IMPLICIT BIAS OF MUON WITH FINITE NEWTON–SCHULZ ITERATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon has recently attracted growing interest as an effective optimizer for large-scale neural network training. However, the theoretical mechanisms underlying Muon, especially the role of Newton–Schulz (NS) iterations in practice, are not yet fully understood. We focus on Muon's implicit bias with finite NS iterations in multi-class classification. We characterize the gap to the optimal spectral margin through a margin bound, which decomposes the gap into a vanishing training term and an additional path-weighted distortion arising from NS iterations. Persistent distortion can leave a nonvanishing margin gap, whereas vanishing weighted distortion guarantees recovery of the optimal spectral margin. This characterization clarifies when finite NS iterations alter or preserve Muon's spectral implicit bias, providing a principled understanding of their role in practical Muon implementations. Finally, extensive experiments across synthetic and real-data settings validate our theoretical findings.

1 INTRODUCTION

Large language models (LLMs) have achieved strong performance across a broad range of language tasks (Brown et al., 2020). Their increasing scale makes efficient training a central concern (Hoffmann et al., 2022). Optimizer choice affects training efficiency, stability, and the quality of the resulting model. Adam and AdamW have long been standard choices for large-scale training (Kingma & Ba, 2015; Loshchilov & Hutter, 2019). Recently, Muon has emerged as a promising alternative, with strong empirical results extending to large-scale LLM training (Jordan et al., 2024; Liu et al., 2025). These results motivate a theoretical understanding of the solutions selected by Muon.

Implicit bias describes how an optimizer selects among solutions that achieve the same limiting training loss. In overparameterized models, different optimizers can drive the loss to zero while selecting different solutions (Soudry et al., 2018; Gunasekar et al., 2018b). For Muon, Fan et al. (2025) establish spectral max-margin selection in separable multiclass linear classification. Gronich & Vardi (2026) extend the analysis of momentum-based spectral geometry to smooth homogeneous neural networks, relating its limiting bias to stationary points of a margin-maximization problem. Li et al. (2026a) study how mini-batch sampling and momentum affect implicit bias under general norm geometries, including the spectral geometry associated with Muon. These analyses model Muon through exact spectral-norm steepest-descent geometry, with matrix directions determined by the polar factor. However, practical Muon approximates the polar factor using finitely many Newton–Schulz (NS) or related polynomial iterations (Jordan et al., 2024; Amsel et al., 2025). In particular, even when the training loss vanishes, it is not known whether finite NS selects the same limiting direction.

To bridge this theory-practice gap, we address the following central question:

How do Newton–Schulz iterations affect the implicit bias of Muon?

We answer this question in separable multiclass linear classification, where the matrix-valued parameterization makes Muon's spectral geometry explicit. This setting permits a direct analysis of the classifier selected as the training loss vanishes. We quantify finite-NS distortion through the realized update amplitude and relative dual-alignment defect, whose accumulation controls the spectral-margin gap. A fixed finite depth can alter the selected spectral margin, whereas calibrated or

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.

adaptive precision can preserve the exact-polar bias. This pathwise analysis distinguishes convergence of the training loss from recovery of the solution geometry associated with exact-polar Muon.

Our Contributions. We characterize when finite Newton–Schulz iterations alter or preserve Muon's spectral implicit bias. First, we derive a finite-time spectral-margin bound that separates two sources of error: a training term that vanishes with optimization, and a path-weighted directional-distortion term that captures how finite NS errors accumulate along the training trajectory. We then characterize when this distortion changes Muon's implicit bias. For every fixed finite depth q, we construct a separable problem where the loss vanishes but the limiting spectral margin retains a persistent $\Theta((2/3)^q)$ gap. Thus, persistent distortion can alter the limiting margin even after successful optimization, whereas vanishing weighted distortion is sufficient to recover the optimal spectral margin. For recovery, we identify two increasing-precision regimes. Horizon-calibrated depth matches the finite-horizon training-time order of exact-polar Muon over a prescribed horizon, whereas adaptive depth recovers the optimal spectral margin along a single continuing trajectory. Finally, linear experiments with certified spectral-margin references test the predicted separation and recovery, while fully connected neural network and Transformer diagnostics examine depth effects beyond linear classification.

2 Related Work

Muon optimizer. Muon is a matrix-aware optimizer for matrix-valued parameters (Jordan et al., 2024). Its idealized update uses the polar factor of the momentum matrix, connecting the optimizer to spectral-norm steepest descent (Bernstein & Newhouse, 2024). Adaptive and matrix-preconditioned methods provide broader optimization context (Duchi et al., 2011; Kingma & Ba, 2015; Gupta et al., 2018), while recent Muon studies examine momentum filtering, spectral dynamics, scaling, and grokking (Li et al., 2026b; Beneventano et al., 2026; Magakyan et al., 2026; Wang, 2026). Together, these studies clarify how Muon's matrix geometry shapes its optimization dynamics across different training regimes. They also motivate understanding the solution geometry selected by the optimizer, beyond its effect on training progress. We examine how its finite NS implementation affects this solution selection as the training loss vanishes.

Implicit bias in linear classification. For separable linear classification, gradient methods can drive the loss to zero while selecting max-margin directions determined by the optimization geometry (Soudry et al., 2018; Gunasekar et al., 2018b; Nacson et al., 2019a; Ji & Telgarsky, 2021). Closely related extensions treat stochasticity, nonseparable data, finite-time risk, early stopping, and multiclass generalization (Nacson et al., 2019b; Ji & Telgarsky, 2019b; Schliserman & Koren, 2023; Shamir, 2021; Wu et al., 2025; Buna et al., 2026; Tewari & Bartlett, 2007; Maurer, 2016). Exact spectral descent and its momentum counterpart select spectral max-margin solutions in multiclass classification, while norm-induced steepest descent has also been studied with sampling and momentum (Fan et al., 2025; Li et al., 2026a). Results for homogeneous networks, deep linear or convolutional models, matrix factorization, wide networks, and adaptive methods provide broader context for our neural diagnostics (Lyu & Li, 2020; Ji & Telgarsky, 2020; Tsilivis et al., 2025; Vardi et al., 2022; Ji & Telgarsky, 2019a; Gunasekar et al., 2018a; 2017; Arora et al., 2019; Chizat & Bach, 2020; Razin & Cohen, 2020; Wang et al., 2021; Gronich & Vardi, 2026). We study how finite NS updates change the max-margin solutions selected in linear classification. Taking the exact spectral target as the reference, we determine how the finite polynomial map changes its selection and when increasing precision recovers it.

Theory of Muon with NS iterations. Practical Muon replaces the exact polar factor by a small number of Newton–Schulz or related polynomial iterations (Jordan et al., 2024; Amsel et al., 2026). Recent theory establishes convergence and stationarity guarantees for Muon with finite NS steps, quantifies how its optimization constants approach those of the SVD-based update (Kim & Oh, 2026), and analyzes inexact polar updates (Shulgin et al., 2026). Finite NS has also been shown to smooth the polar map in nonsmooth nonconvex optimization (Li & Tsuchiya, 2026). Complementary work studies adaptive NS-budget allocation, implementation error, and stochastic Muon with Nesterov momentum and inexact polar decomposition (Zhuang et al., 2026; Do et al., 2026; Choudhury et al., 2026). These results explain optimization or approximation effects of finite NS. Our focus is different: we ask whether the finite polynomial dynamics recover the spectral max-margin bias after the training

loss vanishes. The pathwise directional distortion provides the link between NS precision and the spectral margin selected by the training trajectory.

3 PRELIMINARIES

Notations. Matrices are generally denoted by uppercase letters and vectors by lowercase letters; scalar quantities use Roman or Greek letters as specified below. For matrices A and B of the same shape, we write $\langle A, B\rangle = \mathrm{tr}(A^\top B)$, and denote the Frobenius, spectral, and nuclear norms of A by $\|A\|_{\mathrm{F}}$, $\|A\|_{\mathrm{op}}$, and $\|A\|_*$, respectively. If $A = U_r\Sigma_r V_r^\top$ is a compact singular value decomposition, we write $\mathrm{polar}(A) := U_r V_r^\top$, with $\mathrm{polar}(0) = 0$. The nuclear norm is dual to the spectral norm, namely $\|A\|_* = \max_{\|B\|_{\mathrm{op}}\le 1}\langle A, B\rangle$. For a positive integer r, let $[r] := \{1, \dots, r\}$, and write $\mathbb{N}_0 := \{0, 1, 2, \dots\}$. For a vector $v \in \mathbb{R}^d$, $\|v\|_1 := \sum_{j=1}^d |v_j|$ denotes its ℓ_1 norm. All logarithms are natural. We use $O(\cdot)$, $o(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$ in their standard asymptotic sense, write $x_t \lesssim y_t$ if $x_t \le Cy_t$ for some constant $C > 0$, define $\gtrsim$ analogously, and write $x_t \asymp y_t$ when both relations hold. Unless stated otherwise, hidden constants may depend on the data, initialization, and fixed optimizer parameters, but not on the iteration index or training horizon.

Setup. Consider a multiclass linear classification problem with samples $\{(h_i, y_i)\}_{i=1}^n$, where $h_i \in \mathbb{R}^d$, $y_i \in [k]$, and $k \ge 2$. Here, k is the number of classes and d is the feature dimension. For $c \in [k]$, let $e_c \in \mathbb{R}^k$ denote the cth standard basis vector. We use a linear classifier $f_W(h) = Wh$ parameterized by $W \in \mathbb{R}^{k\times d}$. We assume that every class contains at least one training example. For each example i and competing class $c \ne y_i$, define the pairwise class-score margin by

$$z_{ic}(W) := (Wh_i)_{y_i} - (Wh_i)_c = (e_{y_i} - e_c)^\top Wh_i.$$

It is the difference between the score assigned to the correct class and the score assigned to class c.

We use the following cross-entropy loss to train the classifier:

$$\mathcal{L}(W) = \frac{1}{n}\sum_{i=1}^n \log\left(1 + \sum_{c\ne y_i} e^{-z_{ic}(W)}\right). \tag{1}$$

For nonzero W, define the optimal and normalized spectral margins as

$$\gamma_* := \max_{\|W\|_{\mathrm{op}}\le 1} \min_{i\in[n],\, c\ne y_i} z_{ic}(W), \qquad \widehat{\gamma}(W) := \min_{i\in[n],\, c\ne y_i} \frac{z_{ic}(W)}{\|W\|_{\mathrm{op}}}, \quad W \ne 0. \tag{2}$$

Here, γ_* is the largest minimum class-score margin among matrices with spectral norm at most one; it is our spectral max-margin target. The quantity $\widehat{\gamma}(W)$ is the minimum margin of W divided by its spectral norm. We further write $p := \min\{k-1, d\}$. Since every cross-entropy gradient sums to zero across its class rows, the gradient and momentum matrices have rank at most p; this effective rank enters the uniform finite-NS distortion bounds below.

Muon. Let $G_t := \nabla\mathcal{L}(W_t)$ be the gradient at iteration t. For a momentum parameter $\beta \in [0, 1)$, initialize $M_{-1} = 0$ and consider the Muon updates

$$\begin{aligned} M_t &= \beta M_{t-1} + (1-\beta)G_t, \\ D_t &= \mathrm{NS}(M_t), \\ W_{t+1} &= W_t - \eta_t D_t, \end{aligned} \tag{3}$$

where $\eta_t > 0$ is the learning rate. At iteration t, the NS map defined below uses $q_t \in \mathbb{N}_0$ rounds. When $\beta = 0$, NS is applied directly to G_t.

Newton–Schulz (NS). For $M \ne 0$ and $q \in \mathbb{N}_0$, initialize $X_0 = M/\|M\|_{\mathrm{F}}$. For $j = 0, \dots, q-1$, apply the cubic recursion and return its final iterate:

$$X_{j+1} = \frac{3}{2}X_j - \frac{1}{2}X_jX_j^\top X_j, \qquad \mathrm{NS}(M) := X_q. \tag{4}$$

Choosing $q = 0$ gives $\mathrm{NS}(M) = M/\|M\|_{\mathrm{F}}$; we set $\mathrm{NS}(0) = 0$. If $M = U\Sigma V^\top$, each round applies $\phi(s) = \frac{3}{2}s - \frac{1}{2}s^3$ to the singular values of the current iterate. The output therefore applies $\psi_q := \phi^{\circ q}$ to the Frobenius-normalized singular values of M.

The polynomial ψ_q is odd and strictly increasing on $[0, 1]$, satisfies $s \le \psi_q(s) \le 1$ there, and has a nonincreasing ratio $\psi_q(s)/s$ on $(0, 1]$. Thus finite NS preserves the singular subspaces while changing their relative amplitudes. We analyze the cubic iteration in (4); Theorem B.1 in Appendix B.2 extends the comparison to admissible Taylor–NS polynomials. For comparison, $\mathrm{polar}(M)$ maximizes $\langle M, D\rangle$ over $\|D\|_{\mathrm{op}} \le 1$, with $\langle M, \mathrm{polar}(M)\rangle = \|M\|_*$.

We distinguish three choices of Newton–Schulz depth: (i) **fixed depth**, where one nonnegative integer q is used at every update in a run; (ii) **horizon-calibrated depth**, where one depth is chosen for a prescribed training horizon and kept fixed throughout that run; and (iii) **adaptive depth**, where the depth may vary along one continuing training trajectory. The main results quantify these regimes.

4 Main Results

In this section, we characterize how finite NS depth changes the spectral margin and when increasing depth recovers the exact-polar bias.

4.1 Pathwise Spectral-Margin Bound

To control the training dynamics and relate the loss to the spectral margin, we adopt standard assumptions on linear separability and bounded inputs. These are consistent with Fan et al. (2025, Assumptions 1 and 4) and Li et al. (2026a, Assumptions 3.1 and 3.2).

Assumption 4.1 (Linear separability)**.** There exists $\overline{W} \in \mathbb{R}^{k\times d}$ such that $z_{ic}(\overline{W}) > 0$ for every $i \in [n]$ and $c \in [k] \setminus \{y_i\}$.

Assumption 4.2 (Bounded inputs)**.** There exists $R > 0$ such that $\|h_i\|_1 \le R$ for every $i \in [n]$.

Separability ensures $\gamma_* > 0$, while bounded inputs control changes in pairwise margins. A finite NS update can differ from the polar factor in both magnitude and alignment with momentum. Positive rescaling changes only the update magnitude. We therefore use the operator-normalized quantities

$$e_t := 1 - \frac{\langle M_t, D_t\rangle}{\|M_t\|_*\|D_t\|_{\mathrm{op}}}, \qquad \mathcal{D}_T := \frac{\sum_{t<T}\eta_t\|D_t\|_{\mathrm{op}}e_t}{\sum_{t<T}\eta_t\|D_t\|_{\mathrm{op}}}. \tag{5}$$

Here, e_t measures the loss of dual alignment at iteration t, and $\mathcal{D}_T$ averages it using the realized operator-norm step lengths $\eta_t\|D_t\|_{\mathrm{op}}$ as weights.

Theorem 4.1 (Pathwise spectral-margin bound)**.** *Suppose Assumptions 4.1 and 4.2 hold. Let the Muon iterates $\{W_t\}_{t\ge0}$ follow (3) from any initialization $W_0 \in \mathbb{R}^{k\times d}$, with momentum parameter $\beta \in [0, 1)$, learning rate $\eta_t = \eta_0/\sqrt{t+1}$, $\eta_0 > 0$, and NS depths $q_t \in \mathbb{N}_0$. Then there exist constants $C_0, C_1, C_2, C_3 > 0$ and an integer $T_0 \ge 1$ such that, for every $T \ge T_0$, the following bounds hold:*

$$\begin{aligned} \mathcal{L}(W_T) &\le C_0 e^{-C_1\sqrt{T+1}}, \\ 0 \le \gamma_* - \widehat{\gamma}(W_T) &\le \frac{C_2 + C_3\sum_{t<T}\eta_t^2 + \gamma_*\sum_{t<T}\eta_t\|D_t\|_{\mathrm{op}}e_t}{\|W_0\|_{\mathrm{op}} + \sum_{t<T}\eta_t\|D_t\|_{\mathrm{op}}}. \end{aligned} \tag{6}$$

The constants C_0, C_1, C_2, C_3 and T_0 may depend on $n, k, p, R, \gamma_*, \|W_0\|_{\mathrm{op}}, \beta, \eta_0$, but not on T. They are uniform over the NS depth sequence.

Theorem 4.1 separates a vanishing training term from accumulated NS distortion, giving a sufficient criterion for recovering the exact-polar spectral margin.

Comparison with exact-polar Muon. For exact-polar Muon, Fan et al. (2025, Theorem 2) obtain the spectral-margin bound $O((d\log T + dn)/\sqrt{T})$ under $\eta_t = \eta_0/\sqrt{t+1}$. Our Theorem 4.1 yields a corresponding finite-NS bound with an additional path-weighted distortion term. When $R, \gamma_*, \|W_0\|_{\mathrm{op}}, \beta, \eta_0$ are fixed, the proof gives $C_2 = O(n(1+\log k))$ and $C_3 = O(n)$. The finite-NS update amplitudes satisfy $\|D_t\|_{\mathrm{op}} \ge p^{-1/2}$; this bound is established by Lemma 4.4 in Subsection 4.3.

The learning rate also gives $\sum_{t<T} \eta_t \geq \eta_0\sqrt{T}$ and $\sum_{t<T} \eta_t^2 \leq \eta_0^2(1+\log T)$. Combining these bounds with (6) and the path-weighted defect $\mathcal{D}_T$ defined in (5), we obtain, for $T \geq T_0$,

$$\gamma_* - \widehat{\gamma}(W_T) \leq O\left(\frac{n\sqrt{p}\,(1+\log k+\log T)}{\sqrt{T}}\right) + \gamma_*\mathcal{D}_T. \tag{7}$$

For exact-polar updates, $\mathcal{D}_T = 0$, leaving only the training contribution. For fixed n, k, and d, this contribution has the same $O((1+\log T)/\sqrt{T})$ dependence on the training horizon as the exact-polar bound of Fan et al. (2025). For finite NS, the additional term $\gamma_*\mathcal{D}_T$ measures the directional distortion accumulated along the training trajectory.

From NS depth to implicit bias. We now use the path-weighted defect $\mathcal{D}_T$ to connect NS depth with the selected spectral margin. Equation (7) shows why this quantity matters: the training term vanishes with T, while the effect of finite NS is carried by $\mathcal{D}_T$. The key point is that $\mathcal{D}_T$ reflects errors made throughout training. Early coarse updates become less important as training continues. If later NS errors keep shrinking, their overall effect can disappear; if they persist, a nonzero margin gap may remain. Thus loss convergence alone does not determine whether finite NS preserves the spectral implicit bias. We next study how the choice of NS depth leads to these different behaviors.

4.2 How NS Depth Shapes Muon's Implicit Bias

We now examine how NS depth affects Muon's spectral implicit bias, beginning with fixed depth.

