on linear models and neural networks support our theoretical findings.

1 INTRODUCTION

Deep neural networks, including large language models, have achieved strong performance across a wide range of tasks (LeCun et al., 1998; He et al., 2016; Brown et al., 2020; OpenAI, 2023; Grattafiori et al., 2024). These models are often highly overparameterized, yet can generalize well (Zhang et al., 2017; Belkin et al., 2019). One possible explanation is the implicit bias of optimization algorithms, which describes their preference for particular solutions and can affect generalization despite similar training losses (Gunasekar et al., 2018; Wilson et al., 2017).

A classical result in the theory of implicit bias is the max-margin convergence of gradient descent (GD). On separable linear data, its loss tends to zero while the iterates of GD diverge in norm and converge in direction to the Euclidean max-margin solution (Soudry et al., 2018; Ravi et al., 2024). For homogeneous neural networks, Lyu & Li (2020) show that the normalized margin of GD iterates converges to the value at a KKT point of a margin-maximization problem.

Beyond GD, studies of the implicit bias of mirror descent (MD) investigate how its mirror potential shapes the direction selected by the iterates on separable data. In binary linear classification, analyses of MD establish margin bounds (Li et al., 2021) and, for homogeneous mirror potentials, convergence in direction to the corresponding max-margin solution (Sun et al., 2022; 2023). Subsequent work sharpens margin convergence rates for MD and extends theoretical analysis to mirror flow with broader potentials (Wang et al., 2025; Pesme et al., 2024). Max-margin behavior has also been characterized for mirror flow in homogeneous neural networks and MD in softmax attention models, though only for binary classification (Jacobs & Montúfar, 2026; Julistiono et al., 2026). More broadly, existing implicit-bias analyses of MD on separable data focus on full-gradient updates or continuous-time mirror flow, whereas practical training commonly uses stochastic gradients and momentum. The convergence direction selected by MD for multiclass separable linear data also remains uncharacterized, even with full gradients. These gaps motivate the following questions:

*What is the **implicit bias** of **MD** on multiclass separable data?*
*How do **mini-batch sampling** and **momentum** affect it?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address these questions, we analyze normalized mirror descent (NMD) with and without momentum for multiclass separable linear classification with softmax cross entropy loss. We consider smooth homogeneous mirror potentials formed by combining a Frobenius smoothing term with base potentials such as entrywise ℓ_p, and Schatten-p potentials. We first establish convergence to the corresponding max-margin direction for NMD with full gradients and derive rates for the Bregman error and margin gap. We then analyze how mini-batches and momentum affect this convergence under independent sampling with replacement and random reshuffling.

Contributions. Our main contributions are as follows.

1. We prove that the normalized iterates of full-batch NMD converge to the max-margin direction induced by the mirror potential, with or without momentum. In both cases, the empirical cross entropy loss tends to zero. We also derive rates for the Bregman error and margin gap in terms of the potential's homogeneity degree and the step-size decay exponent. Momentum introduces a lag term that can weaken the bounds when step sizes decay slowly. The threshold depends on the homogeneity degree. Otherwise, the bounds have the same asymptotic order with and without momentum.
2. We show that mini-batch sampling can prevent NMD without momentum from selecting the max-margin direction, while momentum restores directional convergence and vanishing loss at every batch size. On orthogonal inputs with batch size one, the loss vanishes but the normalized iterates converge to an explicit direction distinct from the max-margin direction. For each batch size, we also construct bounded separable data on which the loss diverges and a positive margin gap persists. Both failures occur under uniform independent sampling with replacement and random reshuffling.
3. We evaluate theoretical predictions in separable linear models and neural networks. Under entrywise ℓ_4 and Schatten-4 potentials, mini-batch NMD with momentum exhibits decreasing Bregman error and margin gap under both sampling schemes. Two synthetic datasets show the predicted contrast in loss and margin trajectories with and without momentum. On a subset of MNIST, we observe that a two-layer ReLU network attains positive normalized margins earlier with momentum than without it under both mini-batch sampling schemes.

2 Related Work

Mirror descent. MD was introduced by Nemirovsky & Yudin (1983) as a first-order method for convex optimization. At each step, MD minimizes a linearized objective plus the Bregman divergence of a mirror potential, with the Euclidean quadratic potential yielding GD (Beck & Teboulle, 2003). For Lipschitz convex objectives and strongly convex mirror potentials, Beck & Teboulle (2003) establish convergence guarantees for MD. MD and its variants have been applied to online learning and high-dimensional empirical risk minimization (Duchi et al., 2011) and to sparse neural-network training (Bungert et al., 2022). Other applications include stochastic utility and stochastic max-flow problems (Nemirovski et al., 2009), equilibrium computation (Rakhlin & Sridharan, 2013), and policy optimization in reinforcement learning (Tomar et al., 2022; Yang et al., 2022).

Mirror geometry and max-margin selection. The implicit-bias theory of MD includes results on both the selection of finite interpolants and the limiting directions of iterates on separable data. In overparameterized linear models, Gunasekar et al. (2018) and Azizan & Hassibi (2019) show that, when MD or stochastic MD converges to an interpolant, its limit is the Bregman projection of the initialization onto the set of interpolants. Subsequent analyses establish exponential convergence of stochastic MD to the selected interpolant (Varma & Hassibi, 2025) and examine the implicit bias of entropic MD on linear systems (Malitsky & Posch, 2025). The selection of interpolating solutions by stochastic MD has also been examined in nonlinear models and models with matrix parameters (Azizan et al., 2022; Akhtiamov et al., 2026). On separable data, by contrast, logistic and exponential losses have no finite minimizer, so implicit bias concerns the directions of iterates and normalized margins. In binary linear classification, MD analyses establish margin guarantees and characterize potential-dependent max-margin directions (Li et al., 2021; Sun et al., 2022; 2023; Wang et al., 2025). Max-margin analyses also extend to mirror flow with broader potentials (Pesme et al., 2024), homogeneous neural networks trained by mirror flow (Jacobs & Montúfar, 2026), and

softmax attention optimized by MD (Julistiono et al., 2026). These studies leave open how mirror geometry determines the direction selected by full-gradient NMD on multiclass separable linear data. Our work characterizes this direction for a class of smooth homogeneous mirror potentials.

Stochasticity and momentum in implicit bias. The effects of mini-batch sampling and momentum on implicit bias depend on the algorithm and model. In binary linear classification on separable data, stochastic gradient descent (SGD) and its momentum variant converge in direction to the Euclidean max-margin solution under suitable conditions (Nacson et al., 2019b; Wang et al., 2022). For binary logistic regression on separable data, per-sample Adam can instead select a direction different from full-batch Adam (Baek et al., 2026). Continuous-time heavy-ball and Nesterov dynamics can also change the implicit bias of diagonal linear networks in regression (Lyu et al., 2025). In multiclass linear classification, Fan et al. (2025) establish max-margin convergence for full-gradient normalized steepest descent with and without momentum. Li et al. (2026) show that mini-batch steepest descent without momentum need not approach the max-margin direction. Under random reshuffling, their momentum analysis gives a margin guarantee subject to an effective-margin positivity condition but does not establish directional convergence. For NMD with a class of smooth homogeneous mirror potentials, we show that momentum restores convergence to the induced max-margin direction with mini-batches.

3 PRELIMINARIES

Notations. Let $\mathbb{R}_+ := [0, \infty)$ denote the set of nonnegative real numbers. For a positive integer m, we write $[m] := \{1, \ldots, m\}$. The ith coordinate of a vector $u \in \mathbb{R}^d$ is denoted by u_i, and $A_{i,j}$ denotes the (i,j)th entry of a matrix $A \in \mathbb{R}^{k\times d}$. The Euclidean inner product of vectors $u, v \in \mathbb{R}^d$ is denoted by $\langle u, v\rangle := u^\top v$. For matrices $A, B \in \mathbb{R}^{k\times d}$, we write $\langle A, B\rangle := \operatorname{tr}\left(A^\top B\right) = \sum_{i=1}^k \sum_{j=1}^d A_{i,j}B_{i,j}$ for their Frobenius inner product. The Euclidean and Frobenius norms are denoted by $\|u\|_2 := \left(\sum_{i=1}^d u_i^2\right)^{1/2}$ and $\|A\|_{\mathrm{F}} := \left(\sum_{i=1}^k \sum_{j=1}^d A_{i,j}^2\right)^{1/2}$, respectively. Let $\psi^*(\Theta) := \sup_W\{\langle\Theta, W\rangle - \psi(W)\}$ denote the Fenchel conjugate of ψ. The mirror potential ψ induces the norm $\|\cdot\|_\psi$, whose dual norm is defined by $\|Z\|_{\psi,*} := \sup_{\|W\|_\psi\le 1}\langle Z, W\rangle$ for $Z \in \mathbb{R}^{k\times d}$. The vector $e_l \in \mathbb{R}^k$ denotes the lth standard basis vector for $l \in [k]$, and $\mathbf{1} \in \mathbb{R}^k$ denotes the vector of ones. The probability simplex is denoted by $\Delta^{k-1} := \{p \in \mathbb{R}_+^k : \sum_{l=1}^k p_l = 1\}$. The softmax function $\sigma : \mathbb{R}^k \to \Delta^{k-1}$ is defined componentwise by $\sigma(z)_i := \frac{\mathrm{e}^{z_i}}{\sum_{j=1}^k \mathrm{e}^{z_j}}$ for $i \in [k]$. For functions $f \ge 0$ and $g > 0$, we write $f(t) = O(g(t))$ if $f(t) \le Cg(t)$, and $f(t) = \Theta(g(t))$ if $cg(t) \le f(t) \le Cg(t)$, for all sufficiently large t, where $c, C > 0$ are constants independent of t. And we write $f(t) = \widetilde{O}(g(t))$ if $f(t) = O(g(t)\log^r(t))$ for some $r \ge 0$. The notation $f(t) = o(g(t))$ means that $f(t)/g(t) \to 0$ as $t \to \infty$.

Setup. Consider a multiclass linear classification problem with $k \ge 3$ classes. Let $\mathcal{S} := \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times [k]$ denote a training set of n samples, where y_i denotes the class label associated with x_i. Let $f_W : \mathbb{R}^d \to \mathbb{R}^k$ denote the linear classifier $f_W(x) := Wx$ with weight matrix $W \in \mathbb{R}^{k\times d}$, where $f_W(x)_l$ denotes the score of class l. The classifier predicts a class with the highest score. The softmax output $\sigma(f_W(x)) \in \Delta^{k-1}$ is a probability vector whose lth component is the predicted probability that x belongs to class l.

We learn the weight matrix W by minimizing the empirical cross entropy loss $\mathcal{L}(W) := \frac{1}{n}\sum_{i=1}^n \ell_i(W) = -\frac{1}{n}\sum_{i=1}^n \log\left(\sigma(Wx_i)_{y_i}\right)$, where $\ell_i(W)$ denotes the cross entropy loss for the ith training sample. By the definition of σ, the loss can be rewritten as

$$\mathcal{L}(W) = \frac{1}{n}\sum_{i=1}^n \log\left(1 + \sum_{l\in[k],\, l\ne y_i} \exp\left(-(e_{y_i} - e_l)^\top Wx_i\right)\right). \tag{1}$$

Mirror potential and maximum margin. Throughout the paper, ψ denotes a mirror potential from the following class unless stated otherwise.

Definition 1. Let $L \geq 2$ and $\epsilon > 0$. We define the mirror potential ψ by

$$\psi(W) := \frac{1}{L}\left((L\Phi(W))^{2/L} + \epsilon \|W\|_{\mathrm{F}}^2\right)^{L/2}, \tag{2}$$

where the base potential $\Phi : \mathbb{R}^{k\times d} \to \mathbb{R}_+$ satisfies:

1. $\Phi \in C^2\left(\mathbb{R}^{k\times d} \setminus \{0\}\right)$ is convex.
2. $\Phi(W) > 0$ for every $W \neq 0$.
3. Φ is absolutely homogeneous of degree L, i.e., $\Phi(\alpha W) = |\alpha|^L \Phi(W)$ for all $\alpha \in \mathbb{R}$ and $W \in \mathbb{R}^{k\times d}$.

Similar conditions on the potential have been considered in Sun et al. (2022; 2023); Wang et al. (2025); Julistiono et al. (2026); Jacobs & Montúfar (2026). When $\epsilon = 0$, ψ reduces to the base potential Φ. An important class of base potentials is given by $\Phi(W) = \frac{1}{L}\|W\|^L$, where $\|\cdot\|$ denotes either the Schatten p norm or the entrywise ℓ_p norm for $2 \leq p < \infty$.

The potential ψ induces a norm[1] on $\mathbb{R}^{k\times d}$, defined by

$$\|W\|_\psi := \inf\{r > 0 : \psi(W/r) \leq 1\} = \psi(W)^{1/L}, \tag{3}$$

where the last equality follows from the absolute homogeneity of ψ. We define the maximum normalized margin on the training data $\mathcal{S}$ and its corresponding direction w.r.t. $\|\cdot\|_\psi$ as follows:

$$\begin{aligned} \widehat{\gamma} &:= \max_{\|W\|_\psi \leq 1} \min_{i\in[n], l\neq y_i} (e_{y_i} - e_l)^\top W x_i, \\ \widehat{W} &:= \arg\max_{\|W\|_\psi \leq 1} \min_{i\in[n], l\neq y_i} (e_{y_i} - e_l)^\top W x_i. \end{aligned} \tag{4}$$

Let $\gamma(V) := \frac{\min_{i\in[n], l\neq y_i}\left(e_{y_i} - e_l\right)^\top V x_i}{\|V\|_\psi}$ denote the normalized margin of f_V on $\mathcal{S}$ w.r.t. $\|\cdot\|_\psi$, where $V \in \mathbb{R}^{k\times d} \setminus \{0\}$. To measure the deviation of a normalized weight matrix from the max-margin direction $\widehat{W}$, we introduce the Bregman divergence associated with ψ. For $A, B \in \mathbb{R}^{k\times d}$, this divergence is defined by

$$D_\psi(A, B) := \psi(A) - \psi(B) - \langle \nabla\psi(B), A - B\rangle. \tag{5}$$

By strict convexity of ψ, we have $D_\psi(A, B) \geq 0$, with equality if and only if $A = B$.

Mirror descent. Mirror descent is a generalization of gradient descent in which the Euclidean quadratic penalty is replaced by the Bregman divergence D_ψ (Nemirovsky & Yudin, 1983). With step size $\eta_t > 0$, the update takes the form

$$W_{t+1} = \arg\min_{W\in\mathbb{R}^{k\times d}} \{\eta_t \langle \nabla\mathcal{L}(W_t), W\rangle + D_\psi(W, W_t)\}.$$

Equivalently, it can be written as $\nabla\psi(W_{t+1}) = \nabla\psi(W_t) - \eta_t \nabla\mathcal{L}(W_t)$. We study NMD and its stochastic variants with mini batches. Fix a batch size $s \geq 1$. At iteration t, we select a batch $\mathcal{B}_t = (i_1, \ldots, i_s)$, where $i_j \in [n]$ for every $j \in [s]$. We compute the stochastic gradient as $\widehat{g}_t := \widehat{g}_{\mathcal{B}_t}(W_t)$, where $\widehat{g}_{\mathcal{B}_t}(W) := \frac{1}{s}\sum_{j=1}^s \nabla\ell_{i_j}(W)$. Our analysis covers updates with and without momentum. The updates of NMD without and with momentum are

$$\text{NMD:} \qquad \nabla\psi(W_{t+1}) = \nabla\psi(W_t) - \eta_t \widehat{g}_t / \|\widehat{g}_t\|_{\psi,*}, \tag{6}$$

$$\begin{aligned} \text{NMD with momentum:} \qquad & M_t = \beta_t M_{t-1} + (1-\beta_t)\widehat{g}_t, \\ & \nabla\psi(W_{t+1}) = \nabla\psi(W_t) - \eta_t M_t / \|M_t\|_{\psi,*}, \end{aligned} \tag{7}$$

where M_t denotes the momentum with initialization $M_{-1} = 0$, and $\beta_t \in [0, 1)$ is the momentum parameter. The momentum averages the raw stochastic gradients before normalization. When $\mathcal{B}_t = (1, \ldots, n)$, each sample appears once and $\widehat{g}_t = \nabla\mathcal{L}(W_t)$, so both methods reduce to their full batch counterparts.

[1]The functional $\|\cdot\|_\psi$ is a norm if $\{W : \psi(W) \leq 1\}$ is convex, bounded, and symmetric, with 0 in its interior. These conditions hold for the potentials in Definition 1.

Assumptions. To characterize the implicit bias of NMD, we impose the following conditions on the data, step size, and sampling. We first assume that the training data are linearly separable. It is a standard assumption in the implicit bias literature (Soudry et al., 2018; Gunasekar et al., 2018; Nacson et al., 2019a; Ji & Telgarsky, 2021).

Assumption 1. *There exists* $W^\star \in \mathbb{R}^{k\times d}$ *such that* $\min_{i\in[n],\, l\neq y_i} \left(e_{y_i} - e_l\right)^\top W^\star x_i > 0$.

We also assume that the training data are bounded. Similar assumptions have been used in Zhang et al. (2024); Fan et al. (2025); Li et al. (2026).

Assumption 2. *There exists* $B > 0$ *such that* $\|x_i\|_1 \leq B$ *for every* $i \in [n]$.

Assumption 3. *We choose* $\eta_t = \eta_0 t^{-a}$ *for* $t \geq 1$*, with initial step size* $\eta_0 > 0$ *and* $0 < a \leq 1$.

The step size $\eta_t = \Theta\left(1/t^a\right)$ satisfies $\eta_t \to 0$ and $\sum_{t=0}^{\infty} \eta_t = \infty$, which allows us to control the remainder terms in the directional convergence analysis. These schedules have been employed in convergence and implicit bias analyses of mirror descent (Sun et al., 2023) and Adam (Zhang et al., 2024), as well as steepest descent and its variants (Li et al., 2026), which include SGD (Bottou et al., 2018) and Muon (Fan et al., 2025).

In this work, we consider two schemes for forming mini batches of size s: *independent sampling with replacement* and *random reshuffling*.

Assumption 4. *The batch* $\mathcal{B}_t$ *consists of* s *indices sampled independently with replacement from* $[n]$ *according to a probability vector* $p = (p_1, \ldots, p_n)$ *satisfying* $\sum_{i=1}^n p_i = 1$ *and* $p_i > 0$ *for all* $i \in [n]$*. The batches are independent across iterations.*

Uniform sampling is a special case of Assumption 4, obtained by setting $p_i = 1/n$ for all $i \in [n]$.

The following assumption requires independent random reshuffling of the training samples at each epoch (Nagaraj et al., 2019; Mishchenko et al., 2020; Gürbüzbalaban et al., 2021).

Assumption 5. *We assume that* $n = ms$*, where* s *is the batch size and* m *is the number of batches per epoch. For each epoch* $r \geq 0$*, we draw a permutation* Π_r *of* $[n]$ *uniformly at random and divide the permuted indices into* m *consecutive batches* $\mathcal{B}_{rm}, \ldots, \mathcal{B}_{(r+1)m-1}$*, each of size* s*. The permutations are independent across epochs.*

Note that the m batches of epoch r are processed in order at iterations $t = rm + j$, $j = 0, \ldots, m-1$. The complete pseudocode for both sampling schemes and updates is given in Algorithm B.1.

4 Main Results

In this section, we compare the implicit bias of NMD with and without momentum under full-gradient and mini-batch updates.