Proposition 4.2 (Finite-depth spectral bias)**.** *Fix any finite depth $q \in \mathbb{N}_0$ for the NS map in* (4)*. There exists a strictly separable multiclass linear classification problem with $k = 3$ and $d = 2$, depending on q, with the following property. For every finite initialization $W_0 \in \mathbb{R}^{3\times 2}$, momentum parameter $\beta \in [0, 1)$, and learning-rate scale $\eta_0 > 0$, let the iterates follow* (3) *with NS depth $q_t \equiv q$, and $\eta_t = \eta_0/\sqrt{t+1}$. Then the loss converges to zero and the normalized spectral margin converges with a nonzero $\Theta((2/3)^q)$ gap:*

$$\frac{1}{4\sqrt{2}\,(3/2)^q} \leq \gamma_* - \lim_{t\to\infty} \widehat{\gamma}(W_t) \leq \frac{1}{2\sqrt{2}\,(3/2)^q}. \tag{8}$$

The pathwise upper bound permits a persistent distortion term, but the construction shows that finite depth can change the limiting margin. For every finite depth, Proposition 4.2 gives a separable problem with vanishing loss and a nonzero limiting spectral-margin gap.

What fixed depth changes. Theorem 4.1 shows that vanishing path-weighted distortion is sufficient to recover the optimal spectral margin. Proposition 4.2 complements this result by showing what can happen when the NS depth is kept fixed. For every finite depth q, there exists a separable problem for which the loss converges to zero but the limiting spectral margin remains $\Theta((2/3)^q)$ below the optimum. Thus, finite NS can leave a genuine solution-selection bias even after optimization has succeeded. The two-sided bound in (8) also shows that this worst-case bias decreases geometrically with q, although no fixed finite depth eliminates it in general. This motivates the next question: how should NS precision increase during training to recover the exact-polar spectral margin?

We consider two precision regimes. Horizon-calibrated precision chooses one depth before a prescribed finite run and keeps it fixed throughout that run. Adaptive precision instead allows the depth to vary along a single continuing trajectory. Thus, the former changes precision across runs, whereas the latter changes precision within a run.

Theorem 4.3 (Spectral-margin recovery)**.** *Suppose the assumptions of Theorem 4.1 hold.*

(i) Horizon-calibrated precision. *For each prescribed horizon T, choose $q_T \in \mathbb{N}_0$. Let $\{W_t^{(T)}\}_{t=0}^{T}$ be generated by* (3) *from the common initialization $W_0^{(T)} = W_0$ with $q_t^{(T)} \equiv q_T$ for $0 \leq t < T$, and denote its endpoint by $W_T^{(T)}$. Suppose the NS map with depth q_T has directional defect at most $\rho_T > 0$ for every nonzero matrix M of rank at most p:*

$$1 - \frac{\langle M, \mathrm{NS}(M)\rangle}{\|M\|_*\|\,\mathrm{NS}(M)\|_{\mathrm{op}}} \leq \rho_T.$$

Then, for all sufficiently large T,

$$0 \le \gamma_* - \widehat{\gamma}(W_T^{(T)}) \le C\frac{1+\log(T+1)}{\sqrt{T+1}} + \gamma_*\rho_T. \tag{9}$$

The constant $C > 0$ and the starting horizon are uniform over q_T and ρ_T.

(ii) Adaptive precision. *For a single continuing trajectory $\{W_t\}_{t\ge 0}$, let $q_t \in \mathbb{N}_0$ denote the depth used at iteration t. Suppose $\mathcal{D}_T \to 0$. Then $\widehat{\gamma}(W_T) \to \gamma_*$. For the NS map in* (4)*, $q_t \to \infty$ implies $\mathcal{D}_T \to 0$ and hence guarantees spectral-margin recovery. If $\mathcal{D}_T = O((1+\log(T+1))/\sqrt{T+1})$, the margin gap has the same training-time order as exact-polar Muon.*

The two recovery statements impose accuracy in different ways. Horizon calibration controls each finite run uniformly, while adaptive precision controls distortion accumulated along a single trajectory.

Depth for a prescribed horizon. As shown later in Lemma 4.4 of Subsection 4.3, the finite-NS defect admits an explicit depth bound for any target tolerance. For $p > 1$, it is enough to choose

$$q_T \ge \max\left\{0, \left\lceil \frac{\log((p-1)/(\mathrm{e}\rho_T))}{\log(3/2)} \right\rceil \right\}.$$

Taking $\rho_T \asymp (1+\log T)/\sqrt{T}$ balances the two terms in (9), recovering the exact-polar training-time order with $q_T = O(\log T)$ for fixed p. The sufficient depth grows logarithmically with the horizon but remains constant within each run. When $p = 1$, Frobenius normalization already equals the polar factor, so $q = 0$ suffices.

Depth along one trajectory. A more slowly growing depth suffices for asymptotic recovery: $q_t := \max\{1, \lceil \log_{3/2}\log(t+\mathrm{e})\rceil\} = O(\log\log(t+\mathrm{e}))$ guarantees $\widehat{\gamma}(W_T) \to \gamma_*$. This slowly growing rule guarantees convergence of the spectral margin. To match the finite-time rate, the accumulated defect must satisfy $\mathcal{D}_T = O((1+\log T)/\sqrt{T})$. Proposition B.7 in Appendix B.3 gives an $O(\log t)$ depth rule at this scale, as used in our experiments (Appendix A.1.3). These guarantees use directional accuracy weighted by momentum singular values: very small singular values can retain a large polar residual while contributing little to the margin-controlling defect.

4.3 Proof Sketches

We connect NS depth to solution selection through three lemmas. Lemma 4.4 controls update amplitude and directional defect; Lemma 4.5 transfers momentum alignment to descent of the loss potential. Together, they yield Theorem 4.1 and, with increasing precision, Theorem 4.3. Lemma 4.6 separately identifies the fixed-map geometry underlying Proposition 4.2.

Lemma 4.4 (Spectral control of finite NS)**.** *Let $M \ne 0$ be a momentum matrix of rank at most p, and let $D = \mathrm{NS}(M)$ be its update after q NS rounds. The amplitude satisfies $p^{-1/2} \le \|D\|_{\mathrm{op}} \le 1$, and the momentum alignment satisfies $\langle M, D\rangle/\|M\|_* \ge p^{-1/2}$. The defect obeys*

$$0 \le 1 - \frac{\langle M, D\rangle}{\|M\|_*\|D\|_{\mathrm{op}}} \le \min\left\{1 - \frac{1}{\sqrt{p}}, \frac{p-1}{\mathrm{e}(3/2)^q}\right\}.$$

These bounds provide the local geometric input for the rest of the analysis. The amplitude and alignment bounds drive the pathwise descent argument, while the q-dependent defect bound quantifies the departure from the polar geometry. For normalized positive singular values $s_j = \sigma_j(M)/\|M\|_{\mathrm{F}}$, we have $\sum_j s_j^2 = 1$ and $\max_j s_j \ge p^{-1/2}$. The NS map does not decrease these values. In singular-value coordinates, the defect denominator is at least one, and a largest singular direction contributes no alignment loss. The envelope $1 - \phi^{\circ q}(s) \le (1-s)^{(3/2)^q}$ bounds each remaining weighted residual by $[\mathrm{e}(3/2)^q]^{-1}$. There are at most $p-1$ such terms; see Proposition B.11 in Appendix B.4.

Lemma 4.5 (Momentum tracking and loss potential)**.** *Suppose the assumptions of Theorem 4.1 hold. Momentum tracks the current gradient with relative nuclear-norm error $\|M_t - G_t\|_*/\|G_t\|_* = O(\eta_t)$, uniformly over NS depths. To express accumulated alignment as a descent comparison, define the transformed loss $\mathcal{F}(\ell) := \log(e^{n\ell} - 1)$. For every $T \ge 1$, the constants C_2, C_3 of Theorem 4.1 give*

$$\mathcal{F}(\mathcal{L}(W_T)) \le C_2 + C_3\sum_{t<T}\eta_t^2 - \gamma_*\sum_{t<T}\eta_t\frac{\langle M_t, D_t\rangle}{\|M_t\|_*}. \tag{10}$$

Combining Lemmas 4.4 and 4.5 **completes the pathwise argument underlying Theorem 4.1**. Expanding the momentum recursion expresses $M_t - G_t$ as a geometrically weighted sum of past gradient differences plus an initialization term. The latter decays exponentially relative to $\|G_t\|_*$, while recent differences give an $O(\eta_t)$ relative error. Spectral–nuclear duality transfers alignment to the current gradient; multiplication by the step size yields the $O(\sum_{t<T} \eta_t^2)$ tracking remainder. The corresponding bounds are established in Lemma B.13 and in the proof of Theorem B.1 of Appendix B.4.

With $\eta_t = \eta_0/\sqrt{t+1}$, Lemma 4.4 gives accumulated alignment of order $\sqrt{T}$, while the tracking remainder grows at most logarithmically, yielding the stated loss decay. The comparison $\min_{i,c\neq y_i} z_{ic}(W_T) \geq -\mathcal{F}(\mathcal{L}(W_T))$ gives positive margins and $W_T \neq 0$ eventually. Retaining the initial-potential term in the descent estimate and using $\|W_T\|_{\mathrm{op}} \leq \|W_0\|_{\mathrm{op}} + \sum_{t<T} \eta_t \|D_t\|_{\mathrm{op}}$ gives (6); C_2 covers the initial norm and loss.

With the pathwise bound established, we next characterize the solution geometry selected by a fixed finite NS map.

Lemma 4.6 (Fixed-depth geometry in two columns)**.** *Fix an NS depth $q \in \mathbb{N}_0$ for the map in* (4)*. There exists a strictly convex norm N_q on $\mathbb{R}^{k\times 2}$ such that every nonzero $S \in \mathbb{R}^{k\times 2}$ satisfies*

$$\frac{\mathrm{NS}(S)}{N_q(\mathrm{NS}(S))} \in \arg\max_{N_q(V)\leq 1} \langle S, V\rangle.$$

Suppose the assumptions of Theorem 4.1 hold with $d = 2$ and fixed depth $q_t \equiv q$. Then the normalized iterates converge as $W_t/N_q(W_t) \to W_q^\star$, where $W_q^\star$ is the unique N_q-max-margin solution, the maximizer of $\min_{i,c\neq y_i} z_{ic}(W)$ over $N_q(W) \leq 1$.

The ratio of the two output singular values determines a strictly convex unit boundary with the normalized NS directions as support points. Lemma B.14 in Appendix B.5 constructs this norm and proves direction selection along the momentum trajectory. Lemma 4.6 therefore identifies the solution geometry selected at fixed depth. Applying this characterization to the construction in Theorem B.17 of Appendix B.6.2 yields the $\Theta((2/3)^q)$ spectral-margin gap in (8), **proving Proposition 4.2**.

For increasing precision, no new descent argument is required. In a horizon-calibrated run, the uniform bound $e_t \leq \rho_T$ gives $\mathcal{D}_T \leq \rho_T$, and Theorem 4.1 **yields Theorem 4.3(i)**. Along an adaptive trajectory, Lemma 4.4 gives $e_t \to 0$ whenever $q_t \to \infty$. Since the total realized path length diverges, every finite prefix receives vanishing weight and hence $\mathcal{D}_T \to 0$. Applying Theorem 4.1 **then gives Theorem 4.3(ii)**. The detailed arguments appear in Corollaries B.4 and B.5 of Appendix B.3.

5 Experiments

We test how finite NS separates loss convergence from margin recovery, shifts solution geometry, and permits recovery with increasing precision. Neural diagnostics examine depth effects beyond linear classification. Here T counts parameter updates, q counts NS rounds per update, and $\mathcal{D}_T$ is defined in Section 4.1. Protocols and endpoint summaries are in Appendices A.1.1–A.1.3 and A.2.1.

Setup. We use 24 separable multiclass datasets: four for each $(k, d) \in \{3, 5\} \times \{5, 10, 20\}$. Against exact SVD polar, we compare NS with $q = 1, 2, 4$, calibrated NS, and adaptive NS (Subsection 4.2). They share data, initialization, full-dataset gradients, $\beta = 0.9$, and $\eta_t = 0.2/\sqrt{t+1}$. Appendix A.1.2 provides certified spectral-margin references with primal–dual relative widths below 9.83×10^{-8}.

Finite NS can alter the selected spectral bias. We first test the loss–margin distinction in Theorem 4.1. All 144 linear runs reach cross-entropy below 2.57×10^{-6}, yet their normalized spectral margins remain separated (Figure 1). NS with $q = 1$ has median margin ratio 0.922, versus 0.968 for SVD polar, and ends below its paired SVD endpoint on all 24 datasets (Table 1, Appendix A.2.1). Smaller path-weighted defects are associated with smaller endpoint gaps, consistent with $\mathcal{D}_T$ in the bound. Proposition 4.2 establishes asymptotic separation on its constructed family.

Finite NS changes the solution geometry. We compare endpoints with Frobenius and spectral max-margin references to examine the geometry underlying these gaps (Figure 2; metrics in Appendix A.1.2). Shallow NS remains closer to the Frobenius reference, while increasing precision

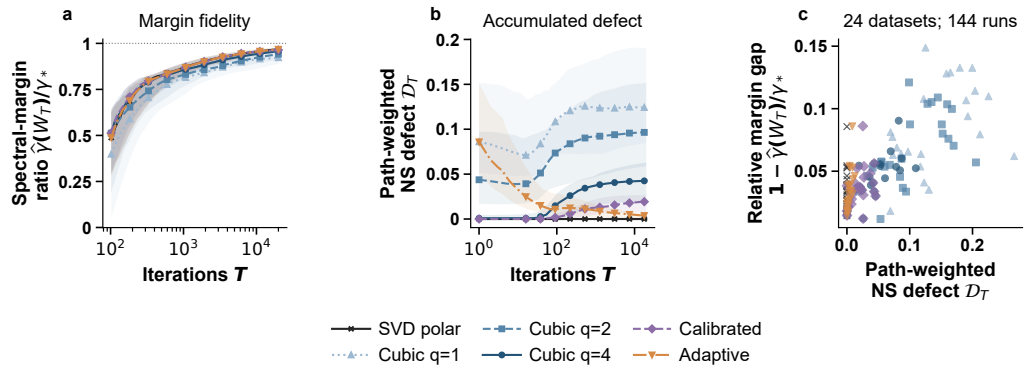


Figure 1: **Finite NS alters spectral margins at near-zero loss.** (a) Spectral-margin ratio. (b) Path-weighted NS defect from (7). (c) Relative endpoint margin gap versus that defect. Lines and bands show medians and interquartile ranges over 24 paired datasets, with certified spectral references.

shifts endpoints toward the spectral reference. Calibrated, adaptive, and SVD endpoints favor the spectral direction on 23, 24, and 24 datasets, respectively (Table 2, Appendix A.2.1). Spectral and rank-one controls appear in Appendices A.2 and A.2.2 (Figures 5 and 9c). These shifts are consistent with NS-induced geometry; Lemma 4.6 proves this mechanism for two-column classifiers.

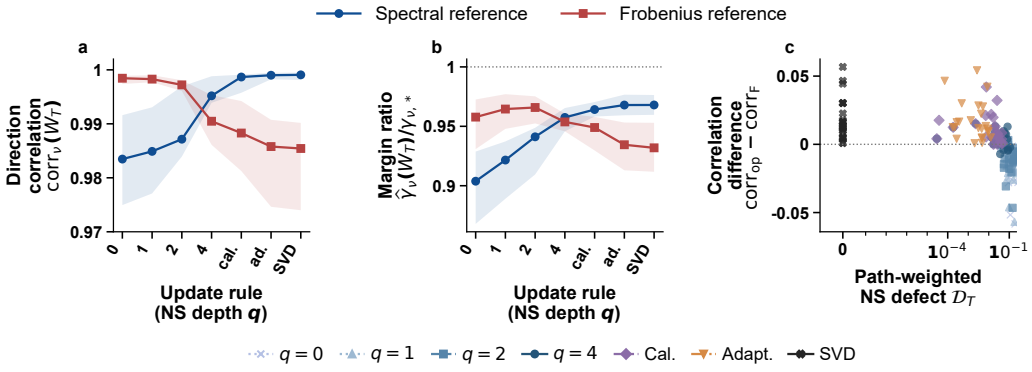


Figure 2: **Finite NS shifts solution geometry with precision.** (a) Direction correlations with Frobenius and spectral max-margin references. (b) Margin fidelity. (c) Correlation difference versus path-weighted NS defect. Metrics are defined in Appendix A.1.2. Method labels give NS depth q or the prescribed precision rule.

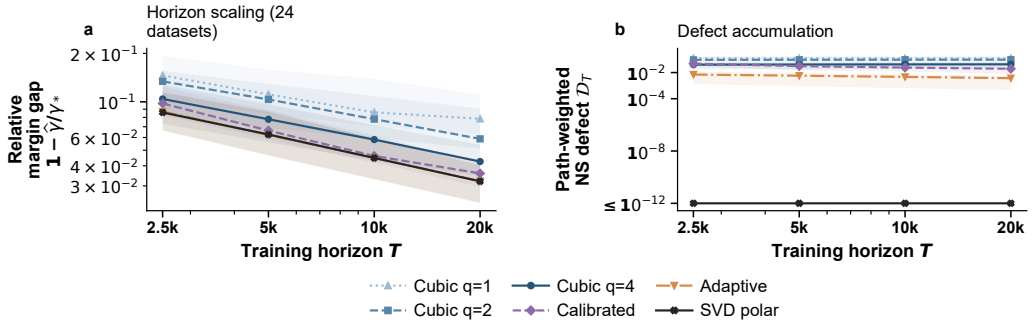


Figure 3: **Increasing precision recovers exact-polar behavior.** (a) Relative spectral-margin gap across horizons T. (b) Path-weighted NS defect, floored at 10^{-12} for display. Calibrated endpoints $W_T^{(T)}$ come from separate runs; other endpoints W_T are prefixes of one continuing trajectory.

Increasing precision preserves the exact-polar bias. We test Theorem 4.3 across $T \in \{2500, 5000, 10000, 20000\}$. Calibrated and adaptive NS track the decreasing spectral-margin gap of SVD polar, whereas shallow fixed depth retains larger defects and gaps (Figure 3). The $q_t = O(\log t)$ adaptive schedule controls $\mathcal{D}_T = O(\log(T+1)/\sqrt{T+1})$ (Proposition B.7, Appendix B.3; schedule in Appendix A.1.3). It therefore tests finite-time rate matching; the slower $O(\log \log t)$ example in Subsection 4.2 suffices for asymptotic recovery.

Neural-network diagnostics. Finally, we test depth effects in a fully connected neural network and a patch Transformer under paired seeds (Appendix A.1.1). Deeper NS generally improves optimization in both architectures, although accuracy and margin statistics need not vary monotonically with depth (Figure 4). Normalizing each nonzero update to unit operator norm retains the main depth trends (Appendix A.2, Figure 6). These results suggest singular-value reweighting, not update magnitude alone, influences training beyond linear models.