4.1 Implicit Bias of NMD without Momentum

We first prove that full-gradient NMD converges to the max-margin direction, with rates for the Bregman error and margin gap in Theorem 1. For stochastic NMD, Theorem 2 constructs bounded separable data on which mini-batches have a persistent margin gap and diverging loss. Proposition 1 identifies the limiting direction on orthogonal data with batch size one, which can differ from the max-margin direction even as the loss vanishes. The detailed proofs are deferred to Appendix C.

Theorem 1. *Suppose that Assumptions 1 to 3 hold. Recall that* $L \geq 2$ *is the homogeneity degree of* ψ*, and* $a \in (0, 1]$ *is the decay exponent in* $\eta_t = \eta_0 t^{-a}$*. Let* $\{W_t\}_{t\geq 0}$ *be the iterates generated by NMD* (6) *with arbitrary initialization* $W_0 \in \mathbb{R}^{k\times d}$ *and* $\widehat{g}_t = \nabla \mathcal{L}\left(W_t\right)$ *for every* $t \geq 0$*. Then the Bregman error and margin gap converge with the following rates:*

$$D_\psi\left(\widehat{W}, W_T / \left\|W_T\right\|_\psi\right) = O(r\left(T\right)), \qquad \widehat{\gamma} - \gamma\left(W_T\right) = O\left(\sqrt{r\left(T\right)}\right), \tag{8}$$

where

$$r\left(T\right) := \begin{cases} \left(\log T\right)/T^{1-a}, & L = 2,\ 0 < a < 1, \\ \left(\log\log T\right)/\log T, & L = 2,\ a = 1, \\ T^{-(1-a)/(L-1)}, & L > 2,\ 0 < a < 1, \\ \left(\log T\right)^{-1/(L-1)}, & L > 2,\ a = 1. \end{cases}$$

Therefore, NMD converges to the max-margin direction, with $W_T / \|W_T\|_\psi \to \widehat{W}$ *and* $\gamma(W_T) \to \widehat{\gamma}$. *Moreover, we have* $\|W_T\|_\psi \to \infty$, *which implies* $\mathcal{L}(W_T) \to 0$.

Theorem 1 shows that full-gradient NMD converges to the max-margin direction w.r.t. $\|\cdot\|_\psi$ and drives the cross-entropy loss to zero. When $0 < a < 1$, the Bregman error decays polynomially in T, up to a logarithmic factor when $L = 2$, and the margin-gap bound is the square root of the Bregman bound. When $a = 1$, both bounds decay logarithmically.

Comparison with mirror descent in the binary setting and steepest descent. Sun et al. (2023) show that the Bregman error of normalized MD with $\eta_t = \eta_0/\sqrt{t+1}$ decays as $O\left(T^{-1/2+\zeta}\right)$ for arbitrarily small $\zeta > 0$ in binary classification with exponential loss, limiting their stated theorem to initializations with sufficiently small loss while leaving $L \geq \left(3+\sqrt{5}\right)/2$ uncovered. Compared with this, Theorem 1 covers the more practical setting of multiclass classification with cross-entropy loss, allows arbitrary initialization, and applies to every $L \geq 2$. Fan et al. (2025) characterize the implicit bias of normalized steepest descent for multiclass classification. When $L = 2$ and $\Phi(W) = \frac{1}{2}\|W\|_{\mathrm{F}}^2$, NMD (6) reduces to normalized gradient descent, which coincides with the normalized steepest descent w.r.t. the Frobenius norm. Theorem 1 therefore extends their Frobenius max-margin convergence result to NMD with mirror potentials outside the scope of their analysis.

Mini-batch updates are common in large-scale training since each step processes only a subset of the data, reducing the computational cost per update (Bottou et al., 2018). This motivates asking whether the max-margin convergence in Theorem 1 persists under mini-batch sampling. However, the next theorem shows that, for every batch size, a bounded linearly separable training set exists on which the normalized margin stays bounded away from $\widehat{\gamma}$ and the training loss diverges.

Theorem 2. *Suppose that* $d \geq k \geq 3$. *There exist constants* $c \geq 0$ *and* $\delta > 0$ *for which the following holds. For every batch size* $s \geq 1$, *integer* $n \geq cs$, *and sampling distribution* p *in Assumption 4, there is a training set* $\mathcal{S}_1$ *of* n *samples such that Assumptions 1 and 2 hold. Let* $\{W_t\}_{t\geq 0}$ *denote the iterates generated by NMD* (6) *on* $\mathcal{S}_1$, *with arbitrary initialization* $W_0 \in \mathbb{R}^{k\times d}$, *step sizes* $\{\eta_t\}_{t\geq 0}$ *satisfying Assumption 3, and mini-batches of size* s *sampled with replacement according to* p. *Then, almost surely, we have*

$$\widehat{\gamma} - \gamma(W_T) \geq \delta > 0 \quad \textit{for all sufficiently large } T, \qquad \mathcal{L}(W_T) \longrightarrow \infty. \tag{9}$$

The same conclusions also hold under random reshuffling in Assumption 5.

For $n \geq cs$, Theorem 2 provides a counterexample to max-margin convergence in which NMD without momentum retains a positive margin gap and its training loss diverges. The construction and proof for both sampling schemes appear in Section C.2.

We conclude this subsection with an example involving orthogonal class inputs in which NMD with batch size one drives the training loss to zero but its normalized iterates converge to a direction different from the max-margin direction.

Potential selection and data construction. Let $d = k \geq 3$, and suppose that the mirror potential ψ satisfies Definition 1 and is invariant under independent permutations within each column, i.e., $\psi([\widetilde{w}_1, \ldots, \widetilde{w}_k]) = \psi([w_1, \ldots, w_k])$ for all $w_1, \ldots, w_k \in \mathbb{R}^k$, where $\widetilde{w}_l$ denotes an arbitrary permutation of the coordinates of w_l. We consider a training set $\mathcal{S}_2 := \{(\alpha_i e_{y_i}, y_i)\}_{i=1}^n$, where $\alpha_i > 0$ and $y_i \in [k]$. Let $n_l := |\{i \in [n] : y_i = l\}| > 0$ for $l \in [k]$. We define

$$\bar{W} := \frac{\nabla\psi^*(G)}{\|\nabla\psi^*(G)\|_\psi},$$

where $G := \sum_{l=1}^k (n_l/n) R_l / \|R_l\|_{\psi,*}$ and $R_l := \left(e_l - k^{-1}\mathbf{1}\right) e_l^\top / \sqrt{(k-1)/k}$ for $l \in [k]$. Let $h_l := \min_{j\in[k],\, j\neq l} (e_l - e_j)^\top \bar{W} e_l$ for $l \in [k]$.

Proposition 1. *Consider NMD with the potential* ψ *and training set* $\mathcal{S}_2$ *specified above, initialization* $W_0 = 0$, *batch size* $s = 1$, *and step sizes* $\{\eta_t\}_{t\geq 0}$ *satisfying Assumption 3. Suppose that Assumption 4 holds with* $p_i = 1/n$ *for every* $i \in [n]$. *Then, almost surely, we have*

$$\frac{W_T}{\|W_T\|_\psi} \longrightarrow \bar{W}, \qquad \mathcal{L}(W_T) \longrightarrow 0. \tag{10}$$

If the values $\min_{i:y_i=1} \alpha_i h_1, \ldots, \min_{i:y_i=k} \alpha_i h_k$ *are not all equal, then* $\bar{W} \neq \widehat{W}$*. The same conclusions also hold under random reshuffling in Assumption 5.*

Potential classes and the SGD case. The conditions imposed on ψ above are satisfied by Frobenius radial potentials, smoothed entrywise ℓ_p potentials with $2 < p < \infty$, and powers of quadratic forms with positive column weights. For $\psi(W) = \|W\|_{\mathrm{F}}^2$, Corollary C.1 recovers Theorem 4.10 in Li et al. (2026) for Normalized-SGD and also covers uniform independent sampling with replacement. Further details are given in Section C.3.

4.2 Implicit Bias of NMD with Momentum

In this subsection, we establish max-margin convergence for NMD with momentum using either full gradients or mini-batch gradients, together with rates for the Bregman error and margin gap. The stochastic analysis covers independent sampling with replacement and random reshuffling for every batch size. We also compare the reshuffling result with stochastic steepest descent. The detailed proofs are deferred to Appendix D.

We first consider NMD with momentum in the full-gradient setting.

Theorem 3. *Suppose that Assumptions 1 to 3 hold. Let* $\{W_t\}_{t\geq 0}$ *be the iterates generated by NMD with momentum* (7) *with arbitrary initialization* $W_0 \in \mathbb{R}^{k\times d}$*, full gradients* $\widehat{g}_t = \nabla \mathcal{L}(W_t)$ *for every* $t \geq 0$*, and a momentum parameter* $\beta \in (0,1)$*. Then the Bregman error and margin gap converge with the following rates:*

$$D_\psi\Big(\widehat{W}, W_T/\|W_T\|_\psi\Big) = O(q(T)), \qquad \widehat{\gamma} - \gamma(W_T) = O\Big(\sqrt{q(T)}\Big), \tag{11}$$

where

$$q(T) := \begin{cases} (\log T)/T^{1-a} + T^{-a}, & L=2,\ 0<a<1, \\ (\log\log T)/\log T, & L=2,\ a=1, \\ T^{-\min\{1-a, L-2+a\}/(L-1)}, & L>2,\ 0<a<1, \\ (\log T)^{-1/(L-1)}, & L>2,\ a=1. \end{cases}$$

Therefore, NMD with momentum converges to the max-margin direction, with $W_T/\|W_T\|_\psi \to \widehat{W}$ *and* $\gamma(W_T) \to \widehat{\gamma}$*. Moreover, we have* $\|W_T\|_\psi \to \infty$*, which implies* $\mathcal{L}(W_T) \to 0$*.*

NMD with momentum converges in direction to $\widehat{W}$ and drives the loss to zero when full gradients are used, just as NMD without momentum does in Theorem 1. Momentum contributes a temporal-lag term to the Bregman-error bound. For $L = 2$, this term dominates when $0 < a < 1/2$, yielding a bound of $O(T^{-a})$ with momentum instead of $O\left((\log T)/T^{1-a}\right)$ without momentum. For $2 < L < 3$, it dominates only when $0 < a < (3-L)/2$, yielding $O\left(T^{-(L-2+a)/(L-1)}\right)$ instead of $O\left(T^{-(1-a)/(L-1)}\right)$. In all other cases, the Bregman-error and margin-gap bounds have the same asymptotic orders with and without momentum.

Mini-batch sampling can prevent NMD without momentum from converging to $\widehat{W}$, as shown in Proposition 1. By contrast, Theorem 4 establishes max-margin convergence and rate bounds for NMD with momentum at every admissible batch size for both sampling schemes.

Theorem 4. *Suppose that Assumptions 1 to 3 hold. Let* $\{W_t\}_{t\geq 0}$ *be the iterates generated by NMD with momentum* (7) *with arbitrary initialization* $W_0 \in \mathbb{R}^{k\times d}$*, mini-batch gradients* $\widehat{g}_t$ *of batch size* $s \geq 1$*, and a momentum parameter* $\beta \in (0,1)$*.*

(i) ***Independent sampling with replacement.*** *Suppose that Assumption 4 holds with* $n \geq 2$*. Let* $\chi := -s\log\left(1 - \min_{i\in[n]} p_i\right) > 0$*. For step size decay exponent* $0 < a < 1$ *and every* $0 < \kappa < \min\{\frac{1-a}{L-1}, \frac{\chi}{\chi+\log(1/\beta)} \min\{\frac{L-2+a}{L-1}, 1-a\}\}$*, we have, almost surely,*

$$D_\psi\Big(\widehat{W}, W_T/\|W_T\|_\psi\Big) = O\left(T^{-\kappa}\right), \qquad \widehat{\gamma} - \gamma(W_T) = O\left(T^{-\kappa/2}\right). \tag{12}$$

When $a = 1$*, the Bregman error and margin gap are almost surely* $O\left((\log T)^{-\kappa}\right)$ *and* $O\left((\log T)^{-\kappa/2}\right)$*, respectively, for every* $0 < \kappa < \min\{\frac{1}{L-1}, \frac{\chi}{\chi+\log(1/\beta)}\}$*.*

(ii) ***Random reshuffling.*** *Suppose that Assumption 5 holds. Then, for every sequence of permutations, we have*

$$D_\psi\Big(\widehat{W}, W_T/\left\|W_T\right\|_\psi\Big) = O(q(T)), \qquad \widehat{\gamma} - \gamma(W_T) = O\Big(\sqrt{q(T)}\Big), \tag{13}$$

where $q(T)$ is defined in Theorem 3.

For every $\beta \in (0,1)$ and batch size, Theorem 4 implies $W_T/\left\|W_T\right\|_\psi \to \widehat{W}$, $\gamma(W_T) \to \widehat{\gamma}$, and $\mathcal{L}(W_T) \to 0$. Random reshuffling covers the data in every epoch and admits rate bounds of the same order as full-gradient NMD with momentum. Independent sampling relies on probabilistic coverage, and its guaranteed exponent depends on the batch size, sampling probabilities, and β.

Effect of momentum. For the data in Proposition 1, the normalized iterates converge to $\bar{W}$ without momentum and to $\widehat{W}$ with momentum. On the data constructed in Theorem 2, the loss diverges without momentum but converges to zero with momentum. These examples illustrate the general guarantee of Theorem 4: with momentum, the normalized iterates converge to $\widehat{W}$ and the loss vanishes for both sampling schemes and all batch sizes.

Comparison with stochastic steepest descent. Li et al. (2026) study stochastic steepest descent with mini-batch gradients. For its momentum variant with random reshuffling, their Theorem 4.4 establishes an approximate max-margin guarantee, but does not prove directional convergence of the normalized iterates. For $\psi(W) = \|W\|_{\mathrm{F}}^2$, our update coincides with normalized stochastic gradient descent with momentum after a constant rescaling of the step size. In this shared setting, Theorem 4 proves $W_T/\left\|W_T\right\|_{\mathrm{F}} \to \widehat{W}$ for every $\beta \in (0,1)$, every batch size, and every sequence of epoch partitions, without the effective-margin positivity condition required by their Theorem 4.4. Except for the Frobenius update, the momentum methods they study fall outside our framework.

5 Numerical Experiments

We compare full-gradient and mini-batch NMD, with and without momentum, on synthetic and real data. All linear-model runs start from $W_0 = 0$ and use $\eta_t = 0.1/\sqrt{\max\{1,t\}}$. The clustered-data runs use 3×10^6 updates, and the three-class runs use 10^6 updates.

Synthetic experiments. Motivated by Fan et al. (2025), we generate datasets with $k = 10$ classes in $d = 25$ dimensions. For each dataset, we draw independent $z_l \sim \mathcal{N}(0, I_d)$ and define $\mu_l = z_l/\left\|z_l\right\|_2$ for $l \in [k]$. Each class contains 50 samples $x_{l,r} = \mu_l + \xi_{l,r}$, $r \in [50]$, where $\xi_{l,r}$ is uniform in the Euclidean ball of radius $\frac{1}{4}\min_{l \neq j}\left\|\mu_l - \mu_j\right\|_2$. We train with smoothed potentials based on the entrywise ℓ_4 and Schatten-4 norms. Full-gradient NMD uses $\beta \in \{0, 0.9\}$, while independent sampling with replacement and random reshuffling use $s = 10$ and $\beta = 0.9$. Figure 1 plots the relative margin gap and Bregman error for both potentials. Both errors decrease over the iterations for full-gradient NMD and for mini-batch NMD under either sampling scheme.

We next test our theoretical predictions for stochastic NMD with and without momentum on two separable three-class datasets. The orthogonal dataset contains $(a_l\,(1 + (r-1)/198)\,e_l, l)$ for $l \in [3]$ and $r \in [100]$, where $(a_1, a_2, a_3) = (1, 2, 4)$. With $s = 1$, NMD without momentum drives the loss toward zero, but its normalized margin approaches $1/\sqrt{2}$, below the max-margin value $4/\sqrt{14}$. Furthermore, the imbalanced dataset consists of $(e_1, 1)$, $(e_3, 3)$, and $(\lambda_r z, 2)$ for $r \in [30]$, where $z = \big(3/4, -\sqrt{7}/4, 0\big)^\top$ and $\lambda_r = 1/2 + (r-1)/58$. With $s = 4$, the margin of mini-batch NMD without momentum remains negative and the loss increases. Both behaviors appear under independent sampling and random reshuffling in Figure 2, illustrating Proposition 1 and Theorem 2. With $\beta = 0.9$, the losses decrease and the margins approach their max-margin values on both datasets.

Real-data experiments. We train both layers of a ReLU neural network with 100 hidden units on a subset of $n = 1000$ MNIST images (LeCun et al., 1998), containing 100 images from each class. Let $\theta_t := (W_{1,t}, W_{2,t})$ denote the pair of weight matrices at iteration t. The neural network model is defined by $f_{\theta_t}(x) := W_{2,t}\,\mathrm{ReLU}(W_{1,t}x)$, where $f_{\theta_t}(x)_l$ is the score of class l. The normalized

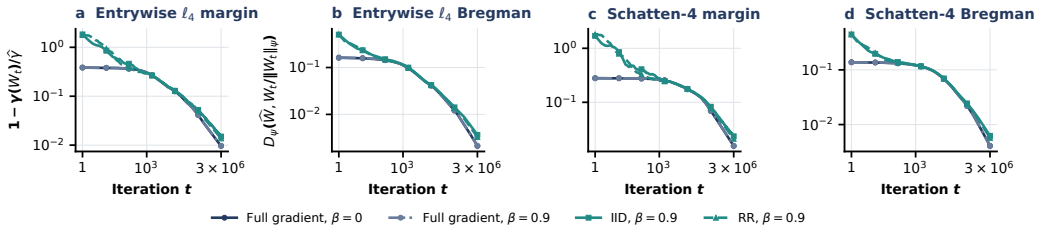

Figure 1: Relative margin gap and Bregman error under smoothed entrywise ℓ_4 and Schatten-4 potentials. Five data seeds generate five datasets. Full-gradient curves are medians across these datasets. For stochastic curves, we take the median of three independent sampling runs on each dataset before taking the median across datasets.

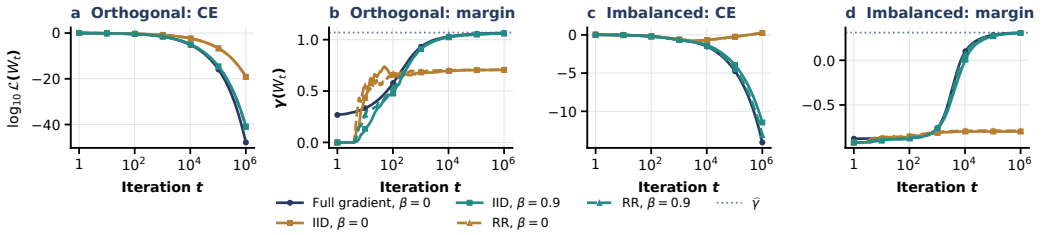

Figure 2: Empirical cross-entropy loss and normalized margin for NMD on the orthogonal data ($n = 300$, $s = 1$) and the imbalanced data ($n = 32$, $s = 4$). Panels (a,c) display the loss as $\log_{10} \mathcal{L}(W_t)$, whereas panels (b,d) display the margin. All runs use the Frobenius mirror potential $\psi(W) = \|W\|_{\mathrm{F}}^2$. Each stochastic curve is the median of ten runs with independent sampling randomness, while the full-gradient curve is deterministic.

margin of this network on the training set is $\gamma_{\mathrm{NN}}(\theta_t) := \min_{i \in [n],\, l \neq y_i} (e_{y_i} - e_l)^\top f_{\theta_t}(x_i) \,/\, \|\theta_t\|_2^2$, where $\|\theta_t\|_2^2 = \|W_{1,t}\|_{\mathrm{F}}^2 + \|W_{2,t}\|_{\mathrm{F}}^2$. We evaluate Euclidean NMD with $\beta \in \{0, 0.9\}$, using full gradients or mini-batches of size $s \in \{20, 100, 500\}$ generated by independent sampling with replacement or random reshuffling. Figure A.1 in the appendix shows the median normalized-margin trajectories across five paired runs per configuration. Under both sampling schemes, the median normalized margin for $s = 20$ becomes positive with $\beta = 0.9$, whereas it remains negative with $\beta = 0$ throughout the displayed horizon. This behavior, which shows that momentum helps small-batch NMD attain a positive normalized margin, is consistent with the theoretical analysis for linear models by Theorem 4.

6 Conclusion

We studied the implicit bias of NMD for separable multiclass linear classification under full and mini-batch gradients. Full-gradient NMD selects the max-margin direction induced by the mirror potential with or without momentum. Under mini-batch sampling without momentum, the loss can diverge or vanish while the iterates approach a non-max-margin direction. In contrast, momentum restores directional convergence and vanishing loss for every batch size under independent sampling with replacement and random reshuffling. We also establish convergence rates for the Bregman error and margin gap. Experiments on linear models illustrate these theoretical results, while the ReLU experiment suggests an advantage of momentum in attaining positive normalized margins. A future direction is to characterize stochastic NMD without momentum on general separable data and try to sharpen the convergence rates under independent sampling. Beyond this separable setting, extending the theory to nonseparable multiclass data and nonlinear models, including diagonal linear networks, self-attention models, and homogeneous neural networks (Lyu et al., 2025; Julistiono et al., 2026; Jacobs & Montúfar, 2026) remains an open challenge. Such an extension would establish whether momentum has a similar effect on implicit bias in nonseparable or nonlinearly parameterized settings. Finally, we will examine whether max-margin convergence persists as $\beta_t \to 0$ in stochastic NMD.

AI USE STATEMENT

In this work, we used generative AI tools to aid manuscript writing, including organization and language editing, and to retrieve and identify related work. We also consulted these tools while refining mathematical claims, proofs, and the experimental design. The authors reviewed all AI-assisted material, and take responsibility for the final manuscript.

REFERENCES

Danil Akhtiamov, Reza Ghane, Omead Pooladzandi, and Babak Hassibi. Implicit bias and convergence of matrix stochastic mirror descent. *arXiv preprint arXiv:2602.18997*, 2026. URL https://arxiv.org/abs/2602.18997.

Navid Azizan and Babak Hassibi. Stochastic gradient/mirror descent: Minimax optimality and implicit regularization. In *International Conference on Learning Representations*, 2019.

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, 2022. doi: 10.1109/TNNLS.2021.3087480.

Beomhan Baek, Minhak Song, and Chulhee Yun. Implicit bias of per-sample Adam on separable data: Departure from the full-batch regime. In *International Conference on Learning Representations*, 2026.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003. doi: 10.1016/S0167-6377(02)00231-6.

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. doi: 10.1073/pnas.1903070116.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Leon Bungert, Tim Roith, Daniel Tenbrinck, and Martin Burger. A bregman learning framework for sparse neural networks. *Journal of Machine Learning Research*, 23(192):1–43, 2022.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and Muon on multiclass separable data. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1832–1841, 2018.

Mert Gürbüzbalaban, Asuman Ozdaglar, and Pablo A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186:49–84, 2021. doi: 10.1007/s10107-019-01440-w.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Tom Jacobs and Guido Montúfar. Implicit bias of mirror flow in homogeneous neural networks: Sparse and dense feature learning. *arXiv preprint arXiv:2605.19458*, 2026. URL https://arxiv.org/abs/2605.19458.

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 772–804, 2021.

Addison Kristanto Julistiono, Davoud Ataee Tarzanagh, and Navid Azizan. Optimizing attention with mirror descent: Generalized max-margin token selection. *Journal of Machine Learning Research*, 27(36):1–61, 2026.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Jichu Li, Xuan Tang, and Difan Zou. The implicit bias of steepest descent with mini-batch stochastic gradient. In *International Conference on Machine Learning*, 2026.

Yan Li, Caleb Ju, Ethan X. Fang, and Tuo Zhao. Implicit regularization of bregman proximal point algorithm and mirror descent on separable data. *arXiv preprint arXiv:2108.06808*, 2021. URL https://arxiv.org/abs/2108.06808.

Bochen Lyu, He Wang, Zheng Wang, and Zhanxing Zhu. Effects of momentum in implicit bias of gradient flow for diagonal linear networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19242–19250, 2025. doi: 10.1609/aaai.v39i18.34118.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Yura Malitsky and Alexander Posch. Entropic mirror descent for linear systems: Polyak's stepsize and implicit bias. *arXiv preprint arXiv:2505.02614*, 2025. URL https://arxiv.org/abs/2505.02614.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3420–3428, 2019a.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3051–3059, 2019b.

Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. SGD without replacement: Sharper rates for general smooth convex functions. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4703–4711, 2019.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. doi: 10.1137/070704277.

A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, 1983.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

Scott Pesme, Radu-Alexandru Dragomir, and Nicolas Flammarion. Implicit bias of mirror flow on separable data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3624.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Hrithik Ravi, Clayton Scott, Daniel Soudry, and Yutong Wang. The implicit bias of gradient descent on separable multiclass data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2585.

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.

Haoyuan Sun, Kwangjun Ahn, Christos Thrampoulidis, and Navid Azizan. Mirror descent maximizes generalized margin and can be implemented efficiently. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. In *International Conference on Learning Representations*, 2022.

K. Nithin Varma and Babak Hassibi. Exponential convergence of stochastic mirror descent in over-parameterized linear models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025.

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? In *Advances in Neural Information Processing Systems*, volume 35, pp. 26764–26776, 2022.

Guanghui Wang, Zihao Hu, Claudio Gentile, Vidya Muthukumar, and Jacob Abernethy. Faster margin maximization rates for generic and adversarially robust optimization methods. *Mathematical Programming*, 2025. doi: 10.1007/s10107-025-02283-4.

Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Long Yang, Yu Zhang, Gang Zheng, Qian Zheng, Pengfei Li, Jianhang Huang, and Gang Pan. Policy optimization with stochastic mirror descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8823–8831, 2022. doi: 10.1609/aaai.v36i8.20863.

Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of Adam on separable data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 23988–24021, 2024. doi: 10.52202/079017-0756.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

APPENDIX

A Supplementary Experiments **13**

B Technical Foundations **13**

B.1 Unified algorithm 14

B.2 Loss and mirror geometry 14

B.3 Bregman estimates 18

C Proofs of the Results for NMD without Momentum **21**

C.1 Proof of Theorem 1 21

C.2 Proof of Theorem 2 22

C.3 Proof of Proposition 1 25

D Proofs of the Results for NMD with Momentum **30**

D.1 Proof of Theorem 3 30

D.2 Proof of Theorem 4 33

A SUPPLEMENTARY EXPERIMENTS

For the MNIST experiment in Section 5, we divide pixel values by 255 and flatten each image into a vector in $\mathbb{R}^{784}$. We initialize the two weight matrices independently as

$$(W_1)_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{2}{784}\right), \qquad (W_2)_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{2}{100}\right).$$

We train for 5×10^4 updates with $\eta_t = 0.5/\sqrt{\max\{1, t\}}$.

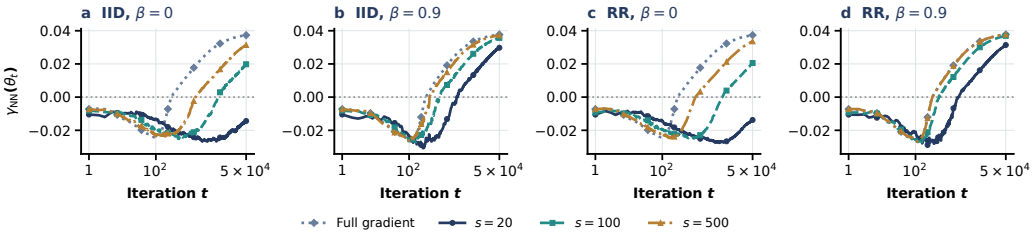

Figure A.1: Normalized margin $\gamma_{\mathrm{NN}}(\theta_t)$ for the two-layer ReLU network on MNIST. Panels (a,b) use independent sampling with replacement, and panels (c,d) use random reshuffling. Momentum is absent in (a,c) and set to $\beta = 0.9$ in (b,d). Each panel shows full-gradient NMD and mini-batches of size $s \in \{20, 100, 500\}$. Curves are medians over five paired runs.

B TECHNICAL FOUNDATIONS

We collect the algorithm, geometric estimates, and a comparison theorem for the Bregman error.

The following epoch formulation includes random reshuffling.

Assumption B.1. *A fixed batch size s divides n, and $m := n/s$. For every epoch $r \geq 0$, each batch $\mathcal{B}_{rm}, \ldots, \mathcal{B}_{rm+m-1}$ contains distinct indices, and the underlying index sets of these batches partition $[n]$. The partitions and their order may vary across epochs.*

B.1 UNIFIED ALGORITHM

Algorithm B.1 implements (6) and (7) under the sampling schemes in Assumptions 4 and 5. It averages raw batch gradients before normalizing the update direction. The choice $\beta_t \equiv 0$ recovers NMD without momentum.

Algorithm B.1 Normalized mirror descent with mini-batches and optional momentum

Require: Training data $\mathcal{S}$, potential ψ, initial weights W_0, updates T, batch size s, step sizes $\{\eta_t\}_{t\geq 0}$ and momentum parameters $\beta_t \in [0, 1)$

Require: Sampling switch mode $\in \{0, 1\}$: mode $= 0$ for independent sampling with replacement from p (Assumption 4), and mode $= 1$ for random reshuffling with $s \mid n$ (Assumption 5)

$\Theta_0 \leftarrow \nabla\psi(W_0)$, $M_{-1} \leftarrow 0$
if mode $= 1$ **then**
  $m \leftarrow n/s$
end if
for $t = 0, \ldots, T-1$ **do**
  if mode $= 0$ **then**
    Draw $i_1, \ldots, i_s$ independently from p ▷ independent sampling
    $\mathcal{B}_t \leftarrow (i_1, \ldots, i_s)$
  else
    if $t \bmod m = 0$ **then**
      $r \leftarrow t/m$
      Draw an independent uniform permutation Π_r of $[n]$ ▷ random reshuffling
      for $j = 0, \ldots, m-1$ **do**
        $\mathcal{B}_{rm+j} \leftarrow (\Pi_r(js+1), \ldots, \Pi_r((j+1)s))$
      end for
    end if
  end if
  Let $\mathcal{B}_t = (i_1, \ldots, i_s)$
  $\widehat{g}_t \leftarrow \frac{1}{s}\sum_{q=1}^{s} \nabla\ell_{i_q}(W_t)$
  $M_t \leftarrow \beta_t M_{t-1} + (1-\beta_t)\widehat{g}_t$
  $\widehat{b}_t \leftarrow M_t / \|M_t\|_{\psi,*}$
  $\Theta_{t+1} \leftarrow \Theta_t - \eta_t \widehat{b}_t$
  $W_{t+1} \leftarrow \nabla\psi^*(\Theta_{t+1})$
end for
return W_T

B.2 LOSS AND MIRROR GEOMETRY

Gradient representation and alignment. The pairwise score differences can be represented by fixed matrices. For $i \in [n]$ and $l \in [k]$, $l \neq y_i$, define

$$\begin{aligned} A_{il} &:= (e_{y_i} - e_l)x_i^\top \in \mathbb{R}^{k\times d}, \\ m_{il}(W) &:= (e_{y_i} - e_l)^\top W x_i = \langle A_{il}, W\rangle . \end{aligned} \tag{B.1}$$

We refer to A_{il} as a margin atom. Let $p_{il}(W) := \sigma(Wx_i)_l > 0$ denote the softmax probability of a competing class. By (1), we have

$$\begin{aligned} \ell_i(W) &= \log\left(1 + \sum_{l\neq y_i} \mathrm{e}^{-m_{il}(W)}\right), \\ -\nabla\mathcal{L}(W) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{l\neq y_i} p_{il}(W)A_{il}. \end{aligned} \tag{B.2}$$

Thus the negative gradient lies in the positive cone generated by the margin atoms. Its coefficient mass is denoted by

$$G(W) := \frac{1}{n}\sum_{i=1}^{n}\sum_{l\neq y_i} p_{il}(W). \tag{B.3}$$

To compare this mass with the gradient norm, let $C_A := \max_{i\in[n],\, l\neq y_i} \|A_{il}\|_{\psi,*}$. Equation (B.11) below gives explicit bounds on C_A for the smoothed potentials. The cone representation in (B.2) bounds the gradient above by $C_A G(W)$. Separability supplies a matching lower bound.

Lemma B.1. *Under Assumptions 1 and 2, the gradient in* (B.2) *satisfies*

$$\widehat{\gamma} G(W) \le \|\nabla\mathcal{L}(W)\|_{\psi,*} \le C_A G(W). \tag{B.4}$$

Consequently, $\nabla\mathcal{L}(W) \neq 0$ *for every finite* W*, and*

$$\left\langle \frac{\nabla\mathcal{L}(W)}{\|\nabla\mathcal{L}(W)\|_{\psi,*}}, \widehat{W} \right\rangle \le -\frac{\widehat{\gamma}}{C_A}. \tag{B.5}$$

Proof. Using (B.2) and the definition of C_A, we have

$$\|\nabla\mathcal{L}(W)\|_{\psi,*} \le \frac{1}{n}\sum_{i,l\neq y_i} p_{il}(W)\,\|A_{il}\|_{\psi,*} \le C_A G(W).$$

Since $\widehat{W}$ attains the maximum in (4), we also have

$$\|\nabla\mathcal{L}(W)\|_{\psi,*} \ge \left\langle -\nabla\mathcal{L}(W), \widehat{W} \right\rangle = \frac{1}{n}\sum_{i,l\neq y_i} p_{il}(W)\left\langle A_{il}, \widehat{W}\right\rangle \ge \widehat{\gamma} G(W).$$

Since $G(W) > 0$, the gradient does not vanish. Dividing the last display by $\|\nabla\mathcal{L}(W)\|_{\psi,*}$ and then applying the upper bound in (B.4), we obtain (B.5). □

Relative stability. For a batch $\mathcal{B} = (i_1, \dots, i_s)$, define

$$G_{\mathcal{B}}(W) := \frac{1}{s}\sum_{j=1}^{s}\sum_{l\neq y_{i_j}} p_{i_j l}(W).$$

Lemma B.2. *For all* W, Z*, let* $\delta := C_A\,\|Z - W\|_{\psi}$*. Then, for every* i *and* $l \neq y_i$*,*

$$\mathrm{e}^{-2\delta} p_{il}(W) \le p_{il}(Z) \le \mathrm{e}^{2\delta} p_{il}(W). \tag{B.6}$$

Consequently,

$$\mathrm{e}^{-2\delta} G(W) \le G(Z) \le \mathrm{e}^{2\delta} G(W), \tag{B.7}$$

$$\|\widehat{g}_{\mathcal{B}}(Z) - \widehat{g}_{\mathcal{B}}(W)\|_{\psi,*} \le C_A\big(\mathrm{e}^{2\delta} - 1\big) G_{\mathcal{B}}(W). \tag{B.8}$$

Proof. For a fixed sample, set

$$D_i(W) := 1 + \sum_{l'\neq y_i} \mathrm{e}^{-m_{il'}(W)}, \qquad p_{il}(W) = \frac{\mathrm{e}^{-m_{il}(W)}}{D_i(W)}.$$

By duality and the definition of C_A, we have $|m_{il'}(Z) - m_{il'}(W)| \le \delta$. Every term in $D_i(Z)$, including the constant term, is therefore between $\mathrm{e}^{-\delta}$ and e^{δ} times the corresponding term in $D_i(W)$. Thus

$$\mathrm{e}^{-\delta} \le \frac{D_i(Z)}{D_i(W)} \le \mathrm{e}^{\delta}.$$

Combining the numerator and denominator comparisons, we obtain (B.6). Summing this comparison, we obtain (B.7). Finally, $|p_{il}(Z) - p_{il}(W)| \le (\mathrm{e}^{2\delta} - 1)p_{il}(W)$. Substituting this inequality into the batch version of (B.2) and using $\|A_{il}\|_{\psi,*} \le C_A$, we obtain (B.8). □

Maximum-margin ray bound.

Lemma B.3. *Under Assumptions 1 and 2, set* $\lambda_{\mathrm{ce}} := \widehat{\gamma}^{-1}\log(2n(k-1))$. *For every* W, *with* $r := \|W\|_{\psi}$, *and every* $q \geq r + \lambda_{\mathrm{ce}}$,

$$\mathcal{L}(q\widehat{W}) \leq \mathcal{L}(W), \qquad \left\langle \frac{\nabla\mathcal{L}(W)}{\|\nabla\mathcal{L}(W)\|_{\psi,*}}, q\widehat{W} - W \right\rangle \leq 0. \tag{B.9}$$

For a weighted loss $\mathcal{L}_w(W) = n^{-1}\sum_i w_i\ell_i(W)$, *the same conclusion holds with its normalized gradient and* $\lambda_{\mathrm{ce}} = \widehat{\gamma}^{-1}\log(2n(k-1)w_+/w_-)$, *uniformly over* $0 < w_- \leq w_i \leq w_+$.

Proof. By $\|W\|_{\psi} = r$ and the definition of $\widehat{\gamma}$, there is a pair (i_0, l_0) such that $m_{i_0 l_0}(W) \leq r\widehat{\gamma}$. From (1), we have

$$\mathcal{L}(W) \geq \frac{1}{n}\log\bigl(1 + \mathrm{e}^{-r\widehat{\gamma}}\bigr).$$

Along the ray $q\widehat{W}$, every margin is at least $q\widehat{\gamma}$, and therefore

$$\mathcal{L}(q\widehat{W}) \leq (k-1)\mathrm{e}^{-q\widehat{\gamma}}.$$

Since $r \geq 0$, we have $0 < \mathrm{e}^{-r\widehat{\gamma}} \leq 1$. Using $\log(1+x) \geq x/2$ for $x \in [0,1]$, we obtain

$$\mathcal{L}(W) \geq \frac{\mathrm{e}^{-r\widehat{\gamma}}}{2n}, \qquad \mathcal{L}(q\widehat{W}) \leq (k-1)\mathrm{e}^{-r\widehat{\gamma}}\mathrm{e}^{-\lambda_{\mathrm{ce}}\widehat{\gamma}} \leq \frac{\mathrm{e}^{-r\widehat{\gamma}}}{2n}.$$

This proves the loss comparison for every radius. By convexity of $\mathcal{L}$, we have

$$\mathcal{L}(q\widehat{W}) \geq \mathcal{L}(W) + \left\langle \nabla\mathcal{L}(W), q\widehat{W} - W \right\rangle.$$

Combining the two inequalities gives (B.9). For the weighted loss, the lower and upper bounds become $w_-\mathrm{e}^{-r\widehat{\gamma}}/(2n)$ and $w_+(k-1)\mathrm{e}^{-q\widehat{\gamma}}$, respectively, giving the stated offset. □

Smoothed mirror potentials.