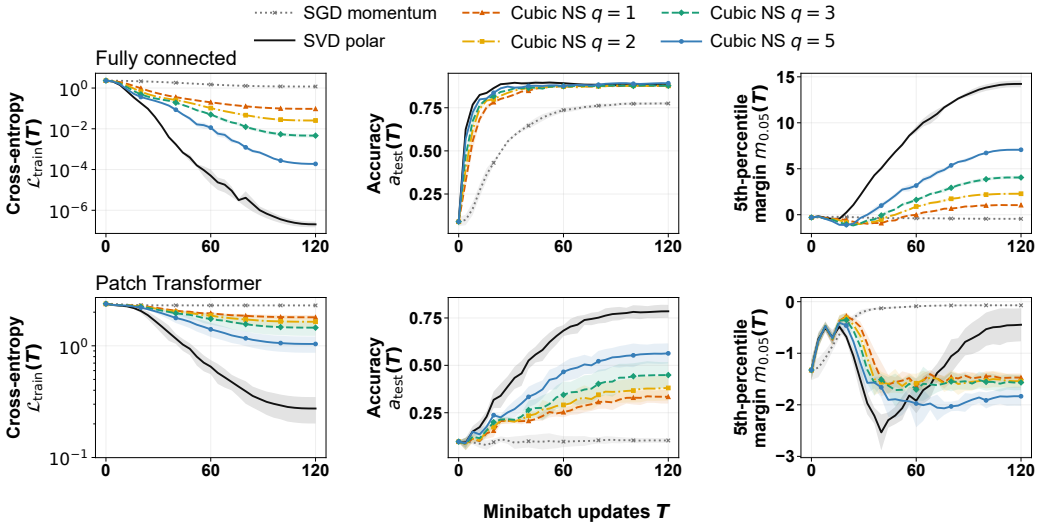


Figure 4: **Finite NS depth affects neural-network training.** Rows: fully connected neural network and patch Transformer. Columns: training cross-entropy $\mathcal{L}_{\text{train}}(T)$; held-out accuracy $a_{\text{test}}(T)$ (a fraction); and the empirical fifth percentile $m_{0.05}(T)$ of training class-score margins (correct-class score minus largest competing-class score). T counts minibatch updates, four per epoch. Curves and bands show means and one standard deviation over five paired seeds; lower loss bands are truncated at the smallest seed loss per update.

6 Discussion and Conclusion

We study how finite Newton–Schulz iterations affect Muon's implicit bias in separable multiclass classification. Our analysis decomposes the spectral-margin gap into a vanishing training term and a path-weighted directional-distortion term induced by finite NS updates. This decomposition separates successful optimization from recovery of the exact-polar solution: a fixed finite depth can leave a persistent margin gap even as the training loss vanishes, whereas increasing precision can recover the optimal spectral margin. More broadly, the effect of finite NS depends not only on the accuracy of individual updates, but also on how their directional errors accumulate along the training trajectory. Horizon-calibrated and adaptive precision control this accumulation in two different ways, across prescribed finite runs and along a single continuing trajectory, respectively.

Our experiments support the predicted distinction between loss convergence, NS depth, and spectral-margin recovery. The neural-network diagnostics further suggest that related depth effects extend beyond the linear setting. Our formal analysis focuses on deterministic separable linear classification, while extending these guarantees to nonlinear and stochastic training remains an important direction. Overall, our results suggest that finite approximations of an ideal optimizer should be assessed not only by the accuracy of each update, but also by how approximation errors accumulate and affect the solution selected over training.

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.

REPRODUCIBILITY STATEMENT

Complete statements and proofs appear in Appendices B.2–B.8, with auxiliary lemmas in Appendices C.1 and C.2. The opening of Appendix A.1 specifies pairing and summaries; Appendices A.1.1–A.1.3 specify data selection, hyperparameters, precision schedules, arithmetic settings, stopping rules, and reference certificates. The accompanying code and results include seeds, outcomes, per-step trajectories, figure source data, dependency versions, data hashes, and the shared configuration. The checks cover pairing, cumulative clocks, and measurement counts.

REFERENCES

Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The Polar Express: Optimal matrix sign methods and their application to the Muon algorithm. In *International Conference on Learning Representations*, 2026.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Pierfrancesco Beneventano, Mahmoud Abdelmoneum, and Tomaso Poggio. The spectral dynamics and noise geometry of muon, 2026. URL https://arxiv.org/abs/2606.08388v1.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology, 2024. URL https://arxiv.org/abs/2409.20325.

Åke Björck and C. Bowie. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364, 1971. doi: 10.1137/0708036.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Alex Buna, Shirley Xiaoqi Liu, and Patrick Rebeschini. Minimax optimal early-stopped gradient descent for gaussian mixture classification, 2026. URL https://arxiv.org/abs/2608.06250v1.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Proceedings of the Thirty-Third Conference on Learning Theory*, 2020.

Sayantan Choudhury, Xiaoran Cheng, Martin Takáč, Sen Na, and Mladen Kolar. Muon with nesterov momentum: Heavy-tailed noise and (randomized) inexact polar decomposition, 2026. URL https://arxiv.org/abs/2605.06884v1.

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

Lijun Ding and Madeleine Udell. A strict complementarity approach to error bound and sensitivity of solution of conic programs. *Optimization Letters*, 17(7):1551–1574, 2023. doi: 10.1007/s11590-022-01942-1.

Thang Do, Steffen Dereich, and Arnulf Jentzen. On MUON optimization: From non-convergence to an error analysis with polar express and the newton–schulz polynomial from implementations, 2026. URL https://arxiv.org/abs/2608.04607v1.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):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, pp. 39622–39669. Curran Associates, Inc., 2025. doi: 10.52202/085713-1322.

Eitan Gronich and Gal Vardi. The implicit bias of Adam and Muon on smooth homogeneous neural networks, 2026. URL https://arxiv.org/abs/2602.16340v3. arXiv version 3.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018a.

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. PMLR, 2018b.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1842–1850, 2018.

Nicholas J. Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986. doi: 10.1137/0907079.

Alan J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4):263–265, 1952.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Jack Rae, Oriol Vinyals, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030. Curran Associates, Inc., 2022. doi: 10.52202/068431-2176.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019a.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1772–1798, 2019b.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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. PMLR, 2021.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. Blog post.

Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton–Schulz. In *International Conference on Learning Representations*, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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, 2026a. URL https://arxiv.org/abs/2602.11557v2. arXiv version 2.

Mingyi Li and Taira Tsuchiya. Muon with finite Newton–Schulz: The smoothing benefit in nonsmooth nonconvex optimization, 2026. URL https://arxiv.org/abs/2608.26288v1.

Xianliang Li, Zihan Zhang, Weiyang Liu, and Han Bao. Denoise first, orthogonalize later: Understanding momentum in muon via spectral filtering, 2026b. URL https://arxiv.org/abs/2606.03899v1.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training, 2025. URL https://arxiv.org/abs/2502.16982.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

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

Gagik Magakyan, Pablo Parrilo, and Asuman Ozdaglar. Spectral scaling laws of muon, 2026. URL https://arxiv.org/abs/2606.04058v2.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory*, pp. 3–17. Springer, 2016. doi: 10.1007/978-3-319-46379-7_1.

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.

Yuji Nakatsukasa and Nicholas J. Higham. Backward stability of iterations for computing the polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, 33(2):460–479, 2012. doi: 10.1137/110857544.

Yuji Nakatsukasa, Zhaojun Bai, and François Gygi. Optimizing Halley's iteration for computing the matrix polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, 31(5): 2700–2720, 2010. doi: 10.1137/090774999.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Matan Schliserman and Tomer Koren. Tight risk bounds for gradient descent on separable data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68749–68759, 2023. doi: 10.52202/075280-3008.

Ohad Shamir. Gradient methods never overfit on separable data. *Journal of Machine Learning Research*, 22(85):1–20, 2021.

Egor Shulgin, Sultan AlRashed, Peter Richtárik, and Francesco Orabona. Beyond the ideal: Analyzing the inexact muon update. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 5167–5175. PMLR, 2026.

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.

Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(36):1007–1025, 2007.

Nikolaos Tsilivis, Gal Vardi, and Julia Kempe. Flavors of margin: Implicit bias of steepest descent in homogeneous neural networks. In *International Conference on Learning Representations*, 2025.

Gal Vardi, Ohad Shamir, and Nathan Srebro. On margin maximization in linear and ReLU networks. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10849–10858, 2021.

Yufeng Wang. The active ingredient in muon's grokking, 2026. URL `https://arxiv.org/abs/2607.20512v1`.

Jingfeng Wu, Peter Bartlett, Matus Telgarsky, and Bin Yu. Benefits of early stopping in gradient descent for overparameterized logistic regression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67081–67110. PMLR, 2025.

Xinlin Zhuang, Panyi Ouyang, Yichen Li, Jiangming Shi, Yizhang Chen, Shuman Liu, Ying Qian, Weiyang Liu, Haibo Zhang, and Imran Razzak. AMO: Adaptive muon orthogonalization, 2026. URL `https://arxiv.org/abs/2605.17806v1`.

Appendix Overview

The appendix provides the experimental protocols and additional results, the complete linear theory and its proofs, and the technical lemmas used in the analysis. The overview links to each section and subsection.

Contents	Page
A. Additional Experiments	**15**
A.1 Experimental Setup	15
A.1.1 Fully connected network and Transformer architecture diagnostic	15
A.1.2 Linear datasets and certified reference margins	16
A.1.3 Linear updates, costs, and paired comparisons	17
A.2 Additional Experimental Results	18
A.2.1 Complete endpoint summaries	18
A.2.2 Unique linear directions and the time obstruction	21
B. Linear Classification	**22**
B.1 Proof Architecture and Algorithmic Interface	22
B.2 Full Linear Theorem	23
B.3 Three Precision Regimes and Direction Selection	25
B.3.1 Centering and full direction selection	27
B.3.2 From a margin gap to a parameter-distance bound	28
B.4 Spectral-Shape Bounds and Proof of the Linear Theorem	30
B.5 Fixed-Map Geometry	34
B.6 Sharpness and Non-Necessity	35
B.6.1 A positive average defect can coexist with exact margin recovery	35
B.6.2 An inverse-slope worst-case bias for fixed Taylor maps	37
B.7 Time Lower Bound	38
B.8 Polynomial Computation Complexity	39
C. Technical Lemmas	**40**
C.1 Polynomial and Bernstein Bounds	40
C.2 Matrix-Circuit Degree Bounds	41

A Additional Experiments

Quantities used in the experimental comparisons. For a nonzero momentum matrix M_t and its finite-NS update D_t, let

$$c_t := \|D_t\|_{\mathrm{op}}, \qquad e_t := 1 - \frac{\langle M_t, D_t \rangle}{\|M_t\|_* c_t}.$$

Here, c_t is the update amplitude, and e_t is the relative loss of dual alignment after the update is normalized to unit operator norm. The accumulated quantities used in the endpoint tables are defined in Appendix A.1.3. These definitions agree with the proof notation in Appendix B.2.

Residual accuracy and solution-selection accuracy. Kim & Oh (2026) control an orthogonality residual and show that, under a uniform bound on the initial residual, the finite-NS optimization constant rapidly approaches its exact-SVD counterpart. Their normalization, $M_t / \max\{1, \|M_t\|_{\mathrm{F}}\}$, differs from the homogeneous Frobenius normalization in (4). We instead use the realized defect e_t to measure the loss of dual alignment in the normalized update direction; Lemma B.13 transfers this control to the current gradient up to the tracking error. Weighting e_t by the realized step lengths yields the pathwise quantity appearing in our margin bound. Polar-factor residual and directional defect need not coincide, and a terminal residual alone cannot capture distortion accumulated earlier in training.

Practical polynomial choices. The original Muon example defaults to five tuned quintic rounds, with a normalization stabilizer and bfloat16 arithmetic (Jordan et al., 2024). Zhuang et al. (2026) allocate NS budgets by operator type after an initial observation phase and then keep the allocation fixed. Our cubic map is initialized by $M/\|M\|_{\mathrm{F}}$; its depth schedules connect the desired spectral-margin accuracy to polynomial precision. Evaluating the realized $\mathcal{D}_T$ uses nuclear and spectral norms, whereas the prescribed depth rules can be applied without measuring it.

Computation accounting. Following the operation-level separation used in finite-NS convergence and polar-decomposition analyses (Kim & Oh, 2026; Amsel et al., 2026), we count only matrix multiplications internal to the polynomial map. Input normalization, gradient evaluation, scalar arithmetic, and diagnostics are excluded. One cubic NS round uses two matrix multiplications, while the quintic Taylor–NS round used in Appendix A.2 uses three. Exact polar is reported as SVD in the cost tables. This accounting defines the equal-polynomial-budget comparisons; N_{matmul} denotes the number of these matrix multiplications. End-to-end runtime additionally depends on matrix shape, hardware, arithmetic precision, and kernel implementation.

A.1 Experimental Setup

We use a common experimental protocol across the corresponding comparisons. The linear experiments use full-dataset deterministic gradients, while the neural diagnostics use the paired mini-batch protocol specified below. Datasets, seeds, method depths, horizons, and stopping rules are specified in the following subsections. The linear figures report medians and interquartile ranges, while the neural diagnostics report means with one-standard-deviation bands. These summaries describe dispersion across datasets or paired seeds.

A.1.1 Fully connected network and Transformer architecture diagnostic

The architecture diagnostic uses a fixed stratified split of the saved MNIST dataset (LeCun et al., 1998): the first 100 examples of each digit form the 1,000-example training set and the next 50 form the 500-example held-out set. The fully connected network is $784 \to 128 \to 10$, with a ReLU and no biases. The Transformer uses nonoverlapping 4×4 patches, a 32-dimensional embedding, fixed sinusoidal positions, one pre-normalized encoder block with four attention heads and a 64-dimensional feed-forward layer, mean pooling, and a bias-free linear head. Dropout is zero. Matrix parameters use the compared update map; the one-dimensional normalization parameters use ordinary momentum.

We compare momentum SGD, exact SVD polar, cubic NS with $q \in \{1, 2, 3, 5\}$, and its operator-norm-matched variant $\widetilde{D}_q = D_q / \|D_q\|_{\mathrm{op}}$. All methods use momentum 0.9, batch size 250, and 30 epochs

(120 minibatch updates), with paired seeds 401–405 and identical initialization and data order. The base learning rates are 0.08 for the fully connected network and 0.035 for the Transformer, followed by 5% linear warm-up and cosine decay. Figures 4 and 6 report means with one-standard-deviation bands; paired differences use 95% t intervals across the five seeds. Optimizer-update time measures only the parameter-update call and excludes forward, backward, and evaluation time.

Which theorem components are examined. The linear comparisons instantiate the three precision regimes and measure finite-time margin fidelity against certified global reference intervals. Appendix A.2.2 additionally tests full direction distances under analytical uniqueness and strict-complementarity certificates, as well as the explicit square-root-time obstruction. The neural experiments test whether depth, update normalization, loss, accuracy, and margin retain related finite-time patterns beyond the proved linear setting.

Across the linear studies, we vary NS depth, precision schedule, horizon, polynomial budget, spectrum, or polynomial family one at a time while holding paired data and initialization fixed. The neural control additionally matches update operator norm within paired fully connected network and Transformer runs.

A.1.2 Linear datasets and certified reference margins

We use 24 datasets: $k \in \{3, 5\}$, $d \in \{5, 10, 20\}$, with four independent instances per pair. A Gaussian teacher is row-centered and rescaled to unit operator norm. Gaussian inputs are labeled by its largest logit; rejection sampling retains 12 inputs per class with teacher margin at least 0.15. Thus the saved, finite dataset has an explicit separator. Instance $j \in \{0, 1, 2, 3\}$ uses seed $30000 + 1000k + 100d + j$; the initialization uses that seed plus 100000. Each W_0 is a row-centered matrix with independent $\mathcal{N}(0, 0.05^2)$ entries before centering. Methods use the same data and initialization W_0.

Index the $m = n(k-1)$ pairs (i, c) with $c \neq y_i$ by $j \in [m]$, and write $A_j = (e_{y_i} - e_c)h_i^\top$. The reference program and its dual are

$$\begin{aligned} \gamma_* &= \max_{W,\gamma} \{\gamma : \langle A_j, W\rangle \geq \gamma \text{ for all } j \in [m], \quad \|W\|_{\mathrm{op}} \leq 1\}, \\ &= \min_{\lambda \geq 0,\, \sum_{j\in[m]} \lambda_j = 1} \left\| \sum_{j\in[m]} \lambda_j A_j \right\|_*. \end{aligned} \tag{11}$$

The primal is modeled in CVXPY (Diamond & Boyd, 2016) using $\left(\begin{smallmatrix} I_k & W \\ W^\top & I_d \end{smallmatrix}\right) \succeq 0$ and a row-centering constraint. CLARABEL supplies candidate primal and dual solutions, with initial feasibility/gap tolerances 10^{-9} and at most one retry at 10^{-11}. We verify the candidate solutions using the following rational witnesses.

Primal lower bound. Round the candidate W to a rational matrix $\widetilde{W}$ on the 10^{-12} grid. Choose a rational $S > \|\widetilde{W}\|_{\mathrm{op}}$ and certify $S^2 I_k - \widetilde{W}\widetilde{W}^\top \succ 0$ by exact rational $LDL^\top$ elimination with positive pivots. Then $\underline{\gamma} = \min_j \langle A_j, \widetilde{W}/S\rangle \leq \gamma_*$ is evaluated as a rational number, treating each saved binary floating-point input as its exact dyadic rational value.

Dual upper bound. Clip and quantize candidate multipliers to nonnegative integers and divide by their exact positive sum. This yields exactly feasible rational λ and $Z = \sum_j \lambda_j A_j$. Construct a rational positive-definite H approximating $(ZZ^\top)^{1/2} + 10^{-8} I$, again checking positive rational $LDL^\top$ pivots. The variational inequality

$$\|Z\|_* \leq \tfrac{1}{2}\{\operatorname{tr} H + \operatorname{tr}(H^{-1} ZZ^\top)\} =: \overline{\gamma}$$

provides a rational upper bound. We save all integer witnesses and both rational endpoints; displayed endpoints are rounded outwards. The largest relative interval width across the 24 datasets is below 9.83×10^{-8}. Plots use $(\underline{\gamma} + \overline{\gamma})/2$ as their reference; the full intervals are stored with the certificates.

For the directional-interpolation control, we also solve

$$\gamma_F = \max_{W,\gamma} \{\gamma : \langle A_j, W\rangle \geq \gamma \text{ for all } j \in [m], \quad \|W\|_{\mathrm{F}} \leq 1\}$$
$$= \min_{\lambda \geq 0,\, \sum_{j\in[m]} \lambda_j = 1} \left\| \sum_{j\in[m]} \lambda_j A_j \right\|_{\mathrm{F}}.$$

The normalized primal iterate gives a feasible lower bound and normalized nonnegative dual multipliers give a feasible dual upper bound. These Frobenius intervals are evaluated in float64 rather than by the rational $LDL^\top$ machinery used for the spectral reference, so we describe them as numerical primal–dual checks. Their maximum observed relative gap is 1.22×10^{-9}. Correlations use the Frobenius inner product after row centering; all compared trajectories remain in the same centered gauge.