Proposition B.1. *Let* Φ *satisfy the conditions on the base potential in Definition 1. For* $\epsilon > 0$, *denote by* $\psi_{\Phi,\epsilon}$ *the potential in* (2). *Then* $\psi_{\Phi,\epsilon}$ *is positive definite, strictly convex, continuously differentiable, and absolutely homogeneous of degree L, with* $\nabla^2\psi_{\Phi,\epsilon}(W) \succ 0$ *for every* $W \neq 0$. *Both* $\psi_{\Phi,\epsilon}$ *and its conjugate are twice continuously differentiable away from the origin.*

Proof. We first write

$$R_\Phi(W) := (L\Phi(W))^{1/L}, \qquad A(W) := R_\Phi(W)^2 + \epsilon\|W\|_{\mathrm{F}}^2.$$

The unit sublevel set of R_Φ is convex, balanced, bounded, and absorbing. By absolute homogeneity, R_Φ is its gauge and therefore a norm. Hence R_Φ^2 is convex and twice continuously differentiable away from zero. In particular,

$$\nabla^2 A(W) \succeq 2\epsilon I, \qquad W \neq 0.$$

For $q(a) := L^{-1}a^{L/2}$, we have

$$q'(a) = \frac{1}{2}a^{(L-2)/2}, \qquad q''(a) = \frac{L-2}{4}a^{(L-4)/2} \geq 0.$$

Since $\psi_{\Phi,\epsilon} = q \circ A$, for every perturbation H we have

$$\begin{aligned}
&\bigl\langle H, \nabla^2\psi_{\Phi,\epsilon}(W)[H]\bigr\rangle_{\mathrm{F}} \\
&\quad = q''(A(W))\,\langle \nabla A(W), H\rangle_{\mathrm{F}}^2 + q'(A(W))\,\bigl\langle H, \nabla^2 A(W)[H]\bigr\rangle_{\mathrm{F}} \\
&\quad \geq \epsilon A(W)^{(L-2)/2}\,\|H\|_{\mathrm{F}}^2.
\end{aligned}$$

Because $L \geq 2$, homogeneity implies that the gradient tends to zero at the origin. Defining it there as zero, we obtain continuous differentiability on the whole space. From the displayed Hessian bound, we obtain strict convexity on segments that avoid zero. On a segment through zero, strict convexity

follows from $\psi_{\Phi,\epsilon}(cW) = |c|^L \psi_{\Phi,\epsilon}(W)$ and $L > 1$. Positive definiteness and absolute homogeneity follow directly from the definition. Moreover,

$$\psi_{\Phi,\epsilon}(W) \geq \frac{\epsilon^{L/2}}{L} \|W\|_{\mathrm{F}}^L,$$

so it is supercoercive. Together with strict convexity, this makes $\nabla\psi_{\Phi,\epsilon}$ bijective with inverse $\nabla\psi^*_{\Phi,\epsilon}$. The inverse function theorem and the positive Hessian bound imply that the conjugate is C^2 with positive definite Hessian away from zero. It is strictly convex, including on lines through zero by its homogeneity of degree $L/(L-1) > 1$. □

Bounds for margin atoms. The smoothed potential in (2) satisfies $\|W\|_\psi \geq L^{-1/L}\sqrt{\epsilon}\,\|W\|_{\mathrm{F}}$. By duality, we obtain

$$\Lambda_\psi := \sup_{\|Z\|_{\mathrm{F}}=1} \|Z\|_{\psi,*} \leq \frac{L^{1/L}}{\sqrt{\epsilon}}, \qquad \|Z\|_{\psi,*} \leq \Lambda_\psi \|Z\|_{\mathrm{F}}. \tag{B.10}$$

Since $\|A_{il}\|_{\mathrm{F}} = \sqrt{2}\,\|x_i\|_2 \leq \sqrt{2}B$, Assumption 2 and (B.10) imply

$$C_A \leq \sqrt{2}\Lambda_\psi B \leq L^{1/L}\sqrt{\frac{2}{\epsilon}}\,B. \tag{B.11}$$

For $\Phi(W) = L^{-1}\|W\|_{\mathrm{F}}^L$, the exact comparison constant is $\Lambda_\psi = L^{1/L}/\sqrt{1+\epsilon}$.

For the entrywise and Schatten examples $\Phi(W) = L^{-1}\|W\|_p^L$ and $\Phi(W) = L^{-1}\|W\|_{S_p}^L$, with $2 \leq p < \infty$, let $q = p/(p-1)$. Since smoothing increases the primal norm, we have, respectively,

$$\|A_{il}\|_{\psi,*} \leq L^{1/L}\|A_{il}\|_q = L^{1/L}2^{1/q}\|x_i\|_q \leq 2L^{1/L}B,$$
$$\|A_{il}\|_{\psi,*} \leq L^{1/L}\|A_{il}\|_{S_q} = L^{1/L}\sqrt{2}\,\|x_i\|_2 \leq \sqrt{2}L^{1/L}B.$$

The first identity uses the two nonzero rows of A_{il}, while the second uses its rank one structure. Both bounds are independent of ϵ.

Homogeneous mirror geometry and margin. Separability implies $\widehat{\gamma} > 0$, so every maximizer in (4) has unit norm. Strict convexity of ψ makes its unit ball strictly convex. If two distinct maximizers existed, their midpoint would have norm less than one, and rescaling it would increase its positive margin. Thus $\widehat{W}$ is unique, $\left\|\widehat{W}\right\|_\psi = 1$, and $\widehat{W} = \widehat{\gamma}W_{\mathrm{hm}}$, where

$$W_{\mathrm{hm}} := \arg\min_W \|W\|_\psi \quad \text{subject to} \quad \min_{i\in[n],l\neq y_i} (e_{y_i} - e_l)^\top W x_i \geq 1. \tag{B.12}$$

The following identities control radial growth and mirror steps.

Lemma B.4. *For ψ as in Definition 1, let $p := L/(L-1)$ and $f(\Theta) := \|\Theta\|_{\psi,*}$. Then*

$$\psi^*(\Theta) = (L-1)L^{-p}f(\Theta)^p, \qquad \|\nabla\psi(W)\|_{\psi,*} = L\|W\|_\psi^{L-1}. \tag{B.13}$$

There is a constant $C_\psi < \infty$, depending only on the potential, such that, for every $\Theta \neq 0$ and every Z,

$$\|\nabla\psi^*(\Theta)\|_\psi = L^{-1/(L-1)}\|\Theta\|_{\psi,*}^{1/(L-1)}, \tag{B.14}$$
$$\left\|\nabla^2\psi^*(\Theta)[Z]\right\|_\psi \leq C_\psi \|\Theta\|_{\psi,*}^{-(L-2)/(L-1)}\|Z\|_{\psi,*}. \tag{B.15}$$

For $\Theta \neq 0$, with $W = \nabla\psi^*(\Theta)$, we also have

$$\frac{W}{\|W\|_\psi} = \nabla f(\Theta), \qquad D_\psi\left(\widehat{W}, W/\|W\|_\psi\right) = L\left(1 - \frac{\left\langle\Theta, \widehat{W}\right\rangle}{f(\Theta)}\right). \tag{B.16}$$

Proof. The unit sublevel set of ψ is convex, centrally symmetric, and has nonempty interior. Its gauge is $\psi^{1/L}$, so this gauge is a norm. Writing $W = rV$ with $\|V\|_\psi = 1$, we have

$$\psi^*(\Theta) = \sup_{r\geq 0}\{rf(\Theta) - r^L\}.$$

The maximizing radius is $(f(\Theta)/L)^{1/(L-1)}$, which proves the first identity in (B.13). If $W \neq 0$, the norming functional $\nabla\|W\|_\psi$ has dual norm one. Differentiating $\psi(W) = \|W\|_\psi^L$, we obtain the second identity in (B.13). Since $\nabla\psi$ and $\nabla\psi^*$ are inverse maps, we obtain (B.14) exactly.

Twice differentiating the degree p homogeneity of ψ^*, we have

$$\nabla^2\psi^*(r\Theta)[Z] = r^{p-2}\nabla^2\psi^*(\Theta)[Z], \qquad r > 0.$$

The operator norm of $\nabla^2\psi^*$, measured from the dual norm to the primal norm, is continuous on the compact dual unit sphere. Its finite maximum and $p - 2 = -(L-2)/(L-1)$ prove (B.15).

Differentiating the first identity in (B.13), we obtain $W = (f(\Theta)/L)^{1/(L-1)}\nabla f(\Theta)$. By (B.14), this proves the first equality in (B.16). Finally, for $V = W/\|W\|_\psi$, Euler's identity and $\psi(\widehat{W}) = \psi(V) = 1$ imply

$$D_\psi\Big(\widehat{W}, V\Big) = L - \Big\langle \nabla\psi(V), \widehat{W}\Big\rangle = L - \frac{L\Big\langle \Theta, \widehat{W}\Big\rangle}{f(\Theta)},$$

where we used $\nabla\psi(V) = \Theta/\|W\|_\psi^{L-1}$ and (B.13). This proves (B.16). □

Lemma B.5. *Suppose that Assumptions 1 and 2 hold. There is a constant $K_\psi < \infty$, depending only on the potential, such that every V with $\|V\|_\psi = 1$ satisfies*

$$\Big\|\widehat{W} - V\Big\|_\psi \leq K_\psi\sqrt{D_\psi\Big(\widehat{W}, V\Big)}, \qquad 0 \leq \widehat{\gamma} - \gamma(V) \leq C_A K_\psi\sqrt{D_\psi\Big(\widehat{W}, V\Big)}. \tag{B.17}$$

Proof. Set $D = D_\psi\Big(\widehat{W}, V\Big)$ and $\theta = \nabla\psi(V)$. By (B.13), we have $\|\theta\|_{\psi,*} = L$. The dual identity (B.16) implies $0 \leq D \leq 2L$. Fix any A with $\|A\|_{\psi,*} \leq 1$. Applying Fenchel's inequality at $\widehat{W}$ and $\theta + hA$, and using $\psi(V) + \psi^*(\theta) = \langle\theta, V\rangle$, we obtain

$$h\Big\langle A, \widehat{W} - V\Big\rangle \leq D + D_{\psi^*}(\theta + hA, \theta) \leq D + \frac{M_\psi h^2}{2}, \qquad 0 < h \leq L/2, \tag{B.18}$$

where the second inequality follows by integrating (B.15) on the segment, whose dual norm is at least $L/2$. We may take $M_\psi = C_\psi(L/2)^{-(L-2)/(L-1)}$. For $D > 0$, choose $h = \sqrt{LD/8} \leq L/2$ in (B.18). Taking the supremum over the dual unit ball, we obtain the first inequality in (B.17), with $K_\psi = \sqrt{8/L} + M_\psi\sqrt{L/32}$. For $D = 0$, we obtain the same conclusion by letting $h \downarrow 0$. For every margin atom, we then have

$$\langle A_{il}, V\rangle \geq \Big\langle A_{il}, \widehat{W}\Big\rangle - C_A\Big\|\widehat{W} - V\Big\|_\psi \geq \widehat{\gamma} - C_A K_\psi\sqrt{D}.$$

Taking the minimum over (i, l), we obtain the upper bound on the margin gap in (B.17). By (4) and $\|V\|_\psi = 1$, we have $\gamma(V) \leq \widehat{\gamma}$, which proves its nonnegativity. □

B.3 Bregman Estimates

We define the cumulative step size, squared step size, and drift scale as

$$S_T := \sum_{t<T}\eta_t, \qquad Q_T := \sum_{t<T}\eta_t^2, \qquad \omega_t := \eta_t(1 + S_t)^{-\frac{L-2}{L-1}}.$$

We also define

$$\rho_L(S) := \begin{cases} \dfrac{\log(\mathrm{e} + S)}{1 + S}, & L = 2, \\ (1 + S)^{-\frac{1}{L-1}}, & L > 2. \end{cases}$$

Alignment, radial growth, and mirror steps.

Lemma B.6. *Suppose that Assumptions 1 and 2 hold. Let $M_{-1} = 0$ and let M_t follow the raw EMA in (7) with any $\beta_t \in [0, 1)$. Then $M_t \neq 0$ and*

$$\left\| \frac{M_t}{\|M_t\|_{\psi,*}} \right\|_{\psi,*} = 1, \qquad \left\langle \frac{M_t}{\|M_t\|_{\psi,*}}, \widehat{W} \right\rangle \leq -\frac{\widehat{\gamma}}{C_A}. \tag{B.19}$$

Proof. Writing $\alpha_t = 1 - \beta_t$ and unrolling the recursion, we have

$$M_t = \sum_{r=0}^{t} \alpha_r \prod_{v=r+1}^{t} \beta_v \, \widehat{g}_{\mathcal{B}_r}(W_r).$$

By (B.2), every summand is a negative combination of the atoms A_{il}. Hence, for some coefficients $h_{il,t} \geq 0$, not all zero,

$$M_t = -\sum_{i,l \neq y_i} h_{il,t} A_{il}.$$

By the definition of $\widehat{W}$ and the bound $\|A_{il}\|_{\psi,*} \leq C_A$, we have

$$\left\langle M_t, \widehat{W} \right\rangle \leq -\widehat{\gamma} \sum_{i,l \neq y_i} h_{il,t}, \qquad \|M_t\|_{\psi,*} \leq C_A \sum_{i,l \neq y_i} h_{il,t}.$$

Thus $M_t \neq 0$, and dividing these bounds gives (B.19). □

Lemma B.7. *Suppose that Assumptions 1 to 3 hold and consider*

$$\Theta_{t+1} = \Theta_t - \eta_t \widehat{b}_t, \qquad W_t = \nabla\psi^*(\Theta_t).$$

Assume that, for fixed constants $B_0, c > 0$,

$$\left\| \widehat{b}_t \right\|_{\psi,*} \leq B_0, \qquad \left\langle \widehat{b}_t, \widehat{W} \right\rangle \leq -c.$$

Set $r_t = \|W_t\|_\psi$ and $R_t = \max_{u \leq t} r_u$. Then, for all sufficiently large t,

$$c'(1 + S_t) \leq \|\Theta_t\|_{\psi,*} \leq C'(1 + S_t), \qquad r_t, R_t = \Theta\big((1 + S_t)^{1/(L-1)}\big). \tag{B.20}$$

Moreover,

$$\|W_{t+1} - W_t\|_\psi \leq C\eta_t (1 + S_t)^{-(L-2)/(L-1)} = C\omega_t, \tag{B.21}$$

$$D_\psi(W_t, W_{t+1}) = D_{\psi^*}(\Theta_{t+1}, \Theta_t) \leq C\eta_t^2 (1 + S_t)^{-(L-2)/(L-1)}. \tag{B.22}$$

For every fixed $\lambda > 0$, the test vector $Z_t = W_t/(r_t + \lambda) - \widehat{W}$ satisfies

$$\|Z_t\|_\psi \leq 2, \qquad \|Z_{t+1} - Z_t\|_\psi \leq C\frac{\eta_t}{1 + S_t}. \tag{B.23}$$

Proof. Summing the dual updates and taking the inner product with $\widehat{W}$, we have

$$\|\Theta_t\|_{\psi,*} \geq \left\langle \Theta_t, \widehat{W} \right\rangle \geq \left\langle \Theta_0, \widehat{W} \right\rangle + cS_t, \qquad \|\Theta_t\|_{\psi,*} \leq \|\Theta_0\|_{\psi,*} + B_0 S_t.$$

Since $S_t \to \infty$, we obtain the dual bounds in (B.20). Applying (B.14), we obtain the radius bounds, including the bound for R_t.

Write $\Delta\Theta_t = \Theta_{t+1} - \Theta_t$. By the reverse triangle inequality and (B.20), for every $v \in [0, 1]$,

$$\|\Theta_t + v\Delta\Theta_t\|_{\psi,*} \geq c'(1 + S_t) - B_0\eta_t \geq \frac{c'}{2}(1 + S_t)$$

for all sufficiently large t, where the last inequality uses $\eta_t/(1 + S_t) \to 0$. Integrating $\nabla^2\psi^*$ on this segment and applying (B.15), we obtain

$$\begin{aligned} \|W_{t+1} - W_t\|_\psi &\leq \int_0^1 \left\| \nabla^2\psi^*(\Theta_t + v\Delta\Theta_t)[\Delta\Theta_t] \right\|_\psi \, \mathrm{d}v \\ &\leq C(1 + S_t)^{-(L-2)/(L-1)} \|\Delta\Theta_t\|_{\psi,*}. \end{aligned}$$

Since $\|\Delta\Theta_t\|_{\psi,*} \le B_0\eta_t$, we obtain (B.21). The integral formula for the dual Bregman divergence also implies

$$\begin{aligned} D_{\psi^*}(\Theta_{t+1}, \Theta_t) &= \int_0^1 (1-v) \left\langle \Delta\Theta_t, \nabla^2\psi^*(\Theta_t + v\Delta\Theta_t)[\Delta\Theta_t] \right\rangle \, \mathrm{d}v \\ &\le C\eta_t^2 (1+S_t)^{-(L-2)/(L-1)}. \end{aligned}$$

By the conjugate Bregman identity, we obtain (B.22).

Finally, $\|Z_t\|_\psi \le r_t/(r_t+\lambda) + 1 \le 2$. By the reverse triangle inequality, $|r_{t+1} - r_t| \le \|W_{t+1} - W_t\|_\psi$. Combining (B.20) and (B.21), we obtain $r_{t+1}/r_t \to 1$ and

$$\begin{aligned} \|Z_{t+1} - Z_t\|_\psi &\le \frac{\|W_{t+1} - W_t\|_\psi}{r_{t+1}+\lambda} + \frac{r_t |r_{t+1} - r_t|}{(r_{t+1}+\lambda)(r_t+\lambda)} \\ &\le C\frac{\omega_t}{r_t} \le C\frac{\eta_t}{1+S_t}, \end{aligned}$$

where the final inequality uses (B.20) and the definition of ω_t. This proves (B.23). □

Cumulative directional error. Fix a ray offset $\lambda > 0$, and let

$$r_t := \|W_t\|_\psi, \qquad R_t := \max_{s \le t} r_s, \qquad Z_t := \frac{W_t}{r_t+\lambda} - \widehat{W}.$$

A reference direction b_t° is admissible whenever it satisfies the ray comparison

$$\left\langle b_t^\circ, (r_t+\lambda)\widehat{W} - W_t \right\rangle \le 0. \tag{B.24}$$

Write the actual direction as

$$\widehat{b}_t = b_t^\circ + q_t. \tag{B.25}$$

Define the cumulative directional error by

$$\mathcal{Q}_T^Z := \max_{1 \le s \le T} \left| \sum_{t<s} \eta_t \langle q_t, Z_t \rangle \right|. \tag{B.26}$$

Theorem B.1. *Fix $\lambda > 0$. Suppose that Assumptions 1 to 3 hold, that* (B.24)–(B.25) *hold, and consider*

$$\Theta_{t+1} = \Theta_t - \eta_t \widehat{b}_t, \qquad W_t = \nabla\psi^*(\Theta_t),$$

where $B_0, c_0 > 0$ are fixed, $\left\|\widehat{b}_t\right\|_{\psi,*} \le B_0$, and $\left\langle \widehat{b}_t, \widehat{W} \right\rangle \le -c_0$. Then

$$r_T, R_T = \Theta\big((1+S_T)^{1/(L-1)}\big).$$

Let

$$\Delta_T := \rho_L(S_T) + \frac{\mathcal{Q}_T^Z}{1+S_T}. \tag{B.27}$$

For every sufficiently large T,

$$D_\psi\Big(\widehat{W}, W_T/r_T\Big) \le C\Delta_T. \tag{B.28}$$

More precisely, writing $[x]_+ := \max\{x, 0\}$, we have

$$D_\psi\Big(\widehat{W}, W_T/r_T\Big) \le C\rho_L(S_T) + \frac{C}{1+S_T}\left[-\sum_{t<T} \eta_t \langle q_t, Z_t \rangle \right]_+. \tag{B.29}$$

Proof. By Lemma B.7 with $c = c_0$, the trajectory satisfies (B.20)–(B.23).