For Figure 2, let $W_{\nu,*}$ be the row-centered max-margin reference for $\nu \in \{\mathrm{F}, \mathrm{op}\}$. We define its optimal margin, the normalized margin of an endpoint, and their direction correlation by

$$\gamma_{\nu,*} := \max_{\|W\|_\nu \leq 1} \min_{i,c\neq y_i} z_{ic}(W),$$
$$\widehat{\gamma}_\nu(W) := \frac{\min_{i,c\neq y_i} z_{ic}(W)}{\|W\|_\nu},$$
$$\mathrm{corr}_\nu(W) := \frac{\langle W, W_{\nu,*}\rangle}{\|W\|_{\mathrm{F}}\|W_{\nu,*}\|_{\mathrm{F}}}.$$

The margin fidelity is $\widehat{\gamma}_\nu(W)/\gamma_{\nu,*}$. Thus $\gamma_{\mathrm{F},*} = \gamma_F$, $\gamma_{\mathrm{op},*} = \gamma_*$, and $\widehat{\gamma}_{\mathrm{op}} = \widehat{\gamma}$. The correlation difference in panel (c) is $\mathrm{corr}_{\mathrm{op}}(W) - \mathrm{corr}_{\mathrm{F}}(W)$.

A.1.3 Linear updates, costs, and paired comparisons

For compact experimental reporting, let

$$A_T := \sum_{t<T} \eta_t c_t, \qquad E_T := \sum_{t<T} \eta_t c_t e_t,$$

where c_t and e_t were defined at the beginning of this appendix. Thus A_T records the sum of the realized update amplitudes, E_T records the accumulated directional defect, and $E_T/A_T = \mathcal{D}_T$ is the path-weighted NS defect appearing in the main-text bound.

The six primary methods run $T = 20000$ updates with $M_{-1} = 0$, $\beta = 0.9$, and $\eta_t = 0.2/\sqrt{t+1}$. The methods are SVD polar, cubic $q = 1, 2, 4$, horizon-calibrated cubic, and adaptive cubic. The interpolation study adds $q = 0$, for which $D_t = M_t/\|M_t\|_{\mathrm{F}}$ and no matrix multiplication is charged. The cubic scalar map is $\phi(s) = 3s/2 - s^3/2$. The calibrated run uses a single depth

$$q_T = \max\left\{1, \left\lceil \frac{\log((p-1)/(\mathrm{e}\rho_T))}{\log(3/2)} \right\rceil \right\}, \qquad \rho_T = (1 + \log T)/\sqrt{T}.$$

This is $q_T = 4$ when $k = 3$ and $q_T = 7$ when $k = 5$ in our grid. Consequently, calibrated and fixed-$q = 4$ trajectories coincide for the $k = 3$ instances. Adaptive runs use $q_t = \max\{1, \lceil \log(t + 2)/(2\log(3/2))\rceil\}$. The effective polynomial cost is two matrix multiplications per cubic round, excluding common forward/backward work and diagnostic SVDs. SVD polar is labeled SVD in the polynomial-cost columns.

The horizon study evaluates $T \in \{2500, 5000, 10000, 20000\}$. Fixed-depth, adaptive, and exact-polar results are taken from prefixes of the same trajectories, whereas the calibrated method is restarted from the common initialization with the depth prescribed for each horizon. For the equal-compute comparison, we use six datasets, one from each (k, d) cell, with budgets of 80,000 and 160,000 polynomial matrix multiplications. Fixed depths $q \in \{1, 2, 4, 8\}$ run for $\lfloor B_{\mathrm{comp}}/(2q)\rfloor$ updates, while calibrated and adaptive runs stop at the largest horizon allowed by the same budget.

The operator-level spectrum experiment uses diagonal matrices with geometric singular values from 1 to $1/\kappa$, followed by Frobenius normalization. The tested ranks and condition numbers are

$$r \in \{2, 4, 8, 16, 32\}, \qquad \kappa \in \{1, 3, 10, 30, 100, 300, 1000\}.$$

The quintic Taylor–NS map is

$$h_2(s) = s\{1 + \tfrac{1}{2}(1 - s^2) + \tfrac{3}{8}(1 - s^2)^2\} = \tfrac{15}{8}s - \tfrac{5}{4}s^3 + \tfrac{3}{8}s^5.$$

A cubic round costs two matrix multiplications and a quintic round costs three; equal-cost comparisons use total budgets divisible by six. We report the same relative dual-alignment defect used in the linear theorem. The experiment is deterministic, so no error bars are reported.

Measurements use the pre-update state at index t. The counters A_t, E_t and cumulative cost include exactly the preceding t updates. The final state $t = T$ is evaluated without an additional parameter update. We record all $T + 1$ pre-update states. Figure 1's margin panel begins at $t = 100$ for readability.

The association analysis is descriptive: six methods on the same dataset are paired. We report pooled rank correlations, a dataset-centered check, and paired endpoint differences. The instantaneous and cumulative defect correlations are reported together.

A.2 Additional Experimental Results

The following experiments examine spectral conditioning, update normalization, and direction selection alongside the four main-text comparisons.

Spectral conditioning and polynomial efficiency. For each nonzero probe matrix M, let $\kappa := \sigma_{\max}(M)/\sigma^{+}_{\min}(M)$, where $\sigma^{+}_{\min}$ is its smallest positive singular value. For update D, define the directional defect $\varepsilon_{\mathrm{NS}} := 1 - \langle M, D\rangle/(\|M\|_*\|D\|_{\mathrm{op}})$. We use $\varepsilon_{\mathrm{cubic}}$ and $\varepsilon_{\mathrm{quintic}}$ for the two maps evaluated at the same multiplication budget. For deterministic rank-controlled spectra, unequal singular values produce a nonzero alignment defect, but the dependence on condition number need not be monotone: very small singular values also carry little weight in the dual-alignment ratio. At 12 matrix multiplications, the quintic Taylor–NS map has lower defect than cubic NS in all 30 nontrivial rank–condition cells, with median equal-cost defect ratio 0.900.

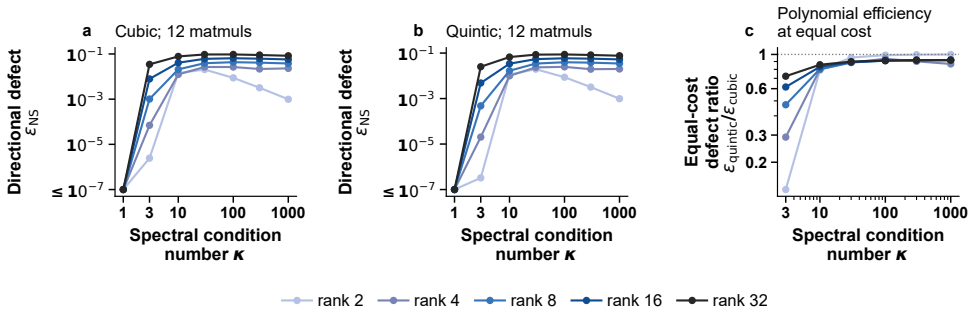


Figure 5: **Spectral conditioning and polynomial efficiency.** Alignment defect versus condition number for cubic and quintic Taylor–NS maps at equal multiplication cost, together with the quintic-to-cubic defect ratio. Values below 10^{-7} are displayed at that floor in (a,b); in particular, equal positive singular values give exactly zero defect. Ratios in (c) use the untruncated values and exclude the zero-defect case.

Operator-norm-matched neural control. Finite-depth NS changes both singular-value shape and update amplitude. We therefore repeat the fully connected network and Transformer experiments after rescaling every nonzero finite-NS update to unit operator norm. The matched runs retain the depth trends, and the raw and matched endpoints become particularly close at the larger tested depths. This comparison supports a contribution from singular-value reweighting to the observed behavior.

A.2.1 Complete endpoint summaries

The linear endpoint summaries report medians across the paired independent datasets. The neural tables specify their seed-wise means or paired differences explicitly. Interquartile ranges and individual endpoints are available in the CSV/trajectory sources.

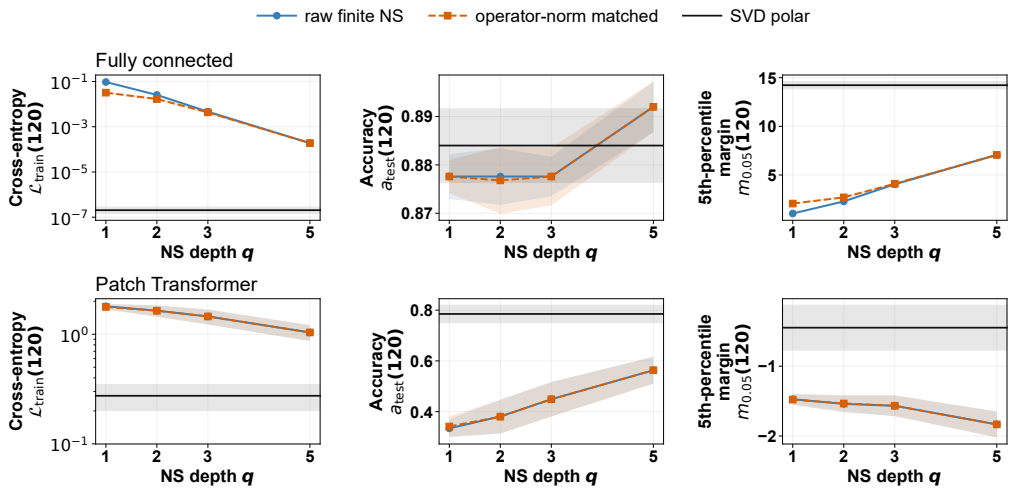


Figure 6: **Operator-norm-matched control.** Endpoints after 120 minibatch updates are compared across NS depths q, using raw updates and $\widetilde{D}_q = D_q/\|D_q\|_{\text{op}}$. Curves show seed-wise means and one-standard-deviation bands, with the same lower-loss-band truncation as Figure 4. Matching update amplitudes retains the main depth trends.

Table 1: Linear endpoints: 24 datasets per method, $T = 20000$ optimization iterations. The polynomial cost excludes SVD, gradients, and diagnostics.

Method	$\widehat{\gamma}(W_T)/\gamma_*$	E_T/A_T	CE	Polynomial matmuls
SVD polar	0.9680	2.66×10^{-16}	1.39×10^{-11}	SVD
NS $q = 1$	0.9216	0.1241	4.28×10^{-11}	40,000
NS $q = 2$	0.9412	0.0965	2.40×10^{-11}	80,000
NS $q = 4$	0.9575	0.0425	1.57×10^{-11}	160,000
Calibrated	0.9641	0.0194	1.40×10^{-11}	220,000
Adaptive	0.9679	0.0037	1.45×10^{-11}	459,438

Table 2: Directional interpolation at 20,000 updates. Correlations are medians over 24 datasets; the final column counts datasets whose endpoint is more correlated with the spectral than with the Frobenius max-margin direction.

Method	Spectral corr.	Frobenius corr.	Spectral closer
Normalization only	0.9835	0.9984	0/24
Cubic $q = 1$	0.9849	0.9983	0/24
Cubic $q = 2$	0.9872	0.9972	4/24
Cubic $q = 4$	0.9952	0.9905	18/24
Calibrated	0.9987	0.9883	23/24
Adaptive	0.9990	0.9858	24/24
SVD polar	0.9991	0.9854	24/24

Equal polynomial-computation budgets. Under matched polynomial-computation budgets, calibration balances update accuracy against the number of available training steps. Figure 7 and Table 3 report the corresponding relative spectral-margin gaps on six fixed dataset strata. The curves show medians and interquartile ranges across these six strata.

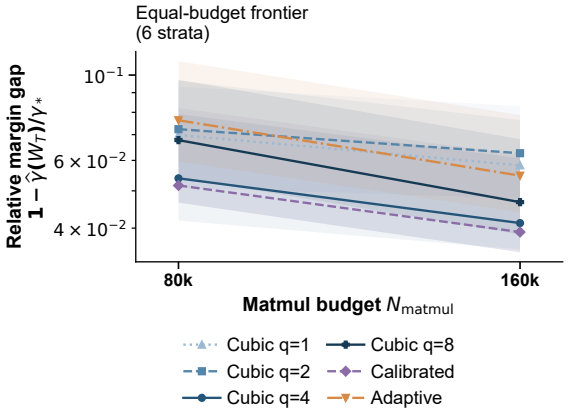


Figure 7: **Precision under equal polynomial-computation budgets.** Relative spectral-margin gaps on six fixed dataset strata at budgets of 80,000 and 160,000 matrix multiplications internal to the NS polynomial map. Gradient evaluation, input normalization, and diagnostic computations are excluded from this accounting.

Table 3: Equal-compute relative spectral-margin gaps on six fixed dataset strata. Parentheses give the median number of completed updates. Smaller is better.

Method	80k matmuls	160k matmuls
Cubic $q = 1$	0.0698 (40,000)	0.0583 (80,000)
Cubic $q = 2$	0.0723 (20,000)	0.0626 (40,000)
Cubic $q = 4$	0.0539 (10,000)	0.0412 (20,000)
Cubic $q = 8$	0.0677 (5,000)	0.0467 (10,000)
Calibrated	**0.0516** (8,333)	**0.0391** (15,714)
Adaptive	0.0763 (4,178)	0.0548 (7,787)

Table 4: Operator-norm-matched control after 120 minibatch updates (30 epochs). Each entry is the mean over five paired seeds; "matched" rescales every nonzero finite-NS matrix update to unit operator norm.

		Train CE		Test acc. (%)		5% margin	
Architecture	q	Raw	Matched	Raw	Matched	Raw	Matched
Fully connected	1	0.0952	0.0322	87.76	87.76	1.042	2.062
Fully connected	2	0.0255	0.0168	87.76	87.68	2.283	2.703
Fully connected	3	0.00462	0.00433	87.76	87.76	4.047	4.098
Fully connected	5	0.000189	0.000189	89.20	89.20	7.065	7.064
Transformer	1	1.800	1.777	33.40	34.16	−1.473	−1.479
Transformer	2	1.642	1.639	38.08	38.00	−1.540	−1.534
Transformer	3	1.454	1.454	44.88	44.88	−1.567	−1.566
Transformer	5	1.039	1.039	56.32	56.32	−1.835	−1.834

Table 5: Paired test-accuracy differences from SVD polar, in percentage points. Intervals are two-sided 95% paired t intervals over seeds 401–405.

Architecture	q	Mean difference	95% CI
Fully connected	1	−0.64	[−1.95, 0.67]
Fully connected	2	−0.64	[−1.78, 0.50]
Fully connected	3	−0.64	[−1.56, 0.28]
Fully connected	5	+0.80	[0.27, 1.33]
Transformer	1	−45.12	[−52.08, −38.16]
Transformer	2	−40.44	[−51.73, −29.15]
Transformer	3	−33.64	[−43.91, −23.37]
Transformer	5	−22.20	[−29.11, −15.29]

Absolute margin units. Figure 8 shows the absolute spectral-margin gap in the pathwise bound. It uses exactly the same 144 endpoints as Figure 1c. The main figure divides each gap by its dataset's certified reference to make differently scaled instances comparable; the absolute-unit view retains that scale variation.

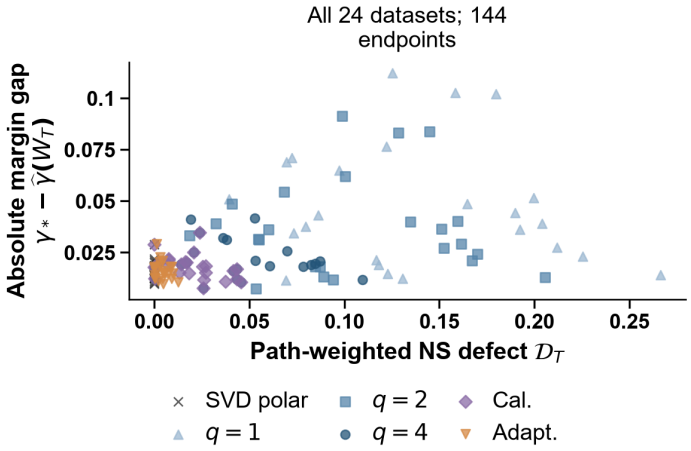


Figure 8: Absolute spectral-margin gap versus path-weighted NS defect, all 24 datasets and six methods. Each reference is the midpoint of its certified primal/dual interval. The observations are paired with those in Figure 1c.

A.2.2 Unique linear directions and the time obstruction

We additionally evaluate linear direction selection and the time obstruction under the settings below.

Direction-selection diagnostic. When the centered spectral max-margin solution is unique, convergence of the normalized margin identifies a unique normalized parameter direction. We test this implication directly on a problem with an analytically unique optimizer and an explicit margin-to-distance certificate.

An analytical reference for direction selection. For $k = d = n \in \{3, 5\}$, take $(h_i, y_i) = (e_i, i)$ and $P = I_k - \mathbf{1}\mathbf{1}^\top/k$. Then $\gamma_* = 1$ and $W_* = P$ is the unique centered spectral maximizer. Indeed, for every centered Z with $\|Z\|_{\mathrm{op}} \le 1$,

$$\begin{aligned} \|Z - P\|_{\mathrm{F}}^2 &\le 2(k-1) - 2\langle P, Z\rangle \\ &\le 2(k-1)\left[1 - \min_{i\in[k],\, c\neq i} z_{ic}(Z)\right]. \end{aligned} \tag{12}$$

The first inequality uses $\operatorname{rank} Z \le k-1$; the second uses $\langle P, Z\rangle/(k-1)$ equal to the average of all correct-class gaps. Equality of the minimum gap to one thus forces $Z = P$. This also supplies a global square-root error bound with an explicit constant.

The stronger conic hypothesis in Proposition B.10 is verifiable here. With the block matrix $S(Z) := \left(\begin{smallmatrix} I_k & Z \\ Z^\top & I_k \end{smallmatrix}\right)$, take

$$\lambda_{ic} = \frac{1}{k(k-1)}, \qquad Y = \frac{1}{2(k-1)} \begin{pmatrix} P & -P \\ -P & P \end{pmatrix}.$$

These satisfy the dual identity, $S(P)Y = 0$, and $S(P) + Y \succ 0$; the smallest positive eigenvalue of Y is $1/(k-1)$, and every active scalar multiplier is positive. Uniqueness and strict complementarity therefore follow analytically.

Five paired Gaussian initializations, seeds 301–305, use $W_0 = 0.2\,\mathcal{N}(0,1)^{k\times k}$ without row centering. For $T = 20000$, $\eta_t = 0.2/\sqrt{t+1}$ and $\beta = 0.9$, we compare SVD polar, cubic $q = 1$, calibrated

cubic, and adaptive cubic. The new calibration uses the conservative certificate $p/[\mathrm{e}(3/2)^q] \le \rho_T$, $p = k-1$ and $\rho_T = (1+\log T)/\sqrt{T}$, giving $q = 6, 8$. The original 24-dataset experiment used the sharper relative-defect bound and depths $4, 7$. The adaptive rule is unchanged.

All 40 trajectories reduce their distance to P (Figure 9a,b). At the endpoint, the median full direction distances for SVD polar and cubic $q = 1$ are 0.00789 and 0.00497 for $k = 3$, and 0.02105 and 0.01245 for $k = 5$. For every saved state with $W_t \neq 0$, we check the constant uncentered component and

$$\left\| \frac{W_t}{\|W_t\|_{\mathrm{op}}} - P \right\|_{\mathrm{F}} \le \sqrt{2(k-1)[1-\widehat{\gamma}(W_t)]} + \frac{\|(I_k - P)W_0\|_{\mathrm{F}}}{\|W_t\|_{\mathrm{op}}}.$$

Thus the experiment measures the full direction and its certified margin-to-distance envelope, beyond a comparison of optimal values. The fixed-depth convergence here reflects the geometry of this symmetric dataset; Appendix B.6 gives contrasting biased instances.

The square-root-time constant. We also run the rank-one example of Theorem B.18 for $T = 100000$, $\eta_0 = 0.2$ and $\beta \in \{0, 0.9, 0.99\}$. Every admissible NS strategy has the same direction on this example. The endpoint values of $\sqrt{T}(1-\widehat{\gamma}(W_T))$ are 0.86989, 0.86998 and 0.87012, respectively, compared with $\log 2/(4\eta_0) = 0.86643$ (Figure 9c). These values approach the predicted asymptotic constant.

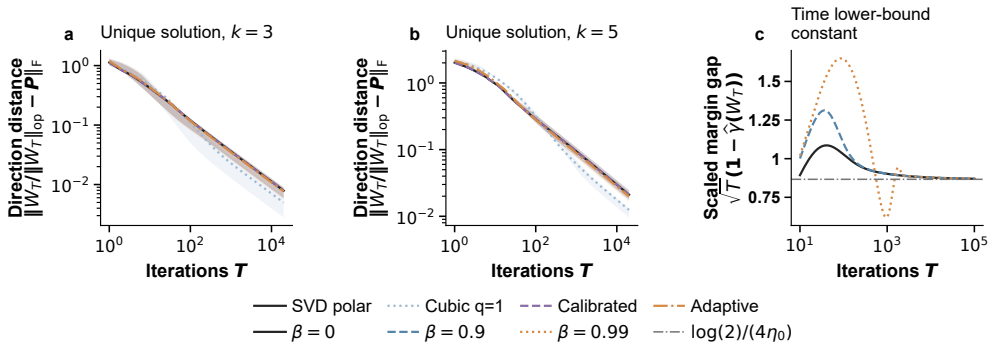


Figure 9: Direct validation of linear direction selection and the time obstruction. (a,b) Full normalized-parameter distance to the analytically unique optimizer, with five paired uncentered initializations per class count and method; lines show medians and bands show interquartile ranges. SVD, calibrated and adaptive curves nearly overlap. (c) The deterministic rank-one lower-bound example at three momentum values; the horizontal line is the proved asymptotic constant.

B Linear Classification

B.1 Proof Architecture and Algorithmic Interface

Appendix B.2 gives the general linear comparison theorem, followed by the precision regimes and the spectral-shape and momentum-tracking bounds used in its proof. We then study direction selection, fixed-map geometry, counterexamples, and computational lower bounds.

We formulate the appendix results for a general polynomial update map, with cubic NS as a special case. Let $G_t = \nabla \mathcal{L}(W_t)$, initialize $M_{-1} = 0$, and consider

$$\begin{aligned} M_t &= \beta M_{t-1} + (1-\beta) G_t, \\ D_t &= P_{\psi_t}(M_t/\|M_t\|_{\mathrm{F}}), \\ W_{t+1} &= W_t - \eta_t D_t. \end{aligned} \tag{13}$$

Here, P_{ψ_t} applies the scalar polynomial ψ_t to the singular values of its input, as defined precisely below, and the update is taken to be zero when $M_t = 0$. The linear theorem will show that this zero case does not occur at a finite iteration on separable data. The polynomial may be fixed, chosen once for a prescribed horizon, or varied along one continuing trajectory. For cubic NS, let $\psi_q := \phi^{\circ q}$ and take $\psi_t = \psi_{q_t}$.

B.2 Full Linear Theorem

Model and margin notation. Consider the multiclass linear problem in Section 3, with training data $(h_i, y_i)_{i=1}^n \subset \mathbb{R}^d \times [k]$ and parameter matrix $W \in \mathbb{R}^{k\times d}$. For each example i and competing class $c \neq y_i$, let

$$A_{i,c} := (e_{y_i} - e_c)h_i^\top.$$

Then $\langle A_{i,c}, W\rangle = z_{ic}(W)$. To simplify the repeated minimum-margin expressions in the proofs, we use the local shorthand $m(W)$ in this appendix. In this notation,

$$\mathcal{L}(W) = \frac{1}{n}\sum_{i=1}^n \log\left(1 + \sum_{c\neq y_i} e^{-\langle A_{i,c}, W\rangle}\right), \tag{14}$$
$$m(W) := \min_{i\in[n],\, c\neq y_i} \langle A_{i,c}, W\rangle,$$
$$\gamma_* = \max_{\|W\|_{\mathrm{op}}\leq 1} m(W), \qquad \widehat{\gamma}(W) = \frac{m(W)}{\|W\|_{\mathrm{op}}} \quad (W \neq 0).$$

Assumptions 4.1 and 4.2 give $\gamma_* > 0$ and the uniform bound

$$\|A_{i,c}\|_* = \sqrt{2}\|h_i\|_2 \leq B := \sqrt{2}R, \qquad p := \min\{k-1, d\}.$$

Thus B is an upper bound on the nuclear norms of the margin matrices; it need not equal their maximum. Spectral–nuclear duality also gives $\gamma_* \leq B$. The comparison theorem allows arbitrary finite initialization and varying signal rank. Strict separability and deterministic gradients of the empirical loss are used to control the loss and parameter norm. Uniqueness is required only when margin recovery is upgraded to convergence of the whole normalized direction.

Newton–Schulz polynomial maps. Classical work on polar iterations, higher-order schemes, and numerical stability provides background for finite approximations of the polar factor (Björck & Bowie, 1971; Higham, 1986; Nakatsukasa et al., 2010; Nakatsukasa & Higham, 2012). Let ψ be an odd polynomial that is strictly increasing on $[0, 1]$ and satisfies

$$s \leq \psi(s) \leq 1, \qquad 0 \leq s \leq 1.$$

We call such a polynomial admissible. For a matrix X with compact singular value decomposition $X = U_r \operatorname{diag}(\sigma_1, \ldots, \sigma_r)V_r^\top$, define

$$P_\psi(X) := U_r \operatorname{diag}(\psi(\sigma_1), \ldots, \psi(\sigma_r))V_r^\top, \qquad P_\psi(0) := 0.$$

Only Frobenius-normalized inputs are used in (13), so their singular values lie in $[0, 1]$. The cubic Newton–Schulz iteration in (4) corresponds to $\psi = \psi_q := \phi^{\circ q}$ with $\phi(s) = 3s/2 - s^3/2$; the case $q = 0$ is the identity polynomial.

The more general Taylor–Newton–Schulz family used in the certificates below consists of the scalar maps

$$f_r(s) = s\sum_{j=0}^r 4^{-j}\binom{2j}{j}(1-s^2)^j, \qquad r \geq 0,$$
$$a_r := f_r'(0) = (2r+1)4^{-r}\binom{2r}{r}. \tag{15}$$

For a composition $\psi = f_{r_q} \circ \cdots \circ f_{r_1}$, its slope at the origin is

$$\Lambda := \psi'(0) = \prod_{j=1}^q a_{r_j}.$$

An empty composition has $\psi(s) = s$ and $\Lambda = 1$. In particular, $a_1 = 3/2$, so q cubic iterations have slope $\Lambda = (3/2)^q$. The scalar envelope used in the spectral-shape bound is

$$\omega(\Lambda) := \max_{0\leq s\leq 1} s(1-s)^\Lambda = \frac{\Lambda^\Lambda}{(\Lambda+1)^{\Lambda+1}} \leq \frac{1}{\mathrm{e}\Lambda}. \tag{16}$$

Learning rates. We analyze (13) from an arbitrary finite W_0, with a fixed momentum parameter $\beta \in [0, 1)$ and learning rates

$$\eta_t = \eta_0(t+1)^{-a}, \qquad \eta_0 > 0, \qquad 0 < a \le 1. \tag{17}$$

The exponent $a = 1/2$ gives the main-text setting. Other exponents change the accumulated step sizes but leave the comparison argument unchanged.

Update amplitudes and directional defects. For a nonzero matrix S, set $D = P_\psi(S/\|S\|_{\mathrm{F}})$ and define

$$e_\psi(S) := 1 - \frac{\langle S, D\rangle}{\|S\|_*\|D\|_{\mathrm{op}}}.$$

This quantity measures the loss of dual alignment of the operator-normalized update. Along the optimization trajectory, write

$$c_t := \|D_t\|_{\mathrm{op}}, \qquad a_t := \eta_t c_t, \qquad e_t := e_{\psi_t}(M_t) = 1 - \frac{\langle M_t, D_t\rangle}{\|M_t\|_* c_t}.$$

Thus c_t is the update amplitude and a_t is the length of the parameter update in operator norm. With $V_t^{\mathrm{dir}} := D_t/c_t$, the update can be written as $W_{t+1} = W_t - a_t V_t^{\mathrm{dir}}$, where $\|V_t^{\mathrm{dir}}\|_{\mathrm{op}} = 1$.

We use the accumulated quantities

$$\begin{aligned} A_T &:= \sum_{t<T} a_t, \quad E_T := \sum_{t<T} a_t e_t, \\ H_T &:= \sum_{t<T} \eta_t, \quad S_T := \sum_{t<T} \eta_t^2, \qquad R_0 := \|W_0\|_{\mathrm{op}}. \end{aligned} \tag{18}$$

Here, A_T is the total realized path length, E_T is the accumulated directional defect, and H_T, S_T depend only on the learning-rate schedule. In particular, E_T/A_T is the path-weighted defect in Theorem 4.1.

To compare the loss with the minimum margin, define $\mathcal{F}(\ell) := \log(e^{n\ell} - 1)$ for $\ell > 0$. The following theorem bounds this transformed loss and the normalized margin using the same accumulated quantities.

Theorem B.1 (Unified linear comparison)**.** *Suppose Assumptions 4.1 and 4.2 hold, and let* $\{(W_t, M_t, D_t)\}_{t\ge 0}$ *be generated by* (13) *with learning rates* (17)*. The polynomial* ψ_t *may be any admissible polynomial at each iteration. Then the iterates are well defined, and* G_t*,* M_t*, and* D_t *are nonzero at every finite iteration.*

There exist constants $C_0, C_1, C_2, C_3 > 0$ *and an integer* $T_0 \ge 1$*, depending only on* $n, k, p, R, \gamma_*, \|W_0\|_{\mathrm{op}}, \beta, \eta_0, a$*, such that* $W_T \ne 0$ *and, for every* $T \ge T_0$*,*

$$\mathcal{F}(\mathcal{L}(W_T)) \le C_2 + C_3 S_T - \gamma_*(A_T - E_T), \tag{19}$$

$$0 \le \gamma_* - \widehat{\gamma}(W_T) \le \frac{C_2 + C_3 S_T + \gamma_* E_T}{R_0 + A_T}. \tag{20}$$

These constants and T_0 *are independent of the polynomial sequence, its degrees and depths, and any prescribed horizon used to select it. For every* $T \ge 1$*, the accumulated quantities also satisfy*

$$\frac{H_T}{\sqrt{p}} \le A_T \le H_T, \qquad A_T - E_T \ge \frac{H_T}{\sqrt{p}}. \tag{21}$$

If $0 < a < 1$*, the loss satisfies*

$$\mathcal{L}(W_T) \le C_0 \exp\bigl(-C_1(T+1)^{1-a}\bigr) \qquad (T \ge T_0). \tag{22}$$

At the endpoint $a = 1$*, the corresponding bound is*

$$\mathcal{L}(W_T) \le C_0(T+1)^{-C_1} \qquad (T \ge T_0). \tag{23}$$

Consequently, for every $a \in (0, 1]$*,* $\mathcal{L}(W_T) \to 0$*,* $m(W_T) \to \infty$*, and* $\|W_T\|_{\mathrm{op}} \to \infty$ *as* $T \to \infty$*.*

B.3 Three Precision Regimes and Direction Selection

We first evaluate the finite-time term in Theorem B.1 for the polynomial learning-rate schedule. For $T \geq 1$, define

$$\vartheta_a(T) := \begin{cases} (T+1)^{-a}, & 0 < a < \frac{1}{2}, \\ \dfrac{1+\log(T+1)}{\sqrt{T+1}}, & a = \frac{1}{2}, \\ (T+1)^{-(1-a)}, & \frac{1}{2} < a < 1, \\ \dfrac{1}{\log(T+1)}, & a = 1. \end{cases} \tag{24}$$

This function records the order of $(1+S_T)/H_T$ and is independent of the chosen NS precision.

Corollary B.2 (Polynomial learning-rate rates)**.** *Under the assumptions of Theorem B.1, there is a constant $C > 0$, independent of the polynomial sequence, such that, for all sufficiently large T,*

$$0 \leq \gamma_* - \widehat{\gamma}(W_T) \leq C\vartheta_a(T) + \gamma_* \frac{E_T}{A_T}.$$

In particular, the contribution from the learning-rate schedule converges to zero for every $a \in (0, 1]$.

Proof. By Theorem B.1,

$$0 \leq \gamma_* - \widehat{\gamma}(W_T) \leq \frac{C_2 + C_3 S_T + \gamma_* E_T}{R_0 + A_T}.$$

Since $A_T \geq H_T/\sqrt{p}$,

$$\gamma_* - \widehat{\gamma}(W_T) \leq \sqrt{p}\,\frac{C_2 + C_3 S_T}{H_T} + \gamma_* \frac{E_T}{A_T}.$$

It remains to estimate $(1+S_T)/H_T$.

If $0 < a < 1/2$, then

$$H_T = \Theta(T^{1-a}), \qquad S_T = \Theta(T^{1-2a}),$$

so $(1+S_T)/H_T = O(T^{-a})$.

If $a = 1/2$, then

$$H_T = \Theta(\sqrt{T}), \qquad S_T = \Theta(\log(T+1)),$$

and therefore

$$\frac{1+S_T}{H_T} = O\left(\frac{1+\log(T+1)}{\sqrt{T+1}}\right).$$

If $1/2 < a < 1$, then $S_T = O(1)$ and $H_T = \Theta(T^{1-a})$, which gives $(1+S_T)/H_T = O(T^{-(1-a)})$.

Finally, if $a = 1$, then $H_T = \Theta(\log(T+1))$ and $S_T = O(1)$, so $(1+S_T)/H_T = O(1/\log(T+1))$. Combining the four cases proves the claim. □

Fixed precision. For a fixed cubic depth q, let

$$\bar{e}_q := \sup_{\substack{M \neq 0 \\ \operatorname{rank}(M) \leq p}} e_{\psi_q}(M).$$

This quantity uniformly bounds the directional defect over all momentum matrices of rank at most p.

Corollary B.3 (Fixed Newton–Schulz depth)**.** *Under the assumptions of Theorem B.1, suppose that $\psi_t = \psi_q$ for a fixed nonnegative integer q. Then*

$$\liminf_{T \to \infty} \widehat{\gamma}(W_T) \geq \gamma_*(1 - \bar{e}_q),$$

where

$$\bar{e}_q \leq \min\left\{1 - \frac{1}{\sqrt{p}}, \frac{p-1}{\mathrm{e}(3/2)^q}\right\}.$$

More generally, the same lower bound holds for any fixed admissible map whose realized defects are uniformly bounded by $\bar{e}_q$.

Proof. Since $E_T \le \bar{e}_q A_T$, Theorem B.1 gives

$$\gamma_* - \widehat{\gamma}(W_T) \le \frac{C_2 + C_3 S_T}{R_0 + A_T} + \gamma_* \bar{e}_q.$$

The first term tends to zero by Corollary B.2. The cubic bound follows from Proposition B.11 and (16). □

Precision chosen for a prescribed horizon. For horizon calibration, each prescribed horizon T defines a separate run with a single depth q_T used throughout that run. The data, initialization, and optimizer parameters are held fixed across horizons.

Corollary B.4 (Horizon-calibrated Newton–Schulz depth)**.** *Assume the setting of Theorem B.1, and keep the data, W_0, β, η_0, and a fixed across runs. For each horizon T, choose a depth q_T and use $q_t^{(T)} \equiv q_T$ for $0 \le t < T$. Let $W_T^{(T)}$ denote the endpoint of this run. If the uniform rank-at-most-p directional defect of this map is at most $\rho_T > 0$, then, for all sufficiently large T,*

$$0 \le \gamma_* - \widehat{\gamma}(W_T^{(T)}) \le C\vartheta_a(T) + \gamma_* \rho_T,$$

where C and the starting horizon are independent of the chosen depths and tolerances. For cubic Newton–Schulz and $p > 1$, the defect condition holds whenever

$$q_T \ge \max\left\{0,\ \left\lceil \frac{\log((p-1)/(\mathrm{e}\rho_T))}{\log(3/2)} \right\rceil \right\}. \tag{25}$$

Proof. For the run prescribed for horizon T, the uniform defect bound gives $E_T \le \rho_T A_T$. The margin conclusion follows from Theorem B.1 and Corollary B.2, whose constants are uniform across the chosen polynomials. For cubic Newton–Schulz, Proposition B.11 gives

$$e_t \le \frac{p-1}{\mathrm{e}(3/2)^{q_T}},$$

which is at most ρ_T under (25). □

Precision along a continuing trajectory. The preceding result concerns endpoints of different runs. Along a single continuing run, the relevant condition is instead the vanishing of the accumulated defect relative to the realized path length.

Corollary B.5 (Adaptive Newton–Schulz depth)**.** *Under the assumptions of Theorem B.1, suppose that*

$$\frac{E_T}{A_T} \to 0.$$

Then $\widehat{\gamma}(W_T) \to \gamma_*$ as $T \to \infty$. For cubic Newton–Schulz, this condition is satisfied by every depth sequence $q_t \to \infty$.

Proof. The first statement follows directly from Corollary B.2, since $\vartheta_a(T) \to 0$. For cubic Newton–Schulz,

$$e_t \le \frac{p-1}{\mathrm{e}(3/2)^{q_t}} \to 0.$$

Since $A_T \to \infty$ and E_T/A_T is the a_t-weighted average of the bounded sequence e_t, weighted Cesàro convergence gives $E_T/A_T \to 0$. □

Remark B.6 (Weighting by realized update lengths). The error term in (20) uses the actual lengths a_t. Since $p^{-1/2}\eta_t \le a_t \le \eta_t$, the vanishing of the a_t-weighted and η_t-weighted averages is equivalent for bounded nonnegative errors. Their finite-time values, however, are different. A smaller defect alone does not make the entire bound numerically smaller at every iteration, because its denominator also depends on the realized path length.

For the square-root learning rate, we can make the adaptive depth requirements explicit. Pointwise growth of q_t is sufficient for recovery, but the accumulated-defect condition also permits schedules that do not increase at every iteration.