We now introduce the directional deficit

$$\mathcal{E}_t^{\mathrm{dir}} := L r_t^{L-1} - \left\langle \Theta_t, \widehat{W} \right\rangle. \tag{B.30}$$

For $r_t > 0$, Euler's identity and (5) imply the exact relation

$$\mathcal{E}_t^{\mathrm{dir}} = r_t^{L-1} D_\psi\Big(\widehat{W}, W_t/r_t\Big). \tag{B.31}$$

It remains to bound $\mathcal{E}_t^{\mathrm{dir}}$ by comparing the actual update with the admissible direction in (B.24).

By (B.24),

$$\Big\langle b_t^\circ, \widehat{W}\Big\rangle \le \frac{\langle b_t^\circ, W_t\rangle}{r_t + \lambda}.$$

For one mirror step, we have the identity

$$\eta_t \Big\langle \widehat{b}_t, W_t\Big\rangle = -(L-1)(r_{t+1}^L - r_t^L) + D_\psi(W_t, W_{t+1}). \tag{B.32}$$

Combining (B.32), (B.25), and (B.24), we obtain

$$\mathcal{E}_{t+1}^{\mathrm{dir}} - \mathcal{E}_t^{\mathrm{dir}} \le \mathcal{R}_t^{\mathrm{fs}} - \eta_t \langle q_t, Z_t\rangle, \tag{B.33}$$

where

$$\mathcal{R}_t^{\mathrm{fs}} := L(r_{t+1}^{L-1} - r_t^{L-1}) - \frac{(L-1)(r_{t+1}^L - r_t^L) - D_\psi(W_t, W_{t+1})}{r_t + \lambda}.$$

For $x = r_t$ and $y = r_{t+1}$, the part of $\mathcal{R}_t^{\mathrm{fs}}$ not containing the Bregman divergence has the integral representation

$$\frac{L(L-1)}{x+\lambda}\int_x^y z^{L-2}(x+\lambda-z)\,\mathrm{d}z.$$

By (B.20) and (B.21), we have $y \in [x/2, 3x/2]$ for all sufficiently large t. Using the integral representation and (B.22), we obtain

$$|\mathcal{R}_t^{\mathrm{fs}}| \le C_\lambda\left[x^{L-3}|y-x| + x^{L-3}|y-x|^2 + \frac{D_\psi(W_t, W_{t+1})}{x}\right] \le C_\lambda\left(\frac{\eta_t}{x} + \frac{\eta_t^2}{x^{L-1}}\right).$$

Absorbing the finite prefix into the constant gives

$$|\mathcal{R}_t^{\mathrm{fs}}| \le C\frac{\eta_t}{1+r_t}. \tag{B.34}$$

Finally, (B.20) and an integral comparison show that

$$\sum_{t<T}\frac{\eta_t}{1+r_t} \le C\begin{cases}\log(\mathrm{e}+S_T), & L=2,\\ (1+S_T)^{(L-2)/(L-1)}, & L>2.\end{cases} \tag{B.35}$$

Summing (B.33) from zero and using (B.35), we obtain

$$\mathcal{E}_T^{\mathrm{dir}} \le C(1+S_T)\rho_L(S_T) - \sum_{t<T}\eta_t\langle q_t, Z_t\rangle.$$

Since $\mathcal{E}_T^{\mathrm{dir}} \ge 0$ and $r_T^{L-1} \asymp 1 + S_T$, dividing by r_T^{L-1} and applying (B.31) proves (B.29). Applying (B.26) to the same bound, we obtain (B.28). □

C Proofs of the Results for NMD without Momentum

For stochastic proofs in this section, $\mathcal{F}_t$ denotes the sigma algebra generated by the initialization and the batches preceding iteration t.

C.1 Proof of Theorem 1

Proof of Theorem 1. We apply Theorem B.1 with $b_t^\circ = \widehat{b}_t = \nabla\mathcal{L}(W_t)/\left\|\nabla\mathcal{L}(W_t)\right\|_{\psi,*}$, $q_t = 0$, and the offset λ_{ce} from Lemma B.3. By Lemma B.3, the reference satisfies (B.24). Its dual norm is one and (B.5) supplies uniform separator alignment. We therefore obtain

$$D_\psi\Big(\widehat{W}, W_T/\left\|W_T\right\|_\psi\Big) \le C\rho_L(S_T).$$

Under Assumption 3, we have $S_T = \Theta(T^{1-a})$ for $a < 1$ and $S_T = \Theta(\log T)$ for $a = 1$. Substituting these scales into the definition of ρ_L, we obtain the Bregman rate in (8). Applying (B.17), we obtain the margin rate and convergence to $\widehat{\gamma}$. The first inequality in (B.17) also gives $\left\|W_T/\left\|W_T\right\|_\psi - \widehat{W}\right\|_\psi \to 0$. Finally, Lemma B.7 ensures $\left\|W_T\right\|_\psi \to \infty$. Since $\gamma(W_T) \to \widehat{\gamma} > 0$, every margin $m_{il}(W_T)$ diverges to positive infinity. Substituting into (1), we obtain $\mathcal{L}(W_T) \to 0$. □

C.2 Proof of Theorem 2

We construct n distinct samples covering all k classes with $\|x_i\|_1 \le 1$. The proof uses a cone of majority-class updates and continuity of the normalized inverse mirror map; for reshuffling, it holds for every permutation sequence. A Frobenius specialization then gives an explicit batch-size threshold.

Proof of Theorem 2. Let u_0 and $(u_l)_{l\in[k],\, l\neq 2}$ be distinct standard basis vectors in $\mathbb{R}^d$. Consider the cone and its normalized convex hull

$$\mathcal{K} := \left\{ \sum_{l\neq 2} b_l (e_2 - e_l) u_0^\top : b_l \ge 0 \right\}, \qquad \mathcal{H} := \operatorname{conv}\{Z \in \mathcal{K} : \|Z\|_{\psi,*} = 1\}. \tag{C.1}$$

The set $\mathcal{H}$ is compact, convex, and does not contain zero. Indeed, for a nonzero element of $\mathcal{K}$, we have $\left\langle Z, e_2 u_0^\top \right\rangle = \sum_{l\neq 2} b_l > 0$. Its minimum over the compact intersection of $\mathcal{K}$ with the dual unit sphere is strictly positive, and the same lower bound holds on the convex hull.

For $Z \neq 0$, define

$$\mathcal{V}(Z) := \frac{\nabla\psi^*(Z)}{\|\nabla\psi^*(Z)\|_\psi}, \qquad f(Z) := \max_{l\neq 2} (e_2 - e_l)^\top \mathcal{V}(Z) u_0.$$

For $Z = \sum_{l\neq 2} b_l (e_2 - e_l) u_0^\top \in \mathcal{H}$, Euler's identity implies

$$\sum_{l\neq 2} b_l (e_2 - e_l)^\top \mathcal{V}(Z) u_0 = \langle Z, \mathcal{V}(Z)\rangle = L \|\nabla\psi^*(Z)\|_\psi^{L-1} > 0.$$

Thus $f(Z) > 0$ on $\mathcal{H}$. By continuity of the inverse mirror map and compactness, we can choose $b > 0$ and $0 < \rho \le 1$ such that

$$\operatorname{dist}_{\psi,*}(Z, \mathcal{H}) \le \rho \quad \Longrightarrow \quad Z \neq 0, \quad f(Z) \ge b/2, \qquad b := \min_{Z\in\mathcal{H}} f(Z) > 0. \tag{C.2}$$

We take ρ smaller than half the dual norm distance of $\mathcal{H}$ from zero.

Let

$$A := \max_{l\neq 2} \left\| (e_l - e_2) u_l^\top \right\|_{\psi,*}, \qquad 0 < \tau \le \min\{1, b/(4A)\}, \qquad c := \max\{k, 8(k-1)/\rho\}.$$

Fix a batch size s, an integer $n \ge cs$, and a sampling distribution p as in Assumption 4. We choose $\mathcal{I} = \{i_l : l \in [k],\, l \neq 2\}$ to contain the indices of the $k-1$ smallest entries of p. Their average is at most the average of all n entries, so we have

$$\sum_{i\in\mathcal{I}} p_i \le \frac{k-1}{n}. \tag{C.3}$$

At these indices, we construct the samples

$$x_{i_l} := \tfrac{1}{2}(u_0 + \tau u_l), \qquad y_{i_l} = l \quad (l \neq 2), \tag{C.4}$$

and place $n - k + 1$ class 2 samples with inputs $\lambda_j u_0/2$ at the remaining indices, where $\lambda_1, \ldots, \lambda_{n-k+1}$ are fixed distinct scalars in $[1/2, 1]$, chosen independently of p. All n samples are distinct and have ℓ_1 norm at most one. To verify separability, choose a weight matrix with rows $w_2 = u_0$ and $w_l = 2u_l/\tau$ for $l \neq 2$. On a class 2 sample, its true score is $\lambda_j/2 > 0$ and every competing score is zero. On the sample in (C.4), the true score is one, the class 2 score is $1/2$, and every other score is zero.

We call the samples in (C.4) rare. Write

$$d_t := -\frac{\widehat{g}_t}{\|\widehat{g}_t\|_{\psi,*}}, \qquad J_t := \mathbf{1}\{\mathcal{B}_t \text{ contains a rare sample}\}, \qquad q_T := \frac{1}{S_T} \sum_{t<T} \eta_t J_t.$$

By Lemma B.6, every d_t has unit dual norm. On a batch with $J_t = 0$, the negative gradient is a nonzero element of $\mathcal{K}$, so $d_t \in \mathcal{H}$. Under Assumption 4, we obtain

$$\mathbb{P}(J_t = 1 \mid \mathcal{F}_t) = 1 - \left(1 - \sum_{i\in\mathcal{I}} p_i\right)^s \le s \sum_{i\in\mathcal{I}} p_i \le \frac{s(k-1)}{n} \le \frac{\rho}{8}. \tag{C.5}$$

Here we used the union bound, (C.3), and $n \geq cs$. For the bounded martingale differences $X_t := J_t - \mathbb{E}[J_t \mid \mathcal{F}_t]$, $\sum_t \eta_t^2/(1+S_t)^2 < \infty$. Martingale convergence and Kronecker's lemma give $S_T^{-1}\sum_{t<T}\eta_t X_t \to 0$ almost surely. Combining this law with (C.5), we obtain, almost surely,

$$q_T \leq \rho/4 \quad \text{for all sufficiently large } T. \tag{C.6}$$

For epoch partitions, at most $k-1$ batches per epoch contain a rare sample. With $m = n/s$, the polynomial step sizes satisfy $\eta_{rm+j}/\eta_{rm} \to 1$ uniformly over $0 \leq j < m$. In particular, for every sufficiently late complete epoch,

$$\frac{\sum_{j=0}^{m-1}\eta_{rm+j}J_{rm+j}}{\sum_{j=0}^{m-1}\eta_{rm+j}} \leq \frac{(k-1)\eta_{rm}}{m\eta_{rm+m-1}} = \frac{k-1}{m} + o(1) \leq \frac{\rho}{8} + o(1).$$

The bound remains valid after taking the weighted average over complete late epochs. The finite prefix and unfinished epoch have weight $o(S_T)$, so we obtain (C.6) for every partition sequence as well.

For arbitrary fixed W_0, let $\Theta_0 := \nabla\psi(W_0)$. For all sufficiently large T, summing (6) and separating the batches according to J_t, we obtain

$$\frac{\Theta_T}{S_T} = \frac{\Theta_0}{S_T} + \frac{1}{S_T}\sum_{t<T}\eta_t d_t = \frac{\Theta_0}{S_T} + (1-q_T)Z_T + q_T Y_T, \tag{C.7}$$

where $Z_T \in \mathcal{H}$, $\|Y_T\|_{\psi,*} \leq 1$, and the term containing Y_T is zero if no rare batch has occurred. By convexity of the dual unit ball, we have $\|Z_T\|_{\psi,*} \leq 1$. Since $S_T \to \infty$ and Θ_0 is fixed, we also have $\|\Theta_0\|_{\psi,*}/S_T \leq \rho/4$ for all sufficiently large T. Combining (C.6) and (C.7), we obtain

$$\operatorname{dist}_{\psi,*}(\Theta_T/S_T, \mathcal{H}) \leq \frac{\|\Theta_0\|_{\psi,*}}{S_T} + q_T\|Y_T - Z_T\|_{\psi,*} \leq \frac{\|\Theta_0\|_{\psi,*}}{S_T} + 2q_T \leq \frac{3\rho}{4} < \rho \tag{C.8}$$

eventually. By the choice of ρ and (C.8), we obtain a positive lower bound on $\|\Theta_T/S_T\|_{\psi,*}$. By (C.7), we also have $\|\Theta_T/S_T\|_{\psi,*} \leq 1 + \|\Theta_0\|_{\psi,*}/S_T \leq 2$ eventually. By (B.14), we obtain $\|W_T\|_\psi = \Theta(S_T^{1/(L-1)}) \to \infty$.

Homogeneity of the inverse mirror map implies $W_T/\|W_T\|_\psi = \mathcal{V}(\Theta_T/S_T)$. Set $V_T := W_T/\|W_T\|_\psi$. By (C.2), some $l_T \neq 2$ satisfies $(e_2 - e_{l_T})^\top V_T u_0 \geq b/2$. For the rare sample with this label, we have

$$\begin{aligned}\gamma(W_T) &\leq (e_{l_T} - e_2)^\top V_T x_{i_{l_T}} \\ &= \tfrac{1}{2}(e_{l_T} - e_2)^\top V_T u_0 + \tfrac{\tau}{2}(e_{l_T} - e_2)^\top V_T u_{l_T} \\ &\leq -b/4 + \tau A/2 \leq -b/8.\end{aligned} \tag{C.9}$$

The last line uses duality, $\|V_T\|_\psi = 1$, and the choice of τ. Taking $\delta = b/8$ and using $\widehat{\gamma} > 0$, we obtain $\widehat{\gamma} - \gamma(W_T) \geq \widehat{\gamma} + \delta > \delta$ eventually, which proves the margin gap bound in (9). At every sufficiently large T, at least one sample therefore has a pairwise margin at most $-\delta\|W_T\|_\psi$. Although its index may vary with T, the empirical loss satisfies

$$\mathcal{L}(W_T) \geq \frac{1}{n}\log\bigl(1 + \exp(\delta\|W_T\|_\psi)\bigr) \longrightarrow \infty.$$

This proves the loss divergence in (9). The constants c, δ depend only on the potential and dimensions; only the assignment of sample indices depends on p. □

Proposition C.1. *Let $k \geq 2$, $n \geq k$, $d = k$, and $0 < \alpha < 1$. There is a linearly separable dataset of n distinct samples, containing every class and satisfying $\|x_i\|_1 \leq \sqrt{2}$, with the following property. Consider NMD without momentum, with $W_0 = 0$, Assumption 3, and $\psi(W) = \nu^L\|W\|_{\mathrm{F}}^L$, where $\nu > 0$ and $L \geq 2$. Then $\mathcal{L}(W_T) \to \infty$ in either of the following cases:*

(i) Uniform independent sampling with replacement with fixed batch size s, almost surely, if

$$1 - \left(1 - \frac{k-1}{n}\right)^s < \frac{\alpha}{2+\alpha}. \tag{C.10}$$

(ii) Epoch partitions into batches of size $s \mid n$, with $m = n/s$, for every sequence of partitions, if

$$\frac{k-1}{m} < \frac{\alpha}{2+\alpha}. \tag{C.11}$$

In particular, the second conclusion holds under random reshuffling. If $k \geq 3$, $\alpha = 3/4$, and $n \geq 4s(k-1)$, both conditions hold in their respective sampling settings, and $\gamma(W_T) \leq -\sqrt{2}/(64\nu)$ for all sufficiently large T. The random reshuffling conclusion holds for every sequence of permutations.

Proof. Let $u_1, \dots, u_k$ be the standard basis of the input space, and put $v := -\alpha u_1 + \sqrt{1-\alpha^2}\, u_2$. The dataset contains $(u_1, 1)$, the $k-2$ samples (u_l, l) for $l = 3, \dots, k$, and $n-k+1$ samples $(-\lambda_j v, 2)$, where the $\lambda_j \in [1/2, 1]$ are distinct. Every input has ℓ_1 norm at most $\sqrt{2}$. Choose a separating matrix with rows $w_1^\star = u_1$, $w_2^\star = -v$, and $w_l^\star = u_l$ for $l \geq 3$. The class 1 and class 2 samples have positive pairwise margins because $\alpha < 1$, and every remaining sample has true score one and all competing scores zero. Thus the data satisfy both standing data assumptions.

Define the unit Frobenius matrix $H := (e_1 - e_2)u_1^\top/\sqrt{2}$, and let $D_t := -\widehat{g}_t/\left\|\widehat{g}_t\right\|_{\mathrm{F}}$. We call the $k-1$ samples outside class 2 the rare samples. For a batch containing no rare sample, its negative gradient has the form $a(-v)^\top$, where $a = \sum_{l\neq 2} b_l(e_2 - e_l)$ with $b_l \geq 0$ and $b := \sum_{l\neq 2} b_l > 0$. Hence

$$\langle a, e_1 - e_2\rangle = -(b + b_1), \qquad \|a\|_2 = \left(b^2 + \sum_{l\neq 2} b_l^2\right)^{1/2} \leq \sqrt{2}\, b.$$

Taking the Frobenius inner product with H, we obtain

$$\langle D_t, H\rangle \leq -\frac{\alpha}{2} \quad \text{if the batch contains no rare sample,} \qquad \langle D_t, H\rangle \leq 1 \quad \text{for every batch.} \tag{C.12}$$

Under uniform independent sampling with replacement, a batch contains a rare sample with probability $q := 1 - (1 - (k-1)/n)^s$, independently of the preceding iterates. By (C.12), we have

$$\mathbb{E}[\langle D_t, H\rangle \mid \mathcal{F}_t] \leq q - \frac{\alpha}{2}(1-q) =: -\delta < 0,$$

where the strict inequality is exactly (C.10). The centered projections are bounded martingale differences. Since $\sum_t \eta_t^2/(1+S_t)^2 < \infty$, the weighted martingale strong law implies

$$\limsup_{T\to\infty} \frac{1}{S_T} \sum_{t<T} \eta_t \langle D_t, H\rangle \leq -\delta \qquad \text{almost surely.} \tag{C.13}$$

For epoch partitions, at most $k-1$ of the m batches in each epoch contain a rare sample. Using (C.12), we obtain

$$\sum_{j=0}^{m-1} \langle D_{rm+j}, H\rangle \leq (k-1) - \frac{\alpha}{2}(m-k+1) = -m\delta, \qquad \delta := \frac{\alpha}{2}\left(1 - \frac{k-1}{m}\right) - \frac{k-1}{m} > 0.$$

For fixed m, Assumption 3 implies $\eta_{rm+j}/\eta_{rm} \to 1$ uniformly over $0 \leq j < m$. Multiplying the preceding bound by η_{rm}, summing over epochs, and bounding the replacement error by $\sum_{r,j} |\eta_{rm+j} - \eta_{rm}| = o(S_T)$, we obtain (C.13) for every epoch partition sequence. The contribution of an unfinished epoch is also $o(S_T)$.

For the stated potential, $\|Z\|_{\psi,*} = \|Z\|_{\mathrm{F}}/\nu$, so $\Theta_T := \nabla\psi(W_T) = \nu \sum_{t<T} \eta_t D_t$. Thus $\|\Theta_T\|_{\mathrm{F}} \leq \nu S_T$, while (C.13) implies $\langle \Theta_T, H\rangle \leq -\nu\delta S_T/2$ eventually. The radial mirror map preserves Frobenius directions, and hence

$$\left\langle H, \frac{W_T}{\|W_T\|_{\mathrm{F}}} \right\rangle = \frac{\langle H, \Theta_T\rangle}{\|\Theta_T\|_{\mathrm{F}}} \leq -\delta/2.$$

Moreover, $\|\Theta_T\|_{\mathrm{F}} \to \infty$ implies $\|W_T\|_{\mathrm{F}} \to \infty$. The class 1 sample therefore has $(e_1 - e_2)^\top W_T u_1 = \sqrt{2}\langle H, W_T\rangle \to -\infty$. Its cross entropy loss is at least $\log(1 + \exp(-(e_1 - e_2)^\top W_T u_1))$, which tends to infinity. This proves the claimed divergence of the empirical loss.

For the stated threshold, we take $\alpha = 3/4$. The condition $n \geq 4s(k-1)$ ensures that

$$1 - \left(1 - \frac{k-1}{n}\right)^s \leq \frac{s(k-1)}{n} \leq \frac{1}{4} < \frac{3}{11} = \frac{\alpha}{2+\alpha}.$$

For random reshuffling, $(k-1)/m = s(k-1)/n$. Thus both (C.10) and (C.11) hold. In both cases, the constant δ in the proof satisfies $\delta \geq (3/8)(3/4) - 1/4 = 1/32$. Using (C.13), we therefore obtain, for all sufficiently large T,

$$\gamma(W_T) \leq \frac{\sqrt{2}}{\nu} \left\langle H, \frac{W_T}{\|W_T\|_{\mathrm{F}}} \right\rangle \leq -\frac{\sqrt{2}}{64\nu}.$$

This proves the stated negative margin bound. □

C.3 Proof of Proposition 1

We characterize the direction selected by sampling frequencies on orthogonal data. Let $u_1, \ldots, u_k \in \mathbb{R}^d$ be orthonormal, with $d \geq k$, and consider

$$x_i = \alpha_i u_{y_i}, \qquad \alpha_i > 0, \qquad n_l := |\{i : y_i = l\}| > 0. \tag{C.14}$$

We write

$$\kappa_k := \sqrt{\frac{k-1}{k}}, \qquad R_l := \frac{e_l - k^{-1}\mathbf{1}}{\kappa_k} u_l^\top, \qquad \alpha_l^{\min} := \min_{i:y_i=l} \alpha_i.$$

Sampling frequencies.

Proposition C.2. *Suppose that Assumption 3 holds, and let the singleton batch at time t contain the index I_t. Assume that there are matrices $D_1, \ldots, D_n$, with $\|D_i\|_{\psi,*} = 1$, such that, along the trajectory,*

$$-\frac{\nabla \ell_i(W_t)}{\|\nabla \ell_i(W_t)\|_{\psi,*}} = D_i \qquad \textit{for every } i \in [n] \textit{ and } t \geq 0. \tag{C.15}$$

Consider the iterates of (6)*, with $\Theta_t := \nabla\psi(W_t)$. If there is a probability vector $\pi \in \mathbb{R}^n_+$ such that*

$$\frac{1}{S_T} \sum_{t<T} \eta_t \mathbf{1}\{I_t = i\} \longrightarrow \pi_i \qquad \textit{for every } i \in [n], \tag{C.16}$$

and $\overline{D}_\pi := \sum_i \pi_i D_i \neq 0$, then

$$\frac{\Theta_T}{S_T} \longrightarrow \overline{D}_\pi, \qquad \frac{W_T}{\|W_T\|_\psi} \longrightarrow \widehat{W}_\pi^\psi := \frac{\nabla\psi^*(\overline{D}_\pi)}{\left\|\nabla\psi^*(\overline{D}_\pi)\right\|_\psi}, \qquad D_\psi\left(\widehat{W}_\pi^\psi, W_T/\|W_T\|_\psi\right) \longrightarrow 0. \tag{C.17}$$

For independent sampling with replacement and batch size one, as in Assumption 4, (C.16) *holds almost surely with $\pi_i = p_i$. It also holds pathwise with $\pi_i = 1/n$ for every sequence of singleton epoch partitions in Assumption B.1.*

Proof. Set $D_t := D_{I_t}$. By (6) and (C.15), we have the exact identity

$$\Theta_T = \Theta_0 + \sum_{t<T} \eta_t D_t. \tag{C.18}$$

Dividing by S_T and applying (C.16), we obtain

$$\frac{\Theta_T}{S_T} = \frac{\Theta_0}{S_T} + \frac{1}{S_T} \sum_{t<T} \eta_t D_t \longrightarrow \overline{D}_\pi.$$

The inverse mirror map is homogeneous of degree $1/(L-1)$. Hence

$$\frac{W_T}{S_T^{1/(L-1)}} \longrightarrow \nabla\psi^*(\overline{D}_\pi),$$

Normalization and continuity of D_ψ give (C.17).

It remains to verify the two sampling claims. Under Assumption 4, for every i, the variables $X_{i,t} := \mathbf{1}\{I_t = i\} - p_i$ are bounded martingale differences. The schedule in Assumption 3 satisfies

$$\sum_{t\geq 0} \frac{\eta_t^2}{S_{t+1}^2} < \infty.$$

The martingale convergence theorem and Kronecker's lemma therefore imply $S_T^{-1} \sum_{t<T} \eta_t X_{i,t} \to 0$ almost surely.

For singleton epoch partitions, let $j_{i,r} \in \{0, \dots, n-1\}$ be the position of sample i in epoch r, and let $\overline{\eta}_r = n^{-1} \sum_{j=0}^{n-1} \eta_{rn+j}$. Polynomial decay and monotonicity imply

$$\sum_{r\geq 0} \left|\eta_{rn+j_{i,r}} - \overline{\eta}_r\right| \leq \sum_{r\geq 0} (\eta_{rn} - \eta_{rn+n-1}) < \infty.$$

Every completed epoch contributes $n\overline{\eta}_r$ to the total clock and one term $\eta_{rn+j_{i,r}}$ to sample i. Dividing by S_T, and noting that the unfinished epoch contributes $o(S_T)$, proves (C.16) with $\pi_i = 1/n$. □

Columnwise permutation symmetry.

Proposition C.3. *Let $d = k \geq 3$, take the data in* (C.14) *with $u_l = e_l$, and assume Assumption 3. Suppose that ψ is invariant under independent permutations within each column: for every $W = [w_1, \dots, w_k] \in \mathbb{R}^{k\times k}$ and permutation matrices $P_1, \dots, P_k$,*

$$\psi([P_1 w_1, \dots, P_k w_k]) = \psi(W). \tag{C.19}$$

Run (6) *from $W_0 = 0$ with batch size one under either Assumption 4 or Assumption B.1. Let $\pi_l := \sum_{i:y_i=l} p_i$ in the independent sampling case and $\pi_l := n_l/n$ in the epoch case. Define*

$$\begin{aligned} d_l &:= \|R_l\|_{\psi,*}, \qquad B_\pi := \sum_{l=1}^{k} \frac{\pi_l}{d_l} R_l, \\ Z_\pi &:= \nabla\psi^*(B_\pi), \quad \widehat{W}_\pi^\psi := \frac{Z_\pi}{\|Z_\pi\|_\psi}. \end{aligned} \tag{C.20}$$

Then, almost surely under independent sampling and for every sequence of singleton epoch partitions, we have

$$\begin{aligned} \frac{\Theta_T}{S_T} &\longrightarrow B_\pi, \qquad \frac{W_T}{S_T^{1/(L-1)}} \longrightarrow Z_\pi, \\ \frac{W_T}{\|W_T\|_\psi} &\longrightarrow \widehat{W}_\pi^\psi, \qquad \mathcal{L}(W_T) \longrightarrow 0. \end{aligned} \tag{C.21}$$

For each $l \in [k]$, define $h_l := (e_l - e_j)^\top \widehat{W}_\pi^\psi e_l$ using any $j \neq l$. This value is independent of j and satisfies $h_l > 0$. The max-margin recovery criterion is

$$\widehat{W}_\pi^\psi = \widehat{W} \quad \Longleftrightarrow \quad \alpha_1^{\min} h_1 = \cdots = \alpha_k^{\min} h_k. \tag{C.22}$$

If the quantities $\alpha_l^{\min} h_l$ are not all equal, then $D_\psi(\widehat{W}, \widehat{W}_\pi^\psi) > 0$, and the limiting margin is strictly smaller than $\widehat{\gamma}$.

Proof. Let $\mathcal{T}$ apply a row permutation independently to each column. This is an orthogonal linear map for the Frobenius inner product. From (C.19) and the definition of the conjugate, we have

$$\psi^*(\mathcal{T}\Theta) = \psi^*(\Theta), \qquad \nabla\psi^*(\mathcal{T}\Theta) = \mathcal{T}\nabla\psi^*(\Theta). \tag{C.23}$$

Suppose that $\Theta = \sum_l b_l R_l/d_l$ with $b_l \geq 0$. For column l, any permutation that fixes row l preserves Θ. By (C.23), all incorrect class coordinates in column l of $W = \nabla\psi^*(\Theta)$ are equal. The negative gradient of any sample from that class is therefore a positive multiple of R_l. Starting from zero and applying induction, we obtain

$$-\frac{\nabla\ell_i(W_t)}{\|\nabla\ell_i(W_t)\|_{\psi,*}} = \frac{R_{y_i}}{d_{y_i}}, \qquad \Theta_T = \sum_{t<T} \eta_t \frac{R_{y_{I_t}}}{d_{y_{I_t}}}. \tag{C.24}$$

Applying Proposition C.2, together with the homogeneity and continuity of $\nabla\psi^*$, we obtain the three parameter limits in (C.21).

We next verify that every limiting class margin is positive. Let $\mathcal{T}$ exchange rows l and $j \neq l$ only in column l. Since ψ^* is strictly convex, its gradient is strictly monotone. The two matrices B_π and $\mathcal{T}B_\pi$ are distinct because $\pi_l > 0$. Using (C.20) and (C.23), we have

$$\begin{aligned} 0 &< \langle B_\pi - \mathcal{T}B_\pi, Z_\pi - \mathcal{T}Z_\pi \rangle \\ &= \frac{2\pi_l}{\kappa_k d_l}(e_l - e_j)^\top Z_\pi e_l, \end{aligned}$$

where the identity follows by restricting the inner product to the two exchanged entries. Thus $h_l > 0$. Combining this positivity with (C.21), we obtain divergence of every sample margin and hence $\mathcal{L}(W_T) \to 0$.

To prove (C.22), let $r_\pi = \|Z_\pi\|_\psi$, $\gamma_\pi = \min_l \alpha_l^{\min} h_l > 0$, and $M = \widehat{W}_\pi^\psi / \gamma_\pi$. This matrix is feasible for (B.12), and homogeneity implies

$$\nabla\psi(M) = (r_\pi \gamma_\pi)^{-(L-1)} B_\pi. \tag{C.25}$$

If $\alpha_l^{\min} h_l > \gamma_\pi$ for some l, then every constraint from that class is strict. The direction $-R_l$ changes only column l, so $M - \tau R_l$ remains feasible for all sufficiently small $\tau > 0$. From (C.25), we have

$$\langle \nabla\psi(M), -R_l \rangle = -(r_\pi \gamma_\pi)^{-(L-1)} \frac{\pi_l}{d_l} < 0.$$

Thus M is not the minimum norm feasible point, which proves necessity in (C.22).

Conversely, suppose that $\alpha_l^{\min} h_l = \gamma_\pi$ for every l. For each class, choose a sample attaining $\alpha_l^{\min}$. All its constraints are active at M. By the identity $R_l = (\kappa_k k)^{-1} \sum_{j \neq l} (e_l - e_j) e_l^\top$, we can rewrite (C.25) as

$$\nabla\psi(M) = \sum_{l=1}^{k} \sum_{j \neq l} \frac{(r_\pi \gamma_\pi)^{-(L-1)} \pi_l}{\kappa_k k d_l \alpha_l^{\min}} \left[\alpha_l^{\min} (e_l - e_j) e_l^\top\right].$$

These are nonnegative multipliers on active constraints and satisfy the KKT conditions for minimizing ψ over the hard-margin constraints. Strict convexity therefore implies $M = W_{\mathrm{hm}}$, which proves sufficiency. Continuity, strict convexity, and uniqueness of the max-margin direction imply the stated positive Bregman error and strict margin gap. □

Proof of Proposition 1. The matrix I_k separates $\mathcal{S}_2$, since every pairwise margin of sample i equals $\alpha_i > 0$. The data are bounded with $B = \max_{i \in [n]} \alpha_i$. Under uniform independent sampling with replacement, the probability of selecting class l is $\pi_l = n_l/n$. Under random reshuffling, every epoch contains n_l samples from class l, so we use the same π_l. We apply Proposition C.3 with $u_l = e_l$ and $\pi_l = n_l/n$. By (C.20), we have $G = B_\pi$ and $\bar{W} = \widehat{W}_\pi^\psi$. By (C.21), we obtain the directional convergence and vanishing loss in (10). The symmetry argument in Proposition C.3 ensures that the class margin factors of $\bar{W}$ equal $h_l > 0$. If $\alpha_1^{\min} h_1, \ldots, \alpha_k^{\min} h_k$ are not all equal, (C.22) gives $\bar{W} \neq \widehat{W}$. □

Frobenius radial potentials. For $\Phi(W) = \|W\|_{\mathrm{F}}^L / L$, the potential in Definition 1 is radial in the Frobenius norm:

$$\psi(W) = \frac{(1+\epsilon)^{L/2}}{L} \|W\|_{\mathrm{F}}^L, \qquad \|W\|_\psi = \nu_{L,\epsilon} \|W\|_{\mathrm{F}}, \qquad \nu_{L,\epsilon} := \frac{\sqrt{1+\epsilon}}{L^{1/L}}. \tag{C.26}$$

Theorem C.1. *Consider* (C.14) *with* $k \geq 3$ *and* (C.26) *with* $L \geq 2$ *and* $\epsilon > 0$. *Use* (6) *with* $W_0 = 0$, *batch size one, and Assumption 3. Under independent sampling with replacement in Assumption 4, let* $\pi_l := \sum_{i:y_i=l} p_i$. *For any sequence of epoch permutations that uses every sample once per epoch, let* $\pi_l := n_l/n$. *Then*

$$\frac{W_T}{\|W_T\|_\psi} \longrightarrow \bar{W} := \frac{\sum_{l=1}^k \pi_l R_l}{\nu_{L,\epsilon} \|\pi\|_2}, \qquad \mathcal{L}(W_T) \longrightarrow 0. \tag{C.27}$$

Moreover, $\|W_T\|_\psi = \Theta(T^{(1-a)/(L-1)})$ *for* $0 < a < 1$*, and* $\|W_T\|_\psi = \Theta((\log T)^{1/(L-1)})$ *for* $a = 1$*. The max-margin direction in* (4) *is*

$$\widehat{W} = \frac{\sum_{l=1}^k (\alpha_l^{\min})^{-1} R_l}{\nu_{L,\epsilon} \sqrt{\sum_{l=1}^k (\alpha_l^{\min})^{-2}}}. \tag{C.28}$$

Consequently,

$$\lim_{T\to\infty} D_\psi\left(\widehat{W}, W_T / \|W_T\|_\psi\right) = L\left(1 - \frac{\sum_{l=1}^k \pi_l (\alpha_l^{\min})^{-1}}{\|\pi\|_2 \sqrt{\sum_{l=1}^k (\alpha_l^{\min})^{-2}}}\right). \tag{C.29}$$

This limit is positive exactly when $(\pi_l)_{l=1}^k \not\propto ((\alpha_l^{\min})^{-1})_{l=1}^k$*. The conclusions hold almost surely under independent sampling with replacement and for every sequence of epoch permutations.*

Proof of Theorem C.1. Write $\nu = \nu_{L,\epsilon}$ and $v_l = e_l - k^{-1}\mathbf{1}$. The matrices $R_l = \kappa_k^{-1} v_l u_l^\top$ are orthonormal in the Frobenius inner product. From (C.26), we have

$$\|Z\|_{\psi,*} = \nu^{-1} \|Z\|_{\mathrm{F}}, \qquad \nabla\psi(W) = L\nu^L \|W\|_{\mathrm{F}}^{L-2} W. \tag{C.30}$$

Thus $\nabla\psi^*(\Theta)$ is a positive multiple of Θ whenever $\Theta \neq 0$.

Suppose that $W = \sum_l a_l R_l$ with $a_l \geq 0$, and consider a sample i from class l. By (C.14), we have $Wx_i = \alpha_i a_l v_l / \kappa_k$. All incorrect class scores are equal, so the negative sample gradient is a positive multiple of R_l. Using (C.30), we obtain

$$-\frac{\nabla \ell_i(W)}{\|\nabla \ell_i(W)\|_{\psi,*}} = \nu R_{y_i}. \tag{C.31}$$

Starting from zero, we retain this form at every iteration. Hence, with $S_{l,T} := \sum_{t<T} \eta_t \mathbf{1}\{y_{I_t} = l\}$, we have

$$\Theta_T = \nu \sum_{l=1}^k S_{l,T} R_l, \qquad \frac{W_T}{\|W_T\|_\psi} = \frac{\sum_{l=1}^k S_{l,T} R_l}{\nu \sqrt{\sum_{l=1}^k S_{l,T}^2}}. \tag{C.32}$$

Applying the sampling conclusions of Proposition C.2 and summing the sample frequencies within each class, we obtain $S_{l,T}/S_T \to \pi_l$. Substituting into (C.32), we obtain the directional limit in (C.27).

From (C.30) and (C.32), we also have

$$L\|W_T\|_\psi^{L-1} = \|\Theta_T\|_{\psi,*} = \sqrt{\sum_{l=1}^k S_{l,T}^2} = S_T \|\pi\|_2 + o(S_T).$$

Since every $\pi_l > 0$, we obtain the claimed radial growth. For every sample i from class l, all its normalized margins converge to $\alpha_i \pi_l / (\kappa_k \nu \|\pi\|_2) > 0$. The unnormalized margins therefore diverge, and substitution into (1) proves $\mathcal{L}(W_T) \to 0$.

We next identify the max-margin direction. Let $U = [u_1, \ldots, u_k]$. The projection $W \mapsto WUU^\top$ preserves every training score and cannot increase the Frobenius norm. We can therefore solve (B.12) over $W = ZU^\top$, where the constraints separate across the columns of Z. For column l, the constraints are

$$(e_l - e_j)^\top Z e_l \geq (\alpha_l^{\min})^{-1} \qquad (j \neq l).$$

By centering this column and averaging its incorrect class coordinates, we preserve feasibility and cannot increase its Euclidean norm. The unique minimum norm column is $Ze_l = (\alpha_l^{\min})^{-1} v_l$. Consequently,

$$W_{\mathrm{hm}} = \kappa_k \sum_{l=1}^k (\alpha_l^{\min})^{-1} R_l, \qquad \|W_{\mathrm{hm}}\|_\psi = \nu\kappa_k \sqrt{\sum_{l=1}^k (\alpha_l^{\min})^{-2}}.$$

Normalizing this matrix, we obtain (C.28).