Proposition B.7 (Online schedules for the square-root learning rate)**.** *Assume the setting of Theorem B.1 with $a = 1/2$. Let each ψ_t be a Taylor–NS composition, and write $\Lambda_t = \psi_t'(0)$. If $\Lambda_t \gtrsim (t+2)^\sigma$ for some $\sigma > 0$, then the precision contribution E_T/A_T is bounded by*

$$\begin{cases} O(T^{-\sigma}), & 0 < \sigma < 1/2, \\ O(\log(T+1)/\sqrt{T+1}), & \sigma = 1/2, \\ O(T^{-1/2}), & \sigma > 1/2. \end{cases}$$

The total margin bound also includes the finite-time term $O((1+\log(T+1))/\sqrt{T+1})$.

For any fixed Taylor order $r \geq 1$, qualitative margin recovery holds with $\psi_t = f_r^{\circ q_t}$ and

$$q_t = \left\lceil \log_{a_r} \log(t+\mathrm{e}) \right\rceil = O(\log\log(t+\mathrm{e})).$$

The same recovery conclusion holds if this schedule is replaced by a fixed depth at iterations $t = 2^j$, $j = 0, 1, 2, \ldots$, and is used unchanged at all other iterations. In that case, the depth need not tend to infinity pointwise.

Proof. Proposition B.11 gives $a_t e_t \leq \eta_t e_t \lesssim (t+1)^{-1/2}(t+2)^{-\sigma}$. Summing this bound and dividing by $A_T \gtrsim \sqrt{T}$ gives the three stated orders. For the logarithmic depth schedule, $\Lambda_t = a_r^{q_t} \geq \log(t+\mathrm{e}) \to \infty$, so the uniform defect bound tends to zero. The additional contribution from the reset iterations is finite, since $e_t \leq 1$ and $\sum_{j\geq 0} \eta_{2^j} < \infty$. Dividing this finite contribution by $A_T \to \infty$ proves the last claim. □

The horizon choice in (25) cannot simply be used as an online pointwise tolerance: a defect of order $e_t \asymp \log(t)/\sqrt{t}$ gives an accumulated allowance of order $\log^2(T)/\sqrt{T}$. A horizon tolerance and an online precision rule therefore require different summations.

B.3.1 Centering and full direction selection

The margin conclusions above do not require a unique maximizer. To identify a unique parameter direction when one exists, we first remove the component that adds the same logit to every class. Let

$$P := I_k - \frac{1}{k}\mathbf{1}\mathbf{1}^\top, \qquad \mathcal{B}_c := \{Z : PZ = Z,\ \|Z\|_{\mathrm{op}} \leq 1\}, \qquad \mathcal{W}_* := \operatorname{argmax}_{Z\in\mathcal{B}_c} m(Z).$$

The matrix P is the orthogonal projection onto the centered class subspace. Since $m(PW) = m(W)$ and $\|PW\|_{\mathrm{op}} \leq \|W\|_{\mathrm{op}}$, the maximum margin over $\mathcal{B}_c$ is still γ_*.

Lemma B.8 (The uncentered component is constant)**.** *Under the assumptions of Theorem B.1, the iterates satisfy*

$$(I_k - P)W_t = (I_k - P)W_0 \qquad \text{for every } t \geq 0.$$

If $\widehat{\gamma}(W_t) \to \gamma_*$, every accumulation point of $W_t/\|W_t\|_{\mathrm{op}}$ belongs to $\mathcal{W}_*$. In particular, if $\mathcal{W}_* = \{W_*\}$, then the whole normalized sequence converges to W_*.

Proof. Each margin matrix sums to zero across its class rows, hence $PG_t = G_t$. The initialization $M_{-1} = 0$ and the momentum recursion give $PM_t = M_t$. An odd spectral polynomial preserves the column space of its input, including at rank deficiency, so $PD_t = D_t$. The update then proves the identity for the uncentered component.

Theorem B.1 gives $r_t := \|W_t\|_{\mathrm{op}} \to \infty$. Consequently, $(I_k - P)W_t/r_t \to 0$. Every convergent subsequence of the compact operator-unit sphere therefore has a centered limit Z with $\|Z\|_{\mathrm{op}} = 1$ and $m(Z) = \gamma_*$. If the centered maximizer is unique, a subsequence remaining a fixed distance from W_* would have a different accumulation point, which is a contradiction. This proves the stated margin-to-direction implication. □

Direction selection under adaptive and calibrated precision. Theorem B.1 gives margin recovery when $E_T/A_T \to 0$, so Lemma B.8 applies to the continuing run. For calibrated endpoints $W_T^{(T)}$, keep the data, W_0, β, η_0, and a fixed across horizons, as in Corollary B.4. The uniform loss bound gives $m(W_T^{(T)}) \to \infty$ and hence $r^{(T)} := \|W_T^{(T)}\|_{\mathrm{op}} \to \infty$. Indeed, $\mathcal{L}(W) \geq n^{-1}\log(1 + e^{-m(W)})$

and $m(W) \le \gamma_* \|W\|_{\mathrm{op}}$. The uniform margin bound and $\rho_T \to 0$ give $\widehat{\gamma}(W_T^{(T)}) \to \gamma_*$. Each finite run also satisfies $(I_k - P)W_T^{(T)} = (I_k - P)W_0$. Thus the same compactness argument applies to the normalized endpoint sequence, even though these endpoints belong to different trajectories. This is endpoint convergence across horizons; the continuing-run conclusion concerns one infinite trajectory. Uniqueness is the only additional geometric hypothesis. Margin recovery is sufficient for direction selection, irrespective of whether the path-weighted defect vanishes.

B.3.2 From a margin gap to a parameter-distance bound

Uniqueness gives qualitative directional convergence by compactness. A quantitative distance bound additionally requires control of the geometry near the maximizer. We first state this requirement explicitly and then give a sufficient conic certificate, in the error-bound setting of Hoffman (1952); Ding & Udell (2023).

Proposition B.9 (A local error bound transfers the linear rate)**.** *Assume the setting of Theorem B.1, and suppose that $\mathcal{W}_* = \{W_*\}$. Assume that there exist constants $K_{\mathrm{eb}} > 0$ and $\vartheta \in (0, 1]$ such that every $Z \in \mathcal{B}_c$ in a neighborhood of W_* satisfies*

$$\|Z - W_*\|_{\mathrm{F}} \le K_{\mathrm{eb}}[\gamma_* - m(Z)]^{\vartheta}.$$

If $E_T/A_T \to 0$, define $g_T := \gamma_* - \widehat{\gamma}(W_T)$ and $r_T := \|W_T\|_{\mathrm{op}}$. Then, for all sufficiently large T,

$$\left\| \frac{W_T}{r_T} - W_* \right\|_{\mathrm{F}} \le K_{\mathrm{eb}} g_T^{\vartheta} + \frac{\|(I_k - P)W_0\|_{\mathrm{F}}}{r_T}. \tag{26}$$

In particular, when $a = 1/2$, this gives

$$\left\| \frac{W_T}{r_T} - W_* \right\|_{\mathrm{F}} = O\left(\left[\frac{1 + \log(T+1)}{\sqrt{T+1}} + \frac{E_T}{A_T} \right]^{\vartheta} \right) + O(T^{-1/2}).$$

Proof. Take $Z_T = PW_T/r_T \in \mathcal{B}_c$. Lemma B.8 gives $Z_T \to W_*$ and $m(Z_T) = \widehat{\gamma}(W_T)$. Apply the assumed local bound to Z_T and add the constant uncentered component to obtain (26). When $a = 1/2$, the loss-rate bound gives $m(W_T) \gtrsim \sqrt{T}$. Since $m(W_T) \le \gamma_* r_T$ and the path-length bound gives $r_T = O(\sqrt{T})$, it follows that $r_T = \Theta(\sqrt{T})$. Substitution of the margin bound from Theorem B.1 proves the displayed rate. □

To give a concrete sufficient condition for this local error bound, index the original margin matrices by j. For a centered spectral maximizer W_*, define the block matrix and active constraint set

$$S(Z) := \begin{pmatrix} I_k & Z \\ Z^\top & I_d \end{pmatrix}, \qquad J := \{j : \langle A_j, W_* \rangle = \gamma_*\}.$$

The condition $S(Z) \succeq 0$ is equivalent to $\|Z\|_{\mathrm{op}} \le 1$. The following proposition uses strict complementarity for both this matrix constraint and the active margin constraints.

Proposition B.10 (A full strict-complementarity certificate)**.** *Suppose that W_* is the unique centered spectral maximizer. Assume that there exist weights $\lambda_j \ge 0$ with $\sum_j \lambda_j = 1$ and a nonzero symmetric matrix $Y \succeq 0$ such that, for every centered matrix Z,*

$$\mathrm{tr}(Y S(Z)) = \gamma_* - \sum_j \lambda_j \langle A_j, Z \rangle. \tag{27}$$

Assume further that

$$\begin{aligned} S(W_*)Y = 0, \qquad & S(W_*) + Y \succ 0, \\ \lambda_j > 0 \quad \textit{for } j \in J, \qquad & \lambda_j = 0 \quad \textit{for } j \notin J. \end{aligned} \tag{28}$$

Then the local error bound in Proposition B.9 holds with exponent $\vartheta = 1/2$.

Proof. The spectral unit ball is exactly $S(Z) \succeq 0$ by the Schur complement. Write $S_* = S(W_*)$. Complementarity gives $\mathrm{range}\, Y \subseteq \ker S_*$; since $S_* + Y \succ 0$, Y is positive definite on $\ker S_*$, so equality holds. Let Q be the orthogonal projector onto this space and put $R = I - Q$. Then

$QS_* = S_*Q = 0$, and S_* is positive definite on $\operatorname{range} R = \ker Y$. Let y_+ be the smallest positive eigenvalue of Y and $\lambda_- := \min_{j\in J} \lambda_j > 0$. The set J is nonempty because the minimum defining $m(W_*) = \gamma_*$ is attained.

The gap controls the exposed face and the active margins. For $Z \in \mathcal{B}_c$, write $g = \gamma_* - m(Z) \geq 0$. The dual identity gives

$$g = \operatorname{tr}(YS(Z)) + \sum_j \lambda_j(\langle A_j, Z\rangle - m(Z)). \tag{29}$$

Both terms are nonnegative. Since $0 \preceq S(Z) \preceq 2I$ and $Y \succeq y_+Q$,

$$\|QS(Z)\|_{\mathrm{F}}^2 = \operatorname{tr}(QS(Z)^2) \leq 2\operatorname{tr}(QS(Z)) \leq 2g/y_+.$$

Here, $S(Z)^2 \preceq 2S(Z)$ and the trace of a product of two positive semidefinite matrices is nonnegative; commutation with Q is not needed. For $j \in J$, (29) gives $0 \leq \langle A_j, Z\rangle - m(Z) \leq g/\lambda_j$, hence

$$-g \leq \langle A_j, Z - W_*\rangle \leq (\lambda_j^{-1} - 1)g.$$

In particular, its absolute value is at most $(1 + \lambda_-^{-1})g$.

Uniqueness makes the joint linear map injective. For a centered matrix Δ, set

$$S_0(\Delta) := \begin{pmatrix} 0 & \Delta \\ \Delta^\top & 0 \end{pmatrix}.$$

The linear map

$$\Delta \mapsto \big(QS_0(\Delta), (\langle A_j, \Delta\rangle)_{j\in J}\big)$$

is injective on centered matrices. To prove this, suppose that a nonzero centered Δ belongs to its kernel and write $H = S_0(\Delta)$. The full matrix equation $QH = 0$ and symmetry of H, Q give $HQ = 0$, so $H = RHR$. Let $\sigma > 0$ be the smallest eigenvalue of the restriction of S_* to $\operatorname{range} R$; thus $S_* \succeq \sigma R$. For $j \notin J$, set $b_j = \langle A_j, W_*\rangle - \gamma_* > 0$. Choose a single $\epsilon > 0$ such that

$$\epsilon\|H\|_{\mathrm{op}} \leq \sigma/2, \qquad \epsilon|\langle A_j, \Delta\rangle| \leq b_j/2 \quad (j \notin J).$$

Such a choice exists because there are finitely many constraints; zero coefficients impose no restriction. Since $H = RHR$,

$$S(W_* \pm \epsilon\Delta) = S_* \pm \epsilon H \succeq (\sigma - \epsilon\|H\|_{\mathrm{op}})R \succeq 0.$$

Both perturbations are centered. Every active margin remains equal to γ_*, and every inactive margin stays strictly larger. Thus both are feasible and have margin exactly γ_*, contradicting uniqueness of W_*.

Injectivity yields the distance estimate. On the compact unit sphere of the centered matrix space, define

$$\kappa_0 := \min_{\substack{P\Delta=\Delta \\ \|\Delta\|_{\mathrm{F}}=1}} (\|QS_0(\Delta)\|_{\mathrm{F}} + \|(\langle A_j, \Delta\rangle)_{j\in J}\|_2) > 0.$$

Continuity and injectivity give the strict positivity. Homogeneity then supplies $K_0 = \kappa_0^{-1}$ with

$$\|\Delta\|_{\mathrm{F}} \leq K_0 \left(\|QS_0(\Delta)\|_{\mathrm{F}} + \|(\langle A_j, \Delta\rangle)_{j\in J}\|_2\right).$$

Since $QS(W_*) = 0$, the preceding estimates give

$$\|Z - W_*\|_{\mathrm{F}} \leq K_0 \left(\sqrt{2g/y_+} + \sqrt{|J|}(1 + \lambda_-^{-1})g\right). \tag{30}$$

For any fixed small $\bar{g} > 0$ and $0 \leq g \leq \bar{g}$, take

$$K_{\mathrm{eb}} := K_0 \left[\sqrt{2/y_+} + \sqrt{|J|}(1 + \lambda_-^{-1})\sqrt{\bar{g}}\right].$$

This proves the asserted local square-root error bound. □

The quantitative bound therefore depends on the exposed-face geometry, the active constraints, and the injectivity constant. These additional conditions are needed only for the parameter-distance bound and not for the main spectral-margin results.

B.4 Spectral-Shape Bounds and Proof of the Linear Theorem

The linear comparison theorem rests on a spectral-shape bound and a momentum-tracking bound. We begin with the spectral effect of a finite polynomial update; the bound does not require a positive lower bound on the smallest nonzero singular value.

Proposition B.11 (A spectral-shape certificate)**.** *Let ψ be an admissible polynomial and $S \neq 0$ a matrix of rank r. Write $s_i := \sigma_i(S)/\|S\|_{\mathrm{F}}$ for its normalized positive singular values, and define*

$$\chi := \sum_{i=1}^{r} s_i, \qquad c := \max_{1\le i\le r} \psi(s_i), \qquad \omega_\psi := \max_{0\le s\le 1} s(1-\psi(s)).$$

Then $c\chi \ge 1$, $c \ge r^{-1/2}$, and the directional defect satisfies

$$0 \le e_\psi(S) \le \min\left\{1 - \frac{1}{c\chi}, \frac{(r-1)\omega_\psi}{c\chi}\right\}. \tag{31}$$

Moreover, $e_\psi(S) = 0$ if and only if all the positive singular values of S are equal. If ψ is a Taylor–NS composition with $\Lambda = \psi'(0)$, then $\omega_\psi \le \omega(\Lambda)$, where ω is defined in (16)*.*

Proof. Since $\sum_i s_i^2 = 1$, $c \ge \max_i s_i \ge r^{-1/2}$ and $c\chi \ge (\max_i s_i)\sum_i s_i \ge 1$. Also,

$$e_\psi(S) = \frac{\sum_i s_i(c - \psi(s_i))}{c\chi}.$$

The numerator is nonnegative. The inequality $\psi(s_i) \ge s_i$ gives the first upper bound. A maximizing singular value contributes zero, and each other summand is at most $s_i(1-\psi(s_i)) \le \omega_\psi$, proving the second bound. Strict monotonicity gives the equality characterization.

For a Taylor stage of order $\ell = 0$, the scalar envelope is an equality. For $\ell \ge 1$, use $f_\ell'(s) = a_\ell(1-s^2)^\ell$ and compare f_ℓ with $g_b(s) = 1-(1-s)^b$, where $b = a_\ell$. The ratio of derivatives is $(1-s)^{\ell-b+1}(1+s)^\ell$; its logarithmic derivative changes sign once because its numerator is $b - 1 - (2\ell - b + 1)s$. The ratio begins at one, first increases, then decreases to zero. Since both functions have the same endpoint values, $f_\ell \ge g_{a_\ell}$. Composing these increasing envelopes and using $g_b \circ g_{b'} = g_{bb'}$ gives $1-\psi(s) \le (1-s)^\Lambda$. Maximizing $s(1-s)^\Lambda$ proves the claim.

For example, four equal normalized singular values $s_i = 1/2$ give $c = 11/16$ after one cubic stage. The raw deficit is $5/16$, whereas $e_\psi(S) = 0$: the update is already a scalar multiple of the polar direction. This example distinguishes a support-direction certificate from an absolute polar-factor approximation. □

Cross-entropy quantities for the descent comparison. For each incorrect class $c \neq y_i$, let

$$p_{i,c}(W) := \frac{e^{-\langle A_{i,c}, W\rangle}}{1 + \sum_{b\neq y_i} e^{-\langle A_{i,b}, W\rangle}}, \qquad \Pi(W) := \frac{1}{n}\sum_{i,c\neq y_i} p_{i,c}(W).$$

Thus $p_{i,c}(W)$ is the model probability of class c on example i, and $\Pi(W)$ is the average total probability assigned to incorrect classes. The gradient has the expansion

$$\nabla\mathcal{L}(W) = -\frac{1}{n}\sum_{i,c\neq y_i} p_{i,c}(W) A_{i,c}.$$

The next proposition converts an alignment guarantee for the current gradient into a loss and margin bound. It allows a finite prefix of updates whose alignment is unfavorable.

Proposition B.12 (Signal transfer and comparison allowing adverse steps)**.** *Consider updates $W_{t+1} = W_t - a_t V_t$ in the separable linear setting, where $0 \le a_t \le \bar{a}$ and $\|V_t\|_{\mathrm{op}} \le 1$. Let $G_t = \nabla\mathcal{L}(W_t)$ and define the current-gradient defect by*

$$\delta_t := 1 - \frac{\langle G_t, V_t\rangle}{\|G_t\|_*}.$$

Then $\delta_t \in [0,2]$. For this sequence, let $A_T := \sum_{t<T} a_t$, $Q_T := \sum_{t<T} a_t^2$, and $R_0 := \|W_0\|_{\mathrm{op}}$. Define

$$\Psi(u) := \gamma_* u + (nB - \gamma_*)(u-1)_+, \qquad C_B := B^2 e^{B\bar{a}}/2,$$

where $(u-1)_+ := \max\{u-1, 0\}$. For any choice of upper bounds $d_t \in [\delta_t, 2]$, the transformed loss satisfies, for every $T \geq 0$,

$$\mathcal{F}(\mathcal{L}(W_T)) \leq \mathcal{F}(\mathcal{L}(W_0)) - \gamma_* A_T + \sum_{t<T} a_t \Psi(d_t) + nC_B Q_T. \tag{32}$$

Whenever $m(W_T) \geq 0$ and $W_T \neq 0$, the spectral-margin gap is at most

$$\frac{\gamma_* R_0 + \mathcal{F}(\mathcal{L}(W_0)) + \sum_{t<T} a_t \Psi(d_t) + nC_B Q_T}{R_0 + A_T}. \tag{33}$$

For the NS update, take $V_t = D_t/c_t$ and define the relative momentum-tracking error by

$$\widetilde{\xi}_t := \frac{\|M_t - G_t\|_*}{\|G_t\|_*}. \tag{34}$$

Whenever $M_t \neq 0$, a valid choice of defect upper bound is

$$d_t := \min\{2,\ e_t + (2 - e_t)\widetilde{\xi}_t\}. \tag{35}$$

In particular, if $A_T \to \infty$, $Q_T = o(A_T)$, and $\sum_{t<T} a_t \Psi(d_t) = o(A_T)$, then the same comparison directly yields both loss convergence and $\widehat{\gamma}(W_T) \to \gamma_*$.

Proof. The positive gradient expansion gives

$$\gamma_* \Pi(W) \leq \|\nabla \mathcal{L}(W)\|_* \leq B\Pi(W).$$

In particular, $\nabla \mathcal{L}(W) \neq 0$ at every finite W, as follows by pairing its negative with a separator. Along a unit-operator-norm direction, the Hessian quadratic form is an average variance and is at most $B^2\Pi(W)$. The total incorrect-class probability obeys $\Pi(W - sV) \leq e^{Bs}\Pi(W)$. Taylor's integral remainder therefore gives

$$\mathcal{L}(W - aV) - \mathcal{L}(W) \leq -a\langle \nabla \mathcal{L}(W), V\rangle + C_B a^2 \Pi(W).$$

The function $\mathcal{F}$ is increasing and concave, and $\mathcal{F}'(\ell) = 1/\zeta(\ell)$, where $\zeta(\ell) := (1 - e^{-n\ell})/n$. Writing $q_i = \sum_{c \neq y_i} p_{i,c}(W)$, the identity $e^{-n\mathcal{L}(W)} = \prod_i (1 - q_i)$ implies $1 \leq \Pi(W)/\zeta(\mathcal{L}(W)) \leq n$. If $\delta \leq 1$, use the lower gradient and ratio bounds on the negative linear term. If $\delta > 1$, use their upper bounds on the positive term. Both cases yield

$$\mathcal{F}(\mathcal{L}(W - aV)) - \mathcal{F}(\mathcal{L}(W)) \leq -\gamma_* a + a\Psi(\delta) + nC_B a^2.$$

Since Ψ is increasing, summation proves (32). One group contains a minimum-margin atom, hence $n\mathcal{L}(W) \geq \log(1 + e^{-m(W)})$, or $m(W) \geq -\mathcal{F}(\mathcal{L}(W))$. Using $\|W_T\|_{\mathrm{op}} \leq R_0 + A_T$ and the nonnegative margin proves (33). Its initial constant is nonnegative because $-\mathcal{F}(\mathcal{L}(W_0)) \leq m(W_0) \leq \gamma_* R_0$.

For signal transfer,

$$\begin{aligned}\langle G_t, V_t\rangle &\geq (1 - e_t)\|M_t\|_* - \|M_t - G_t\|_* \\ &\geq [1 - e_t - (2 - e_t)\widetilde{\xi}_t]\|G_t\|_*.\end{aligned}$$

Together with $\delta_t \leq 2$, this proves (35). Finally,

$$\mathcal{L}(W_T) = \frac{1}{n}\log\left(1 + e^{\mathcal{F}(\mathcal{L}(W_T))}\right) \leq \frac{1}{n} e^{\mathcal{F}(\mathcal{L}(W_T))}.$$

The stated cumulative conditions force $\mathcal{F}(\mathcal{L}(W_T)) \to -\infty$. The loss and margin conclusions follow from the same inequalities. Zero-length steps may be omitted. □

The signal-transfer bound reduces the momentum analysis to control of $\widetilde{\xi}_t$. The following lemma gives the required estimate uniformly over all update directions with operator norm at most one, and hence uniformly over the NS depths.

Lemma B.13 (Momentum tracking under polynomial learning rates)**.** *Consider the momentum recursion in* (13)*, with $M_{-1} = 0$, $\beta \in [0, 1)$, and learning rates $\eta_t = \eta_0(t+1)^{-a}$ for $a \in (0, 1]$. Every trajectory satisfying $\|D_t\|_{\mathrm{op}} \le 1$ obeys*

$$\widetilde{\xi}_t = \frac{\|M_t - G_t\|_*}{\|G_t\|_*} \le \frac{B}{\gamma_*}\mathcal{R}_t = O(\eta_t), \tag{36}$$

where

$$\mathcal{R}_t := \beta^{t+1} + (1-\beta)\sum_{j=0}^{t}\beta^j\left[\exp\left(2B\sum_{r=t-j}^{t-1}\eta_r\right) - 1\right].$$

In particular, define

$$K_{\mathrm{tr}} := 1 + \sup_{t\ge 0}\frac{\mathcal{R}_t}{\eta_t}, \qquad C_\xi := \frac{B}{\gamma_*}K_{\mathrm{tr}}. \tag{37}$$

Then $K_{\mathrm{tr}} < \infty$ and $\widetilde{\xi}_t \le C_\xi\eta_t$ for every $t \ge 0$. These constants depend only on $B, \gamma_*, \beta, \eta_0, a$, with no additional dependence on n, k, d, p or the NS depths. Moreover, $M_t \ne 0$ at every finite iteration.

Proof. The nuclear norm of $\nabla \log p_{i,c}(W)$ is at most $2B$. Therefore, for $s \le t$,

$$\|G_s - G_t\|_* \le B\Pi(W_t)\left[\exp\left(2B\sum_{r=s}^{t-1}\eta_r\right) - 1\right],$$

while strict separability gives $\|G_t\|_* \ge \gamma_*\Pi(W_t)$.

Expanding the momentum recursion yields

$$M_t - G_t = (1-\beta)\sum_{j=0}^{t}\beta^j(G_{t-j} - G_t) - \beta^{t+1}G_t.$$

Dividing by $\|G_t\|_*$ gives the bound through $\mathcal{R}_t$.

For $\beta = 0$ the result is immediate, so suppose $\beta \in (0, 1)$. Split the sum at $j = t/2$. For $j \le t/2$, monotonicity of η_t gives

$$\sum_{r=t-j}^{t-1}\eta_r \le j\eta_{t-j} \le 2^a j\eta_t.$$

Let $K = 2^{a+1}B$. Then $2B\sum_{r=t-j}^{t-1}\eta_r \le Kj\eta_t$. Choose $\rho \in (\beta, 1)$. Since $\eta_t \to 0$, for all sufficiently large t, $\beta e^{K\eta_t} \le \rho$. Using $e^x - 1 \le xe^x$,

$$\begin{aligned}\sum_{j\le t/2}\beta^j(e^{Kj\eta_t} - 1) &\le K\eta_t\sum_{j\le t/2} j\beta^j e^{Kj\eta_t}\\ &\le K\eta_t\sum_{j=0}^{\infty} j\rho^j = O(\eta_t).\end{aligned}$$

For $j > t/2$,

$$\sum_{j>t/2}\beta^j\left[\exp\left(2B\sum_{r=t-j}^{t-1}\eta_r\right) - 1\right] \le \frac{\beta^{t/2}}{1-\beta}\exp\left(2B\sum_{r<t}\eta_r\right).$$

If $0 < a < 1$, then $\sum_{r<t}\eta_r = O(t^{1-a}) = o(t)$, and hence

$$\beta^{t/2}\exp(O(t^{1-a})) = o(t^{-a}) = o(\eta_t).$$

If $a = 1$, then $\sum_{r<t}\eta_r = O(\log t)$, so

$$\beta^{t/2}t^{O(1)} = o(t^{-1}) = o(\eta_t).$$

Also $\beta^{t+1} = o(\eta_t)$. Therefore $\mathcal{R}_t = O(\eta_t)$. Since each η_t is positive and each $\mathcal{R}_t$ is finite, the finite prefix and this tail bound give $K_{\mathrm{tr}} < \infty$. The envelope involves only B, β, η_0, a; the factor B/γ_* yields the stated dependence of C_ξ. Finally, pairing the positive momentum weights with any fixed separator shows that $M_t \ne 0$ at every finite iteration. □

The proof combines the spectral-shape bound, momentum tracking, and the descent comparison. We first establish loss decay and eventual positive margins, and then derive the normalized margin bound.

Proof of Theorem B.1. The positive gradient expansion and strict separability give $G_t \neq 0$. The momentum is a positive weighted sum of these gradients, and pairing its negative with a separator gives $M_t \neq 0$. An admissible polynomial has $\psi_t(s) \geq s > 0$ on $(0, 1]$, so its update is also nonzero. Thus the recursion is well defined at every finite iteration.

The gradients and momentum sum to zero across their class rows, so their ranks are at most p, regardless of W_0. Proposition B.11 gives $p^{-1/2} \leq c_t \leq 1$, $e_t \leq 1 - p^{-1/2}$, and

$$c_t(1 - e_t) = \frac{\sum_i s_i \psi_t(s_i)}{\sum_i s_i} \geq \frac{1}{\sum_i s_i} \geq p^{-1/2}.$$

These inequalities prove (21).

Use the all-iteration constant C_ξ from (37). The transfer bound (35) uses $d_t = \min\{2, e_t + (2 - e_t)\widetilde{\xi}_t\}$. Because $0 \leq e_t \leq 1$ and $\gamma_* \leq B$, we have

$$d_t \leq e_t + 2\widetilde{\xi}_t, \qquad (d_t - 1)_+ \leq 2\widetilde{\xi}_t, \qquad \Psi(d_t) \leq \gamma_* e_t + 2nB\widetilde{\xi}_t.$$

Thus, at every iteration, without a rank-dependent prefix,

$$a_t \Psi(d_t) \leq \gamma_* a_t e_t + 2nBC_\xi \eta_t^2. \tag{38}$$

We now choose the same positive constants for the loss and margin comparisons. Write $R_0 = \|W_0\|_{\mathrm{op}}$ and $L_0 = \mathcal{L}(W_0)$. Since each initial margin is at least $-BR_0$,

$$L_0 \leq \log\bigl(1 + (k - 1)e^{BR_0}\bigr) \leq \log k + BR_0, \qquad \mathcal{F}(L_0) \leq nL_0.$$

Take

$$\begin{aligned} C_2 &:= 1 + n \log k + (nB + \gamma_*)R_0, \\ C_3 &:= \frac{nB^2 e^{B\eta_0}}{2} + 2nBC_\xi. \end{aligned} \tag{39}$$

In particular, $C_2 \geq \gamma_* R_0 + \mathcal{F}(L_0) + 1 > \mathcal{F}(L_0)$. The two terms in C_3 bound the loss-curvature remainder and the momentum-tracking penalty. Taking $\bar{a} = \eta_0$ and $Q_T \leq S_T$ in Proposition B.12, together with (38), proves (19) for every $T \geq 1$.

The path bound $A_T - E_T \geq H_T/\sqrt{p}$ consequently yields

$$\mathcal{L}(W_T) \leq \frac{1}{n} \exp\left(C_2 + C_3 S_T - \frac{\gamma_*}{\sqrt{p}} H_T \right). \tag{40}$$

For every $a \in (0, 1]$, $H_T \to \infty$ and $S_T = o(H_T)$. Indeed, $H_T = \Theta(T^{1-a})$ for $a < 1$, while S_T is $\Theta(T^{1-2a})$ for $a < 1/2$, $\Theta(\log(T + 1))$ for $a = 1/2$, and bounded for $a > 1/2$. At $a = 1$, $H_T = \Theta(\log(T + 1))$. We may therefore choose a starting time, independently of the polynomial sequence, after which $C_3 S_T \leq \gamma_* H_T/(2\sqrt{p})$. For $0 < a < 1$, the integral bound gives $H_T \geq \eta_0((T + 1)^{1-a} - 1)/(1 - a)$, and hence $H_T \geq \eta_0(T + 1)^{1-a}/(2(1 - a))$ eventually. For $a = 1$, $H_T \geq \eta_0 \log(T + 1)$. Thus valid loss constants are

$$C_0 := \frac{e^{C_2}}{n}, \qquad C_1 := \begin{cases} \dfrac{\gamma_* \eta_0}{4(1 - a)\sqrt{p}}, & 0 < a < 1, \\ \dfrac{\gamma_* \eta_0}{2\sqrt{p}}, & a = 1. \end{cases} \tag{41}$$

These choices prove (22) and (23). In particular, when $a = 1/2$, $C_1 = \gamma_* \eta_0/(2\sqrt{p})$, as in the main theorem.

The same potential bound gives $m(W_T) \geq -\mathcal{F}(\mathcal{L}(W_T)) \to \infty$. Since $m(W_T) \leq \gamma_* \|W_T\|_{\mathrm{op}}$, the parameter norm diverges. Enlarging T_0 if necessary ensures $m(W_T) > 0$ and $W_T \neq 0$ uniformly over all polynomial sequences. Only now do we apply (33), using (38) and the bound on $\gamma_* R_0 + \mathcal{F}(L_0)$ already included in C_2. This proves (20) with the same C_2, C_3.

For fixed $R, \gamma_*, R_0, \beta, \eta_0, a$, the choices above give $C_2 = O(n(1 + \log k))$ and $C_3 = O(n)$. Neither comparison constant has additional dependence on p or d. The explicit rank factor in the margin rate comes from $A_T \geq H_T/\sqrt{p}$; the loss exponent and the starting time can also depend on p. □

B.5 FIXED-MAP GEOMETRY

A fixed finite map can be described through the solution geometry it induces. For two-column matrices, the normalized polynomial updates are support points of a strictly convex matrix norm, yielding the associated maximum-margin selection rule.

Lemma B.14 (Fixed-map selection with two columns)**.** *Let ψ be a fixed admissible polynomial. There exists a strictly convex matrix norm N_ψ on $\mathbb{R}^{k\times 2}$ such that, for every nonzero matrix $S \in \mathbb{R}^{k\times 2}$,*

$$\frac{P_\psi(S/\|S\|_{\mathrm{F}})}{N_\psi(P_\psi(S/\|S\|_{\mathrm{F}}))} \in \operatorname{argmax}_{N_\psi(V)\le 1}\langle S, V\rangle.$$

Consider any strictly separable linear classification problem with $d = 2$. Let the iterates be generated by (13) *using this fixed map, an arbitrary finite W_0, $M_{-1} = 0$, $\beta \in [0, 1)$, and the learning rates* (17)*. Then $W_t/N_\psi(W_t)$ converges to the unique maximizer of $\max_{N_\psi(W)\le 1} m(W)$. Consequently, the operator-normalized iterates also converge to the operator-normalized version of that maximizer.*

Proof. For $0 \le \lambda \le 1$, define

$$R_\psi(\lambda) := \frac{\psi(\lambda/\sqrt{1+\lambda^2})}{\psi(1/\sqrt{1+\lambda^2})}, \qquad \rho_\psi := R_\psi^{-1},$$

and let

$$x_\psi(s) := \exp\left(-\int_0^s \frac{\rho_\psi(u)}{1+u\rho_\psi(u)}\,du\right).$$

The ratio R_ψ is continuous and strictly increasing from zero to one. The curve $x_\psi(s)(1, s)$, its coordinate permutations, and its sign reflections form a symmetric convex unit boundary. To verify this, its tangent is $x_\psi(s)(-\rho_\psi(s), 1)/(1 + s\rho_\psi(s))$, so the outward normal has strictly increasing slope $\rho_\psi(s)$. The endpoint normals are $(1, 0)$ and $(1, 1)$, and the reflected pieces join tangentially. Every nonzero normal has a unique support point; the resulting symmetric gauge n_ψ is strictly convex. The support point for normal $(1, \lambda)$ has coordinate ratio $R_\psi(\lambda)$.

Set $N_\psi(Z) := n_\psi(\sigma_1(Z), \sigma_2(Z))$. Singular-value weak majorization and the symmetric-gauge triangle inequality make this a matrix norm. Von Neumann's trace inequality and the preceding support construction give, for nonzero S,

$$V_\psi(S) := \frac{P_\psi(S/\|S\|_{\mathrm{F}})}{N_\psi(P_\psi(S/\|S\|_{\mathrm{F}}))} \in \operatorname{argmax}_{N_\psi(V)\le 1}\langle S, V\rangle.$$

This matrix support point is unique. Distinct positive singular values fix the paired singular subspaces; repeated positive singular values have equal output singular values, so the allowed rotations cancel. At rank one the second output singular value is zero, as for the axial normal of the vector gauge. Thus the matrix unit ball is strictly convex. Its strictly separable maximum-margin solution is unique: the average of two different maximizers would be strictly inside the unit ball and could be scaled up to increase all margins.

We next prove the selection claim for the actual momentum updates. Write $N = N_\psi$, let N_* be its dual norm, and put

$$b_t := N(D_t), \qquad V_t := D_t/b_t, \qquad \alpha_t := \eta_t b_t.$$

Since $n_\psi(1, 0) = 1$, one has $\|Z\|_{\mathrm{op}} \le N(Z) \le \|Z\|_*$ and $1/\sqrt{2} \le b_t \le 2$. Fixed norm equivalence and Lemma B.13 imply

$$\frac{N_*(M_t - G_t)}{N_*(G_t)} = O(\eta_t), \qquad 0 \le \delta_t := 1 - \frac{\langle G_t, V_t\rangle}{N_*(G_t)} \le 2\frac{N_*(M_t - G_t)}{N_*(G_t)}.$$

In particular, $0 \le \delta_t \le 1$ after a finite prefix. Let

$$\gamma_N := \max_{N(W)\le 1} m(W) > 0, \qquad B_N := \max_{i,c\ne y_i} N_*(A_{i,c}).$$

The positive gradient expansion gives $N_*(G_t) \ge \gamma_N \Pi_t$, where $\Pi_t := \Pi(W_t)$ is the average total incorrect-class probability. Along an N-unit direction, the cross-entropy Hessian is at most $B_N^2\Pi$, and $\Pi(W - sV) \le e^{B_N s}\Pi(W)$. Taylor's formula therefore yields

$$\mathcal{L}(W_{t+1}) - \mathcal{L}(W_t) \le -\alpha_t\langle G_t, V_t\rangle + C_N\alpha_t^2\Pi_t, \qquad C_N := B_N^2 e^{2B_N\eta_0}/2.$$

For $\mathcal{F}(\ell) := \log(e^{n\ell} - 1)$, set $\zeta(\ell) := (1 - e^{-n\ell})/n$. The product of the group partition functions gives $1 \le \Pi_t/\zeta(\mathcal{L}(W_t)) \le n$. Concavity of $\mathcal{F}$ and the eventual bound $\delta_t \le 1$ imply

$$\mathcal{F}(\mathcal{L}(W_{t+1})) - \mathcal{F}(\mathcal{L}(W_t)) \le -\gamma_N \alpha_t (1 - \delta_t) + n C_N \alpha_t^2.$$

Since b_t is bounded above and below by positive constants, $\sum_{t<T} \alpha_t \asymp H_T$. Moreover,

$$\sum_{t<T} \alpha_t \delta_t + \sum_{t<T} \alpha_t^2 = O(1 + S_T) = o(H_T).$$

Summing the transformed-loss inequality from the finite prefix, and using $m(W_T) \ge -\mathcal{F}(\mathcal{L}(W_T))$ and $N(W_T) \le N(W_{t_0}) + \sum_{t=t_0}^{T-1} \alpha_t$, therefore gives $m(W_T)/N(W_T) \to \gamma_N$ and $N(W_T) \to \infty$. Compactness and uniqueness of the maximizer imply convergence of the whole N-normalized direction, and hence of its operator-normalized version. □

B.6 Sharpness and Non-Necessity

We next examine the sharpness of the path-weighted defect condition. Two complementary constructions are useful. A positive limiting defect can coexist with exact margin recovery, while a fixed finite map can also produce a nonzero limiting spectral-margin gap.