For $\|A\|_\psi = \|B\|_\psi = 1$, we have

$$D_\psi(A, B) = L\left(1 - \nu^2 \langle A, B\rangle\right).$$

Substituting (C.27) and (C.28), and using the orthonormality of the matrices R_l, we obtain (C.29). The equality condition in the Cauchy–Schwarz inequality proves the stated criterion for max-margin recovery. □

Corollary C.1. *Under the data, initialization, and step size conditions of Theorem C.1, consider normalized SGD*

$$W_{t+1} = W_t - \eta_t \frac{\nabla \ell_{I_t}(W_t)}{\|\nabla \ell_{I_t}(W_t)\|_{\mathrm{F}}}. \tag{C.33}$$

Under uniform independent sampling with replacement or any sequence of epoch permutations, we have

$$\frac{W_T}{\|W_T\|_{\mathrm{F}}} \longrightarrow \frac{\sum_{l=1}^k n_l R_l}{\sqrt{\sum_{l=1}^k n_l^2}}, \qquad \mathcal{L}(W_T) \longrightarrow 0. \tag{C.34}$$

The limit equals the Frobenius max-margin direction if and only if $(n_l)_{l=1}^k \propto ((\alpha_l^{\min})^{-1})_{l=1}^k$. The conclusion for independent sampling with replacement holds almost surely.

Proof of Corollary C.1. Set $L = 2$ and $\epsilon = 1$ in (C.26). Then $\psi(W) = \|W\|_{\mathrm{F}}^2$, $\nu_{2,1} = 1$, and $\Theta_t = 2W_t$. The mirror update with step size $2\eta_t$ is exactly (C.33). This rescaling preserves Assumption 3. Applying Theorem C.1 with $\pi_l = n_l/n$, we obtain (C.34) and the recovery criterion. □

Smoothed entrywise p norm potentials.

Corollary C.2. *Use the data, algorithm, and initialization in Proposition C.3, and let*

$$\psi(W) = \frac{1}{L}\left(\|W\|_p^2 + \epsilon \|W\|_{\mathrm{F}}^2\right)^{L/2}, \qquad \|W\|_p := \left(\sum_{i=1}^k \sum_{j=1}^k |W_{i,j}|^p\right)^{1/p}, \tag{C.35}$$

where $2 < p < \infty$, $L \geq 2$, and $\epsilon > 0$. For arbitrary positive class probabilities π_l, the conclusions of Proposition C.3 hold. If $\pi_l = 1/k$ for every l, then

$$\widehat{W}_\pi^\psi = \frac{M_r}{\|M_r\|_\psi}, \qquad M_r := (r+1)I_k - \mathbf{1}\mathbf{1}^\top, \tag{C.36}$$

where $r \in ((k-1)^{1/(p-1)}, k-1)$ is the unique solution of

$$c_p(r)\big(r^{p-1} - (k-1)\big) + \epsilon\big(r - (k-1)\big) = 0, \qquad c_p(r) := \big(kr^p + k(k-1)\big)^{(2-p)/p}. \tag{C.37}$$

Moreover,

$$\widehat{W}_\pi^\psi = \widehat{W} \iff \alpha_1^{\min} = \cdots = \alpha_k^{\min}. \tag{C.38}$$

Thus unequal class minimum scales produce a limiting direction distinct from the max-margin direction for every potential in (C.35) *under balanced sampling, while the training loss converges to zero.*

Proof. The potential in (C.35) has the form in Definition 1 with $\Phi(W) = \|W\|_p^L / L$, and it is invariant under independent permutations within every column. We can therefore apply Proposition C.3. The values d_l in (C.20) are identical because the potential is also invariant under column permutations.

When $\pi_l = 1/k$, the matrix B_π is a positive multiple of $I_k - k^{-1}\mathbf{1}\mathbf{1}^\top$. For every permutation matrix P, we have $PB_\pi P^\top = B_\pi$. Invariance of the potential implies $PZ_\pi P^\top = Z_\pi$, so all diagonal entries of Z_π are equal and all off diagonal entries are equal. The entrywise gradient of the potential has the sign of the corresponding entry of Z_π. Since $\nabla\psi(Z_\pi) = B_\pi$, we can write $Z_\pi = bM_r$ for some $b > 0$ and $r > 0$. Equating the ratio between the diagonal and off diagonal gradient magnitudes to $k-1$, we obtain (C.37). Let $F(r)$ denote the left-hand side of (C.37). Its derivative is

$$F'(r) = \frac{c_p(r) r^{p-2}}{r^p + k - 1}\big[r^p + (p-1)(k-1) + (p-2)(k-1)r\big] + \epsilon > 0 \qquad (r > 0).$$

Moreover, $F((k-1)^{1/(p-1)}) < 0$ and $F(k-1) > 0$. The intermediate value theorem and strict monotonicity therefore give a unique root in the stated interval.

By (C.36), every class has the same positive margin factor $h_l = (r+1)/\|M_r\|_\psi$. Substituting into (C.22), we obtain (C.38) and the remaining conclusions. □

Quadratic potentials with column weights.

Corollary C.3. *Use the data, algorithm, initialization, and sampling in Proposition C.3, with $L = 2$ and the potential*

$$\psi_q(W) := \sum_{l=1}^{k} q_l \|We_l\|_2^2, \qquad q_l > 0. \tag{C.39}$$

Then the limiting direction and the corresponding max-margin direction are

$$\widehat{W}_\pi^{\psi_q} = \frac{\sum_{l=1}^k \pi_l q_l^{-1/2} R_l}{\|\pi\|_2}, \qquad \widehat{W} = \frac{\sum_{l=1}^k (\alpha_l^{\min})^{-1} R_l}{\sqrt{\sum_{l=1}^k q_l (\alpha_l^{\min})^{-2}}}. \tag{C.40}$$

In particular,

$$\widehat{W}_\pi^{\psi_q} = \widehat{W} \iff \frac{\alpha_1^{\min}\pi_1}{\sqrt{q_1}} = \cdots = \frac{\alpha_k^{\min}\pi_k}{\sqrt{q_k}}. \tag{C.41}$$

The training loss converges to zero. If (C.41) *fails, the limiting Bregman error is strictly positive and the limiting margin is strictly smaller than $\widehat{\gamma}$.*

Proof. For any $q_l > 0$, choose $0 < \epsilon < 2\min_l q_l$ and set $\Phi(W) = \sum_l (q_l - \epsilon/2)\|We_l\|_2^2$. We then have $\psi_q(W) = \Phi(W) + \epsilon\|W\|_{\mathrm{F}}^2/2$, which is the representation in Definition 1 for $L = 2$. The potential also satisfies (C.19). Its dual norm and inverse mirror map satisfy

$$\|Z\|_{\psi,*}^2 = \sum_{l=1}^k q_l^{-1}\|Ze_l\|_2^2, \qquad \nabla\psi_q^*(\Theta)e_l = \frac{\Theta e_l}{2q_l}.$$

Thus $d_l = q_l^{-1/2}$, and from (C.20), we obtain

$$B_\pi = \sum_{l=1}^k \pi_l\sqrt{q_l}R_l, \qquad Z_\pi = \frac{1}{2}\sum_{l=1}^k \pi_l q_l^{-1/2}R_l, \qquad \|Z_\pi\|_\psi = \frac{1}{2}\|\pi\|_2.$$

Normalizing Z_π, we obtain the first formula in (C.40). For the hard margin problem, the positive scalar q_l does not change the minimum norm feasible column. Hence $W_{\mathrm{hm}} = \kappa_k \sum_l (\alpha_l^{\min})^{-1}R_l$. Normalizing in $\|\cdot\|_\psi$, we obtain the second formula in (C.40). Comparing their coefficients, we obtain (C.41). The loss and gap conclusions follow from Proposition C.3. □

D Proofs of the Results for NMD with Momentum

We compare the momentum with an average of past batch gradients evaluated at the current iterate. Bounded sample coverage gives positive weights for this reference and yields the full-gradient and reshuffling results. For independent sampling, we instead control the weighted frequency of windows that fail to cover every sample.

D.1 Proof of Theorem 3

Assumption D.1. *For every batch $\mathcal{B}_t$ of fixed size s, let $a_{i,t}$ be the multiplicity of sample i divided by s. Thus $a_{i,t} \geq 0$, $\sum_i a_{i,t} = 1$, and $\widehat{g}_{\mathcal{B}_t}(W) = \sum_i a_{i,t}\nabla\ell_i(W)$. There are integers $K \geq 1$, $t_{\mathrm{c}} \geq K-1$, and a constant $\mu > 0$ such that, for every $i \in [n]$ and $t \geq t_{\mathrm{c}}$,*

$$\sum_{r=t-K+1}^{t} a_{i,r} \geq \mu. \tag{D.1}$$

Lemma D.1. *Suppose that Assumptions 1 to 3 and D.1 hold. Let* (7) *use a fixed* $\beta_t = \beta \in (0,1)$. *For every sufficiently large* T, *we have*

$$D_\psi\Big(\widehat{W}, W_T/\left\|W_T\right\|_\psi\Big) \le C_{\beta,K,\mu,t_c}\left[\rho_L(S_T) + \frac{\sum_{t<T}\eta_t\omega_t}{1+S_T}\right]. \tag{D.2}$$

The constant depends only on β, K, μ, t_c *and the standing problem parameters, and is uniform in* T *and the batch sequence. Consequently, the Bregman divergence converges to zero and* $W_T/\left\|W_T\right\|_\psi \to \widehat{W}$. *Moreover,*

$$\gamma(W_T) \longrightarrow \widehat{\gamma}, \qquad \mathcal{L}(W_T) \longrightarrow 0. \tag{D.3}$$

Proof of Lemma D.1. Fix a batch sequence satisfying Assumption D.1. We apply Theorem B.1 using the frozen reference

$$\begin{aligned}\widetilde{M}_t &:= (1-\beta)\sum_{r=0}^{t}\beta^{t-r}\widehat{g}_{\mathcal{B}_r}(W_t)\\ &= \sum_{i=1}^{n} q_{i,t}\nabla\ell_i(W_t),\end{aligned} \tag{D.4}$$

where

$$q_{i,t} = (1-\beta)\sum_{r=0}^{t}\beta^{t-r}a_{i,r}.$$

Using (D.1) for the last K terms and $a_{i,r} \le 1$ for the entire geometric series, we obtain, for every $t \ge t_c$,

$$q_- := (1-\beta)\beta^{K-1}\mu \le q_{i,t} \le q_+ := 1. \tag{D.5}$$

Thus $\widetilde{M}_t$ is the gradient of a positively reweighted cross entropy risk. Applying the weighted part of Lemma B.3 with weights $w_{i,t} = nq_{i,t}$ and (D.5), we obtain

$$\left\langle \frac{\widetilde{M}_t}{\left\|\widetilde{M}_t\right\|_{\psi,*}}, (r_t+\lambda_{\mathrm{cov}})\widehat{W} - W_t \right\rangle \le 0 \tag{D.6}$$

for all $t \ge t_c$, with the fixed choice

$$\lambda_{\mathrm{cov}} = \widehat{\gamma}^{-1}\log(2n(k-1)q_+/q_-)\,.$$

Combining (D.5) with the positive cone argument in the proof of (B.4), we also obtain

$$n\widehat{\gamma}q_- G(W_t) \le \left\|\widetilde{M}_t\right\|_{\psi,*} \le nC_A q_+ G(W_t). \tag{D.7}$$

We next bound the freezing error. By (B.21) and regular variation of ω_t, for $t/2 \le r \le t$, we have

$$\left\|W_t - W_r\right\|_\psi \le C(t-r)\omega_t. \tag{D.8}$$

For arbitrary batch weights $a_{i,r} \le 1$, we have

$$G_{\mathcal{B}_r}(W_t) = \sum_i a_{i,r}\sum_{l\ne y_i} p_{il}(W_t) \le nG(W_t). \tag{D.9}$$

Applying (B.8), (D.8), and (D.9), and then summing the geometric kernel, we obtain, for all large t,

$$\begin{aligned}&(1-\beta)\sum_{r=\lceil t/2\rceil}^{t}\beta^{t-r}\left\|\widehat{g}_{\mathcal{B}_r}(W_r) - \widehat{g}_{\mathcal{B}_r}(W_t)\right\|_{\psi,*}\\ &\quad\le CG(W_t)(1-\beta)\sum_{h\ge 0}\beta^h\big(\mathrm{e}^{Ch\omega_t}-1\big) \le C_\beta\omega_t G(W_t),\end{aligned}$$

where we use $\omega_t \to 0$ to bound the geometric sum for fixed $\beta < 1$. For the older half of the EMA, we obtain an $O(\beta^{t/2})$ contribution from boundedness of the sample gradients. By the softmax formula

and $r_t = O((1+S_t)^{1/(L-1)}) = o(t)$, we have $G(W_t) \geq c\exp(-Cr_t)$. Hence $\beta^{t/2} = o(\omega_t G(W_t))$, so we can absorb the old contribution. Combining the recent and old parts, we obtain

$$\left\| M_t - \widetilde{M}_t \right\|_{\psi,*} \leq C_\beta \omega_t G(W_t). \tag{D.10}$$

By the triangle inequality, (D.7), and (D.10),

$$\left\| \frac{M_t}{\|M_t\|_{\psi,*}} - \frac{\widetilde{M}_t}{\left\|\widetilde{M}_t\right\|_{\psi,*}} \right\|_{\psi,*} \leq \frac{2\left\|M_t - \widetilde{M}_t\right\|_{\psi,*}}{\left\|\widetilde{M}_t\right\|_{\psi,*}} \leq C_{\beta,K,\mu}\omega_t. \tag{D.11}$$

We apply Theorem B.1 with actual signal $\widehat{b}_t = \frac{M_t}{\|M_t\|_{\psi,*}}$, and the pathwise reference

$$b_t^\circ = \begin{cases} \frac{\nabla\mathcal{L}(W_t)}{\|\nabla\mathcal{L}(W_t)\|_{\psi,*}}, & t < t_{\mathrm{c}}, \\ \frac{\widetilde{M}_t}{\left\|\widetilde{M}_t\right\|_{\psi,*}}, & t \geq t_{\mathrm{c}}, \end{cases} \qquad q_t = \frac{M_t}{\|M_t\|_{\psi,*}} - b_t^\circ.$$

By (B.19), the actual direction is bounded and uniformly aligned. Using the global comparison in Lemma B.3 before t_{c} and (D.6) thereafter, we verify the ray premise at every time with the same enlarged λ_{cov}. From $\|Z_t\|_\psi \leq 2$ and (D.11), we have

$$\mathcal{Q}_T^Z \leq C_{\beta,K,\mu,t_{\mathrm{c}}} \left(1 + \sum_{t<T} \eta_t\omega_t\right).$$

Substituting into (B.28), we obtain (D.2). Here the finite prefix contributes $O((1+S_T)^{-1})$, which is absorbed by $\rho_L(S_T)$. Since $\omega_t \to 0$ and $S_T \to \infty$, the weighted Toeplitz lemma gives $\sum_{t<T}\eta_t\omega_t/(1+S_T) \to 0$. Thus the Bregman divergence vanishes. By (B.17), $W_T/\|W_T\|_\psi \to \widehat{W}$, hence $\gamma(W_T) \to \widehat{\gamma} > 0$. Radial divergence in (B.20) then gives

$$\min_{i\in[n],l\neq y_i} m_{il}(W_T) = r_T\gamma(W_T) \longrightarrow \infty.$$

Substituting into (1), we obtain the loss limit in (D.3). □

Full-gradient specialization. We collect the mirror comparison term and the temporal lag term in

$$\mathcal{R}_T := \rho_L(S_T) + \frac{\sum_{t<T}\eta_t\omega_t}{1+S_T}. \tag{D.12}$$

Proof of Theorem 3. For full gradients, $a_{i,t} = 1/n$, so Assumption D.1 holds with $K = 1$, $\mu = 1/n$, and $t_{\mathrm{c}} = 0$. Thus Lemma D.1 gives

$$D_\psi\left(\widehat{W}, W_T/\|W_T\|_\psi\right) \leq C_\beta\mathcal{R}_T.$$

For full gradients, we have

$$\widetilde{M}_t := (1-\beta)\sum_{j=0}^{t}\beta^{t-j}\nabla\mathcal{L}(W_t) = (1-\beta^{t+1})\nabla\mathcal{L}(W_t). \tag{D.13}$$

Hence (D.11) gives $\left\| \frac{M_t}{\|M_t\|_{\psi,*}} - \frac{\nabla\mathcal{L}(W_t)}{\|\nabla\mathcal{L}(W_t)\|_{\psi,*}} \right\|_{\psi,*} \leq C_\beta\omega_t$.

For $0 < a < 1$, Assumption 3 and the definitions of S_T and ω_t imply

$$S_T = \Theta(T^{1-a}), \qquad \eta_t\omega_t = \Theta\left(t^{-\frac{L-2+La}{L-1}}\right).$$

By integral comparison, we obtain

$$\frac{\sum_{t<T}\eta_t\omega_t}{1+S_T} = \begin{cases} O\left(T^{-\frac{L-2+a}{L-1}}\right), & 0 < a < 1/L, \\ O\left(T^{-(1-a)}\log T\right), & a = 1/L, \\ O\left(T^{-(1-a)}\right), & 1/L < a < 1. \end{cases} \tag{D.14}$$

For $L > 2$, the logarithmic term at $a = 1/L$ decays faster than $\rho_L(S_T)$. For $a = 1$, we have $S_T = \Theta(\log T)$ and $\sum_{t<T} \eta_t \omega_t = O(1)$, so the lag term is $O((\log T)^{-1})$. Substituting (D.14) and the $a = 1$ estimate into (D.12), and using the definition of ρ_L, we obtain the Bregman rates in (11).

By (B.17), we have

$$0 \le \widehat{\gamma} - \gamma(W_T) \le C_A K_\psi \sqrt{D_\psi\Big(\widehat{W}, W_T / \|W_T\|_\psi\Big)} \le C_\beta \sqrt{\mathcal{R}_T},$$

where the first inequality follows from (4). The directional and loss limits follow from Lemma D.1. □

D.2 Proof of Theorem 4

We use the epoch partitions defined in Assumption B.1. Every realization of Assumption 5 satisfies that assumption.

Part (ii): epoch sampling.

Proof of part (ii) of Theorem 4. Every trailing window of $2m$ batches, once available, contains a complete epoch. We therefore verify Assumption D.1 with $K = 2m$, $\mu = 1/s$, and $t_c = 2m - 1$. Applying Lemma D.1 and (B.17), we obtain

$$\begin{aligned} D_\psi\Big(\widehat{W}, W_T / \|W_T\|_\psi\Big) &\le C_{\beta,m} \mathcal{R}_T, \\ 0 \le \widehat{\gamma} - \gamma(W_T) &\le C_{\beta,m} \sqrt{\mathcal{R}_T}, \end{aligned} \tag{D.15}$$

The constant is uniform in the batch sequence and T. Evaluating $\mathcal{R}_T$ as in Section D.1, including $a = 1$, gives (13); Lemma D.1 gives the directional and loss limits. □

Part (i): independent sampling with replacement.

Intermittent coverage.

Lemma D.2. *Suppose that Assumptions 1 to 3 hold. Use* (7) *with fixed $\beta \in (0, 1)$ and an arbitrary sequence of batches of fixed size s, allowing repeated indices. For an integer $K \ge 1$, define*

$$J_{t,K} := \begin{cases} 1, & t < K - 1, \\ \mathbf{1}\left\{ \bigcup_{j=t-K+1}^{t} \mathcal{B}_j \ne [n] \right\}, & t \ge K - 1. \end{cases} \tag{D.16}$$

Define

$$q_K := \frac{(1-\beta)\beta^{K-1}}{s}, \qquad \lambda_K := \widehat{\gamma}^{-1} \log\left(\frac{2n(k-1)}{q_K}\right), \qquad \mathcal{A}_T := \frac{1 + \sum_{t<T} \eta_t \omega_t}{1 + S_T}. \tag{D.17}$$

There are finite constants C_β, T_β, independent of K, s and the batch sequence, such that for every $K \ge 1$ and $T \ge T_\beta$,

$$\begin{aligned} D_\psi\Big(\widehat{W}, W_T / \|W_T\|_\psi\Big) &\le C_\beta \left[(1 + \lambda_K)\rho_L(S_T) + q_K^{-1} \mathcal{A}_T\right] + \frac{C_0}{1 + S_T} \sum_{t<T} \eta_t J_{t,K}, \\ C_0 &:= \frac{8LC_A}{\widehat{\gamma}}. \end{aligned} \tag{D.18}$$

In particular, for each fixed K, a finite constant C_K depending on β, s, K and the standing constants satisfies

$$D_\psi\Big(\widehat{W}, W_T / \|W_T\|_\psi\Big) \le C_K \left[\rho_L(S_T) + \frac{\sum_{t<T} \eta_t \omega_t}{1 + S_T}\right] + \frac{C_0}{1 + S_T} \sum_{t<T} \eta_t J_{t,K}. \tag{D.19}$$

The coefficient C_0 is independent of K, s, β.

Proof. By (B.19), the actual direction $M_t/\left\|M_t\right\|_{\psi,*}$ has unit dual norm and $\left\langle M_t/\left\|M_t\right\|_{\psi,*}, \widehat{W}\right\rangle \le -\widehat{\gamma}/C_A$, uniformly in the batch sequence and β. Applying Lemma B.7, we obtain (B.20)–(B.22). Writing $r_t = \|W_t\|_\psi$, the exact radius identity (B.14) also implies, for all sufficiently large t,

$$r_t^{L-1} = \frac{\|\Theta_t\|_{\psi,*}}{L} \ge \frac{\widehat{\gamma}}{2LC_A}(1+S_t). \tag{D.20}$$

Indeed, summing the aligned updates, we have $\|\Theta_t\|_{\psi,*} \ge \left\langle \Theta_0, \widehat{W}\right\rangle + (\widehat{\gamma}/C_A)S_t$. The finite starting time in (D.20) can be chosen independently of K, s, β.

Define $\widetilde{M}_t$ by (D.4), with empirical batch weights $a_{i,j}$. Unrolling the memory gives

$$\widetilde{M}_t = \sum_i q_{i,t}\nabla\ell_i(W_t), \qquad q_{i,t} = (1-\beta)\sum_{j=0}^{t}\beta^{t-j}a_{i,j}, \qquad 0 \le q_{i,t} \le 1. \tag{D.21}$$

The positive cone argument gives $\widetilde{M}_t \neq 0$. The derivation of (D.10), using (D.8), (D.9), bounded sample gradients, and $G(W_t) \ge c\mathrm{e}^{-Cr_t}$, yields

$$\left\|M_t - \widetilde{M}_t\right\|_{\psi,*} \le C_\beta \omega_t G(W_t) \tag{D.22}$$

with constants and starting time independent of K, s and the batch sequence.

Fix K. If $J_{t,K} = 0$, every sample occurs at least once in the last K batches, so (D.21) implies

$$q_{i,t} \ge q_K := \frac{(1-\beta)\beta^{K-1}}{s}. \tag{D.23}$$

The weighted ray comparison in Lemma B.3, applied with weights $nq_{i,t}$, therefore holds with the offset λ_K in (D.17). By the cone estimate, we also have $\left\|\widetilde{M}_t\right\|_{\psi,*} \ge n\widehat{\gamma}q_K G(W_t)$ on these windows. Applying normalization stability to (D.22), we obtain

$$\left\| \frac{M_t}{\|M_t\|_{\psi,*}} - \frac{\widetilde{M}_t}{\left\|\widetilde{M}_t\right\|_{\psi,*}} \right\|_{\psi,*} \le C_\beta q_K^{-1}\omega_t \qquad \text{when } J_{t,K} = 0.$$

At uncovered times we use $\frac{\nabla\mathcal{L}(W_t)}{\|\nabla\mathcal{L}(W_t)\|_{\psi,*}}$ as the reference. More precisely, set

$$b_t^\circ = \begin{cases} \frac{\widetilde{M}_t}{\left\|\widetilde{M}_t\right\|_{\psi,*}}, & J_{t,K} = 0, \\ \frac{\nabla\mathcal{L}(W_t)}{\|\nabla\mathcal{L}(W_t)\|_{\psi,*}}, & J_{t,K} = 1, \end{cases} \qquad e_t = \frac{M_t}{\|M_t\|_{\psi,*}} - b_t^\circ, \qquad Z_t^{(K)} = \frac{W_t}{r_t + \lambda_K} - \widehat{W}.$$

Since $\lambda_K \ge \lambda_{\mathrm{ce}}$, both references satisfy (B.24) with offset λ_K. For all t after a fixed starting time independent of K, s, we have

$$\|e_t\|_{\psi,*} \le C_\beta q_K^{-1}\omega_t + 2J_{t,K}, \qquad \left\|Z_t^{(K)}\right\|_\psi \le 2. \tag{D.24}$$

Applying (B.33) with offset λ_K, we obtain

$$\mathcal{E}_{t+1}^{\mathrm{dir}} - \mathcal{E}_t^{\mathrm{dir}} \le \mathcal{R}_{t,K}^{\mathrm{fs}} - \eta_t\left\langle e_t, Z_t^{(K)}\right\rangle.$$

To check the uniformity in K, write $x = r_t$, $y = r_{t+1}$. After a starting time independent of K, s, β, radial growth and (B.21) ensure $x > 0$ and $y \in [x/2, 3x/2]$. For every $\lambda \ge \lambda_{\mathrm{ce}} > 0$, the integral in the finite step remainder satisfies

$$\left| \frac{L(L-1)}{x+\lambda}\int_x^y z^{L-2}(x+\lambda-z)\,\mathrm{d}z \right| \le Cx^{L-3}\big(\lambda|y-x| + |y-x|^2\big).$$

The constant is independent of λ. Applying (B.21)–(B.22), with $\lambda = \lambda_K$, we obtain

$$|\mathcal{R}_{t,K}^{\mathrm{fs}}| \le C\left(\frac{\lambda_K\eta_t}{r_t} + \frac{\eta_t^2}{r_t^{L-1}}\right).$$

Moreover, $\sum_t \eta_t^2 / r_t^{L-1} < \infty$ after this prefix, because $r_t^{L-1} \asymp 1 + S_t$. On the prefix, bounded dual increments and $r_t + \lambda_K \geq \lambda_{\mathrm{ce}}$ bound the finite step remainders uniformly in K, s. Normalized directions have dual norm one, so their prefix work is uniformly bounded as well. Applying (B.35) and (D.24), we consequently obtain

$$\begin{aligned} \mathcal{E}_T^{\mathrm{dir}} &\leq C(1+\lambda_K)(1+S_T)\rho_L(S_T) \\ &\quad + C_\beta q_K^{-1}\left(1 + \sum_{t<T} \eta_t \omega_t\right) + 4\sum_{t<T} \eta_t J_{t,K}. \end{aligned} \tag{D.25}$$

The constants and starting time are uniform in K, s and the batch sequence. Dividing (D.25) by r_T^{L-1}, and using (B.31) and (D.20), we obtain (D.18), with $C_0 = 8LC_A/\widehat{\gamma}$. Since $(1+S_T)^{-1} \leq C\rho_L(S_T)$, fixing K also proves (D.19). □

Proof of the convergence conclusions in part (i) of Theorem 4. Write $p_{\min} := \min_{i\in[n]} p_i > 0$. Fix an integer $K \geq 1$. Under Assumption 4 with fixed batch size s, the batches are independent and identically distributed. For $t \geq K-1$, let π_K denote the common mean of the indicator in (D.16). The probability that sample i is absent from the sK independent draws equals $(1-p_i)^{sK}$. Applying a union bound over the samples, we obtain

$$\pi_K \leq \sum_{i=1}^{n} (1-p_i)^{sK} \leq n(1-p_{\min})^{sK}. \tag{D.26}$$

For each residue class modulo K, the windows are disjoint, so their indicators are i.i.d. after a finite prefix. The strong law gives $\sum_{t<T}(J_{t,K} - \pi_K) = o(T)$ almost surely. Since η_t is nonincreasing and $S_T \to \infty$, summation by parts then gives $\sum_{t<T} \eta_t (J_{t,K} - \pi_K) = o(S_T)$. Thus

$$\frac{1}{1+S_T} \sum_{t<T} \eta_t J_{t,K} \longrightarrow \pi_K \qquad \text{almost surely.} \tag{D.27}$$

This also covers $a = 1$.

We intersect the probability one events in (D.27) over the countable set $K = 1, 2, \ldots$. On that event, fix K and apply Lemma D.2. Since $\rho_L(S_T) \to 0$, $\omega_t \to 0$, and the weighted Toeplitz lemma implies $(1+S_T)^{-1}\sum_{t<T} \eta_t \omega_t \to 0$, we obtain

$$\limsup_{T\to\infty} D_\psi\left(\widehat{W}, W_T/\left\|W_T\right\|_\psi\right) \leq C_0 \pi_K \leq C_0 n(1-p_{\min})^{sK} \qquad \text{for every } K.$$

The coefficient C_0 is independent of K. Since $p_{\min} > 0$, letting $K \to \infty$, we obtain zero Bregman error. By (B.17), $W_T/\left\|W_T\right\|_\psi \to \widehat{W}$ and $\gamma(W_T) \to \widehat{\gamma} > 0$. Finally, radial divergence in (B.20) and (1) imply $\mathcal{L}(W_T) \to 0$. □

Proposition D.1. *Suppose that Assumptions 1 to 4 hold. Consider* (7) *with arbitrary initialization, fixed batch size s, and fixed momentum parameter $\beta \in (0,1)$. Let $1 \leq K \leq T$ be deterministic integers, $T \geq T_\beta$, and $\delta \in (0,1)$. Write $\pi_K = \mathbb{P}(J_{t,K} = 1)$ for $t \geq K-1$. With probability at least $1-\delta$,*

$$\begin{aligned} D_\psi\left(\widehat{W}, W_T/\left\|W_T\right\|_\psi\right) &\leq C_\beta\left[(1+\lambda_K)\rho_L(S_T) + q_K^{-1}\mathcal{A}_T\right] \\ &\quad + C_0\left[\pi_K + \frac{S_{K-1} + \sqrt{\frac{1}{2}KQ_T\log(K/\delta)}}{1+S_T}\right]. \end{aligned} \tag{D.28}$$

Here $\pi_K \leq n(1-p_{\min})^{sK}$ by (D.26)*. The constants, window scales, and starting time are those of Lemma D.2.*

Proof. For $r = 0, \ldots, K-1$, define

$$\mathcal{I}_r := \{t : K-1 \leq t < T,\ t \equiv r \pmod K\}, \qquad Q_{T,r} := \sum_{t\in\mathcal{I}_r} \eta_t^2.$$

Within each $\mathcal{I}_r$, the covering windows are disjoint. Hence their indicators are independent, have mean π_K, and take values in $[0,1]$. Weighted Hoeffding concentration implies

$$\mathbb{P}\left(\sum_{t\in\mathcal{I}_r}\eta_t(J_{t,K}-\pi_K) > \sqrt{\tfrac{1}{2}Q_{T,r}\log(K/\delta)}\right) \le \frac{\delta}{K}.$$

A union bound and Cauchy–Schwarz imply, on an event of probability at least $1-\delta$,

$$\sum_{t=K-1}^{T-1}\eta_t(J_{t,K}-\pi_K) \le \sqrt{\tfrac{1}{2}\log(K/\delta)}\sum_{r=0}^{K-1}\sqrt{Q_{T,r}} \le \sqrt{\tfrac{1}{2}KQ_T\log(K/\delta)}.$$

The first $K-1$ indicators equal one by definition and contribute S_{K-1}. Thus

$$\sum_{t<T}\eta_t J_{t,K} \le \pi_K S_T + S_{K-1} + \sqrt{\tfrac{1}{2}KQ_T\log(K/\delta)}.$$

Substitution into (D.18), together with (D.26), proves (D.28). □

Rates under independent sampling with replacement.

Corollary D.1. *Suppose that Assumptions 1 to 4 hold with $n \ge 2$. Consider* (7) *with arbitrary initialization, fixed batch size s, and fixed momentum parameter $\beta \in (0,1)$. Let $\chi_{p,s} := -s\log(1-p_{\min})$. For $0<a<1$, define*

$$\begin{aligned}\zeta_{L,a} &:= \min\left\{\frac{L-2+a}{L-1}, 1-a\right\},\\ r_{\mathrm{cov}} &:= \min\left\{\frac{1-a}{L-1}, \frac{\chi_{p,s}\zeta_{L,a}}{\chi_{p,s}+\log(1/\beta)}\right\}.\end{aligned} \tag{D.29}$$

Then, almost surely,

$$D_\psi\Big(\widehat{W}, W_T/\left\|W_T\right\|_\psi\Big) = \widetilde{O}(T^{-r_{\mathrm{cov}}}), \qquad 0 \le \widehat{\gamma}-\gamma(W_T) = \widetilde{O}(T^{-r_{\mathrm{cov}}/2}). \tag{D.30}$$

When $a=1$, for every $0<\kappa<\min\{1/(L-1), \chi_{p,s}/(\chi_{p,s}+\log(1/\beta))\}$, almost surely,

$$D_\psi\Big(\widehat{W}, W_T/\left\|W_T\right\|_\psi\Big) = O\big((\log T)^{-\kappa}\big), \qquad \widehat{\gamma}-\gamma(W_T) = O\Big((\log T)^{-\kappa/2}\Big). \tag{D.31}$$

Proof of Corollary D.1. Write $\chi = \chi_{p,s}$ and $\ell_\beta = \log(1/\beta)$. We first consider $0<a<1$ and write $\zeta = \zeta_{L,a}$. For each deterministic horizon T, we choose

$$K_T := \min\left\{T, \max\left\{1, \left\lceil \frac{\zeta}{\chi+\ell_\beta}\log(\mathrm{e}+T)\right\rceil\right\}\right\}. \tag{D.32}$$

For $0<a<1$, we write $\vartheta_{L,a} := (L-2+a)/(L-1)$. By Assumption 3 and the definitions of S_T and ω_t, we have

$$S_T \asymp T^{1-a}, \qquad \omega_T \asymp T^{-\vartheta_{L,a}}, \qquad \mathcal{A}_T = \widetilde{O}(T^{-\zeta}).$$

The last estimate follows by summing $t^{-a-\vartheta_{L,a}}$, with a logarithmic factor at $a+\vartheta_{L,a}=1$. By (D.17) and (D.32), we have

$$\begin{aligned} q_{K_T}^{-1} &= \frac{s}{1-\beta}\mathrm{e}^{\ell_\beta(K_T-1)} = O\Big(T^{\ell_\beta\zeta/(\chi+\ell_\beta)}\Big),\\ \pi_{K_T} &\le n\mathrm{e}^{-\chi K_T} = O\Big(T^{-\chi\zeta/(\chi+\ell_\beta)}\Big), \qquad \lambda_{K_T} = O(\log T).\end{aligned}$$

Consequently, both $q_{K_T}^{-1}\mathcal{A}_T$ and π_{K_T} are $\widetilde{O}(T^{-\chi\zeta/(\chi+\ell_\beta)})$. Also,

$$(1+\lambda_{K_T})\rho_L(S_T) = \widetilde{O}\Big(T^{-(1-a)/(L-1)}\Big).$$

We apply Proposition D.1 with $K=K_T$ and $\delta_T=(T+1)^{-2}$. By integral comparison of the sums defining S_T, Q_T, we obtain

$$\frac{S_{K_T-1}+\sqrt{\frac{1}{2}K_TQ_T\log(K_T/\delta_T)}}{1+S_T} = \widetilde{O}\Big(T^{-\min\{1/2,1-a\}}\Big).$$

This term is no slower than $T^{-r_{\mathrm{cov}}}$. Indeed, if $\xi = (1-a)/(L-1)$, then $\vartheta_{L,a} = 1-\xi$, and

$$r_{\mathrm{cov}} \le \min\{\xi, \zeta\} \le \min\{\xi, 1-\xi, 1-a\} \le \min\{1/2, 1-a\}.$$

Substitution into (D.28) and the first Borel–Cantelli lemma yield the Bregman rate almost surely, since $\sum_T \delta_T < \infty$. Equation (B.17) gives the margin rate in (D.30). For every $0 < \kappa < r_{\mathrm{cov}}$, we absorb the logarithmic factors into $T^{r_{\mathrm{cov}}-\kappa}$ and obtain (12).

We next consider $a = 1$. For integers $j \ge 1$, we choose

$$T_j := 2^j, \qquad K_j := \max\left\{1, \left\lceil \frac{\log(\mathrm{e}+j)}{\chi + \ell_\beta} \right\rceil \right\}, \qquad \delta_j := (j+1)^{-2}. \tag{D.33}$$

For all sufficiently large j, we have $K_j \le T_j$. Applying the concentration estimate in the proof of Proposition D.1 at horizon T_{j+1} with window K_j, we obtain, with probability at least $1-\delta_j$,

$$\sum_{t<T_{j+1}} \eta_t J_{t,K_j} \le \pi_{K_j} S_{T_{j+1}} + S_{K_j-1} + \sqrt{\tfrac{1}{2} K_j Q_{T_{j+1}} \log(K_j/\delta_j)}. \tag{D.34}$$

Since $\sum_j \delta_j < \infty$, the first Borel–Cantelli lemma implies that (D.34) holds for all sufficiently large j, almost surely.

Let $b := \chi/(\chi+\ell_\beta) \in (0,1)$. By Assumption 3, the definitions of S_T and Q_T, and (D.33), we have $1 + S_T \asymp j$ uniformly for $T_j \le T < T_{j+1}$, $Q_{T_{j+1}} = O(1)$, and $K_j = O(\log(\mathrm{e}+j))$. Using the nonnegativity of J_{t,K_j}, (D.26), and (D.34), we obtain, almost surely,

$$\frac{\sum_{t<T} \eta_t J_{t,K_j}}{1+S_T} = O\left(j^{-b} + \frac{\log(\mathrm{e}+j)}{j} \right) \qquad (T_j \le T < T_{j+1}). \tag{D.35}$$

This estimate is uniform over each interval because its numerator is bounded by the sum through $T_{j+1}-1$, while its denominator is bounded below by $1 + S_{T_j}$.

By the definition of ω_t and (D.17), we have $\omega_t = O(1/t)$ and $\sum_{t<T} \eta_t \omega_t = O(1)$ when $a = 1$. Combining the definition of ρ_L, (D.17), and (D.33), we obtain, uniformly for $T_j \le T < T_{j+1}$,

$$q_{K_j}^{-1} \mathcal{A}_T = O(j^{-b}), \qquad (1+\lambda_{K_j}) \rho_L(S_T) = O\left((\log(\mathrm{e}+j))^2 j^{-1/(L-1)} \right). \tag{D.36}$$

Substituting (D.35) and (D.36) into (D.18), we obtain, almost surely,

$$D_\psi\left(\widehat{W}, W_T / \|W_T\|_\psi \right) = O\left(j^{-b} + (\log(\mathrm{e}+j))^2 j^{-1/(L-1)} \right) \qquad (T_j \le T < T_{j+1}).$$

For every $0 < \kappa < \min\{1/(L-1), b\}$, we absorb the logarithmic factor and use $\log T \asymp j$ to obtain the Bregman rate in (D.31). Applying (B.17), we obtain its margin rate. When $n = 1$, every batch gradient equals the full gradient, so Theorem 3 applies directly. This completes the rate proof for part (i) of Theorem 4. □