Throughout this subsection, the loss is the grouped cross-entropy in (14), the momentum parameter is fixed in $[0, 1)$, and

$$G_t = \nabla \mathcal{L}(W_t), \qquad M_t = (1 - \beta) \sum_{j=0}^{t} \beta^j G_{t-j}, \qquad \eta_t = \frac{\eta_0}{\sqrt{t+1}}, \quad \eta_0 > 0.$$

The update is $W_{t+1} = W_t - \eta_t D_t$, with $D_t = P_{\psi_t}(M_t/\|M_t\|_{\mathrm{F}})$. Every polynomial is admissible, and all statements concern exact arithmetic. We continue to use the quantities $c_t, a_t, e_t, A_T, E_T, H_T$ defined in Appendix B.2. Theorem B.1 ensures nonzero gradients, momentum signals, and updates at every finite state, vanishing loss, and eventual divergence of the parameter norm. Parameter directions are used only when $W_t \neq 0$, and E_T/A_T is used only for $T \ge 1$.

Both the first construction and the time lower bound reduce part of the dynamics to a scalar recursion with a stable zero. The following lemma gives the convergence property needed for these recursions, without requiring square-summable learning rates.

Lemma B.15 (Vanishing-step scalar trapping)**.** *Let a scalar sequence satisfy*

$$z_{t+1} = z_t + \eta_t (F(z_t) + \epsilon_t),$$

where F is continuous and bounded, $\eta_t > 0$, $\eta_t \to 0$, $\sum_t \eta_t = \infty$, and $\epsilon_t \to 0$. Suppose that F has a unique zero z_*, is positive to the left of z_*, and is negative to its right. Assume further that, for every $\varepsilon > 0$,

$$\inf_{z \le z_* - \varepsilon} F(z) > 0, \qquad \inf_{z \ge z_* + \varepsilon} [-F(z)] > 0.$$

Then $z_t \to z_*$ as $t \to \infty$.

Proof. Fix $\varepsilon > 0$. At sufficiently large times, the error is smaller than half the two displayed drift bounds for $\varepsilon/2$, and the absolute step is smaller than $\varepsilon/2$. Outside $[z_* - \varepsilon/2, z_* + \varepsilon/2]$, the drift points inward. The divergence of $\sum_t \eta_t$ forces an entry into this interval: otherwise the increments would have a fixed sign and magnitude at least a positive constant times η_t, contradicting remaining on that side. A crossing from one side to the other also enters the interval because the step is small. After entry, an exit from the smaller interval can overshoot by less than $\varepsilon/2$, and the subsequent drift points inward until re-entry. Thus the larger ε-interval is eventually invariant. Applying this argument to every $\varepsilon > 0$ proves convergence. □

B.6.1 A positive average defect can coexist with exact margin recovery

We first give a symmetric three-class example for which the spectral max-margin direction is recovered even though the weighted defect has a positive limit. This example concerns a single fixed polynomial and a specified initialization.

Theorem B.16 (The vanishing-defect condition is not necessary)**.** *Consider $k = 3$ classes in $d = 2$ dimensions, with $y_i = i$ and*

$$h_1 = (1/\sqrt{3}, 0), \qquad h_2 = (1/\sqrt{3}, 1), \qquad h_3 = (1/\sqrt{3}, -1).$$

Let

$$Q := \begin{pmatrix} 2/\sqrt{6} & 0 \\ -1/\sqrt{6} & 1/\sqrt{2} \\ -1/\sqrt{6} & -1/\sqrt{2} \end{pmatrix},$$

whose columns form an orthonormal basis of the centered class subspace. Run (13) *from $W_0 = 0$ and $M_{-1} = 0$, using any fixed $\beta \in [0, 1)$, the learning rate $\eta_t = \eta_0/\sqrt{t+1}$, and one fixed admissible polynomial ψ.*

There is a unique $r_\psi \in (0, 1)$ satisfying

$$\frac{\psi(r_\psi/\sqrt{1+r_\psi^2})}{\psi(1/\sqrt{1+r_\psi^2})} = \frac{1}{2}. \tag{42}$$

The optimal spectral margin and the limiting parameter direction are

$$\gamma_* = \frac{1}{2\sqrt{2}}, \qquad \frac{W_t}{\|W_t\|_{\mathrm{op}}} \to Q\,\mathrm{diag}(1/2, 1). \tag{43}$$

Nevertheless, the instantaneous and path-weighted defects satisfy

$$e_t \to \frac{r_\psi}{2(1+r_\psi)} > 0, \qquad \frac{E_T}{A_T} \to \frac{r_\psi}{2(1+r_\psi)}. \tag{44}$$

The maximum-margin solution is unique in the centered space $\mathbf{1}^\top W = 0$. No uniqueness in the uncentered parameter space is asserted.

Proof. Since $Q^\top Q = I_2$, write a centered matrix as $W = QZ$, where $Z = \left(\begin{smallmatrix} u & x \\ y & v \end{smallmatrix}\right)$. Averaging the two constraints of sample 1 and, separately, the class-1 competing constraints of samples 2 and 3 gives

$$m(W) \le \min\{u/\sqrt{2}, (v-u)/\sqrt{2}\} \le \frac{v}{2\sqrt{2}} \le \frac{\|Z\|_{\mathrm{op}}}{2\sqrt{2}}.$$

The matrix $Z_* = \mathrm{diag}(1/2, 1)$ attains the bound; its other two margins are $\sqrt{2}$. At a centered optimum with $\|Z\|_{\mathrm{op}} \le 1$, equality requires $v = 1$ and $u = 1/2$. The row and column Euclidean norm bounds then force $x = y = 0$. Centering is an orthogonal projection that preserves all margins and cannot increase the operator norm, proving the claimed value of γ_* in the full space as well.

Reflection symmetry preserves $W_t = Q\,\mathrm{diag}(u_t, v_t)$ and the same form for the momentum. For this proof, put

$$P_m := u/\sqrt{2}, \qquad R_m := (v-u)/\sqrt{2}, \qquad z := R_m - P_m.$$

The exact grouped cross-entropy loss is

$$\mathcal{L}(u, v) = \frac{1}{3}\log(1 + 2e^{-P_m}) + \frac{2}{3}\log(1 + e^{-R_m} + e^{-2(P_m+R_m)}).$$

Define

$$A := \frac{2e^{-P_m}}{3(1+2e^{-P_m})},$$

$$B_m := \frac{2e^{-R_m}}{3(1+e^{-R_m}+e^{-2(P_m+R_m)})},$$

$$C := \frac{2e^{-2(P_m+R_m)}}{3(1+e^{-R_m}+e^{-2(P_m+R_m)})}.$$

The two coordinates of the negative gradient are $(A - B_m, B_m + 2C)/\sqrt{2}$. Theorem B.1 gives $P_{m,t}, R_{m,t} \to \infty$. Consequently, its normalized direction differs from

$$\frac{(e^{z_t} - 1, 1)}{\sqrt{(e^{z_t}-1)^2 + 1}}$$

by a quantity tending to zero uniformly in z_t along the trajectory. Indeed, $A = (2/3)e^{-P_m}(1+o(1))$, $B_m = (2/3)e^{-R_m}(1+o(1))$, and $C/e^{-R_m} \to 0$. The norm of $(e^z - 1, 1)$ controls $1 + e^z$ up to an absolute factor. Lemma B.13 transfers this conclusion to the negative momentum signal. Normalizing a nonzero signal is continuous under a vanishing relative error, and the fixed polynomial ψ is Lipschitz on $[-1, 1]$. Therefore,

$$z_{t+1} - z_t = \eta_t\{F_\psi(z_t) + o(1)\},$$

where

$$F_\psi(z) := \frac{1}{\sqrt{2}}\left[\psi\left(\frac{1}{\sqrt{1+(e^z-1)^2}}\right) - 2\psi\left(\frac{e^z-1}{\sqrt{1+(e^z-1)^2}}\right)\right].$$

For $z < 0$, oddness gives $F_\psi(z) > 0$. For $z \geq 0$, the ratio

$$r \mapsto \frac{\psi(r/\sqrt{1+r^2})}{\psi(1/\sqrt{1+r^2})}$$

is strictly increasing because its numerator strictly increases and its positive denominator strictly decreases. It goes from zero to one on $[0, 1]$. Hence F_ψ has exactly one zero, $z_\psi = \log(1 + r_\psi)$. Its limits at the two infinite endpoints are $3\psi(1/\sqrt{2})/\sqrt{2} > 0$ and $-\sqrt{2}$, respectively. Lemma B.15 gives $z_t \to z_\psi$. It follows that $v_t - 2u_t \to \sqrt{2}z_\psi$, while $u_t, v_t \to \infty$, proving (43).

The two singular values of the negative momentum signal have limiting ratio r_ψ, and those of its update have limiting ratio $1/2$. Substitution into the definition of e_t gives $e_t \to r_\psi/[2(1 + r_\psi)]$. Since $A_T \to \infty$, weighted averaging proves the second assertion in (44). □

For the identity map, $r_\psi = 1/2$ and the limiting defect is $1/6$. For one cubic NS stage it is approximately 0.1344500. For this fixed-map trajectory from the zero initialization, exact margin recovery coexists with a positive limiting defect. Thus the vanishing-defect condition in Theorem B.1 is sufficient but not necessary.

B.6.2 An inverse-slope worst-case bias for fixed Taylor maps

We now turn to the opposite behavior. For every fixed finite Taylor–NS map, we construct a separable problem with a nonzero limiting spectral-margin gap. The data depend only on the slope at the origin, and the selection conclusion holds from every finite initialization.

Theorem B.17 (Realizable fixed-map inverse-slope lower bound)**.** *Let ψ be any fixed finite Taylor–NS composition, including the identity as an empty composition. Write $\Lambda = \psi'(0)$ and $r = (2\Lambda)^{-1}$. Consider the three-class dataset with $y_i = i$ and*

$$h_1 = (3, 0), \qquad h_2 = (-r/\sqrt{3}, 1), \qquad h_3 = (-r/\sqrt{3}, -1).$$

Let Q be the matrix in Theorem B.16, and define

$$d_\psi := \frac{\psi(r/\sqrt{1+r^2})}{\psi(1/\sqrt{1+r^2})}.$$

Run (13) *using this fixed map, $M_{-1} = 0$, any $\beta \in [0, 1)$, the learning rate $\eta_t = \eta_0/\sqrt{t+1}$ with $\eta_0 > 0$, and any finite $W_0 \in \mathbb{R}^{3\times 2}$. Then*

$$\frac{W_t}{\|W_t\|_{\mathrm{op}}} \to Q\,\mathrm{diag}(d_\psi, 1), \qquad \gamma_* = (1+r)/\sqrt{2},$$

and the limiting spectral-margin gap satisfies

$$\frac{1}{4\sqrt{2}\Lambda} \leq \gamma_* - \lim_{t\to\infty}\widehat{\gamma}(W_t) = \frac{r(1-d_\psi)}{\sqrt{2}} \leq \frac{1}{2\sqrt{2}\Lambda}. \tag{45}$$

The inputs satisfy $\max_i \|h_i\|_1 = 3$, so we may take $R = 3$ and $B = 3\sqrt{2}$. Moreover, $\gamma_* \geq 1/\sqrt{2}$, independently of the selected map.

Proof. The average of $A_{2,1}$ and $A_{3,1}$ is $\overline{A} = Q\,\mathrm{diag}(r,1)/\sqrt{2}$. Hence $m(W) \le \langle \overline{A}, W\rangle$. The unique support direction of the norm in Lemma B.14 at this atom is proportional to $Q\,\mathrm{diag}(d_\psi, 1)$. At the latter matrix, the six margins have three values,

$$3\sqrt{3/2}\,d_\psi, \qquad (rd_\psi + 1)/\sqrt{2}, \qquad \sqrt{2}.$$

By Lemma C.1, $\psi(s) \ge s$, and $r \le 1/2$,

$$1 - e^{-1/\sqrt{5}} \le d_\psi \le \Lambda r = 1/2.$$

For the lower bound, the denominator is at most one and $\Lambda r/\sqrt{1+r^2} \ge 1/\sqrt{5}$. For the upper bound, bound the numerator by $\Lambda r/\sqrt{1+r^2}$ and the denominator below by $1/\sqrt{1+r^2}$. In particular, $(3\sqrt{3} - r)d_\psi > 1$ and $rd_\psi < 1$. The first inequality also follows without numerical approximation from $1 - e^{-1/\sqrt{5}} > 1/(1+\sqrt{5})$ and $3\sqrt{3} - 1/2 > 1 + \sqrt{5}$. Thus the averaged constraints are exactly the minimum margins at this support direction. The norm support bound is attained, so Lemma B.14 identifies the actual limiting direction for every finite initialization.

For the operator norm, the same atom has nuclear norm $(1+r)/\sqrt{2}$, and QI_2 attains this upper bound while satisfying the remaining constraints. This proves the asserted γ_*. Subtracting the limiting candidate margin gives (45); the inequalities $0 \le d_\psi \le 1/2$ give its two bounds. The maximum input ℓ_1 norm is three, so $R = 3$ and $B = 3\sqrt{2}$ are valid choices. □

Taking $\psi = \psi_q = \phi^{\circ q}$ gives $\Lambda = (3/2)^q$ and proves Proposition 4.2. The order in (45) is a worst-case asymptotic statement: for each fixed map, its hard dataset is selected before the run. The dataset depends on ψ, and the conclusion concerns its limiting margin under that fixed map.

B.7 Time Lower Bound

The preceding lower bound is caused by finite-map distortion. A separate obstruction remains even when every update has zero directional defect. In a rank-one binary example, all admissible NS maps coincide, so the margin gap is determined by the learning-rate schedule rather than by polynomial precision.

Theorem B.18 (A square-root-time obstruction)**.** *Consider $k = 2$ classes in $d = 2$ dimensions, with training data*

$$(h_1, y_1) = ((1,1)/\sqrt{2}, 1), \qquad (h_2, y_2) = (-(1,1)/\sqrt{2}, 2), \qquad (h_3, y_3) = ((1,-1)/\sqrt{2}, 1).$$

Run (13) *from $W_0 = 0$ and $M_{-1} = 0$, with any fixed $\beta \in [0,1)$, learning rate $\eta_t = \eta_0/\sqrt{t+1}$ for $\eta_0 > 0$, and any sequence of admissible NS maps. Then $p = 1$, $c_t = 1$, and $e_t = 0$ at every iteration. Nevertheless,*

$$\gamma_* = 1, \qquad \gamma_* - \widehat{\gamma}(W_T) \sim \frac{\log 2}{4\eta_0\sqrt{T}} \qquad \text{as } T \to \infty. \tag{46}$$

Proof. Write $Q_2 = (1,-1)^\top/\sqrt{2}$ and $W = Q_2(u,v)$. The two distinct effective margin atoms are $(1,1)$ and $(1,-1)$; the first occurs twice. Hence

$$\mathcal{L}(u,v) = \frac{1}{3}\left[2\log(1+e^{-u-v}) + \log(1+e^{-u+v})\right],$$
$$m(W) = u - |v|, \qquad \|W\|_{\mathrm{op}} = \sqrt{u^2+v^2}.$$

The operator-unit maximum margin is one, attained at $(u,v) = (1,0)$. Every nonzero centered two-class gradient and momentum signal has rank one. Its sole normalized singular value is one, and $\psi_t(1) = 1$ for every admissible map. All such maps therefore give the same update direction.

Theorem B.1 gives $u_t \pm v_t \to \infty$. Up to a uniformly vanishing relative error, the negative gradient is proportional to $(2e^{-v_t} + e^{v_t}, 2e^{-v_t} - e^{v_t})$. Lemma B.13 gives the same normalized direction for the negative momentum signal, up to $o(1)$. Thus the second-coordinate recursion is

$$v_{t+1} - v_t = \eta_t\{F(v_t) + o(1)\}, \qquad F(v) := \frac{2e^{-v} - e^{v}}{\sqrt{(2e^{-v} + e^{v})^2 + (2e^{-v} - e^{v})^2}}.$$

This continuous bounded drift has a unique stable zero $v_* = (\log 2)/2$, and its limits at the two endpoints are $1/\sqrt{2}$ and $-1/\sqrt{2}$. Lemma B.15 gives $v_t \to v_*$. The first-coordinate update divided by η_t then tends to one, so weighted averaging gives $u_T/H_T \to 1$. Since $H_T \sim 2\eta_0\sqrt{T}$,

$$1 - \frac{u_T - |v_T|}{\sqrt{u_T^2 + v_T^2}} \sim \frac{v_*}{H_T} \sim \frac{\log 2}{4\eta_0\sqrt{T}}.$$

□

This excludes a general $o(T^{-1/2})$ margin guarantee for the stated learning-rate schedule. Determining whether the logarithm in the general upper bound is necessary remains open.

B.8 Polynomial Computation Complexity

We conclude the linear theory by asking how much polynomial computation is required to control the directional defect uniformly over normalized spectra. This is a worst-case approximation question, distinct from controlling the defect along one realized training trajectory. For an admissible polynomial ψ and an integer $p \geq 1$, define

$$\overline{e}_p(\psi) := \sup_{\substack{1 \leq r \leq p,\ s_i > 0 \\ \sum_i s_i^2 = 1}} \left[1 - \frac{\sum_i s_i \psi(s_i)}{\psi(s_{\max}) \sum_i s_i}\right], \qquad s_{\max} := \max_i s_i.$$

Thus $\overline{e}_p(\psi)$ is a worst-case pointwise directional certificate over matrices of rank at most p.

A permitted circuit starts from an already Frobenius-normalized matrix and uses fixed coefficients, linear combinations, transposes, and matrix multiplications. We count only matrix multiplications. Input normalization, arithmetic on fixed coefficients or input-independent bookkeeping, and additions are excluded from this count. Entrywise extraction, input-dependent scalar multiplication and reassembly, SVD, matrix inversion, and spectrum-dependent free coefficient selection are not permitted circuit operations. All subsequent operations remain polynomial in the normalized input entries.

We first lower-bound the polynomial degree required for uniform defect control and then translate this requirement into matrix-multiplication complexity. Cubic NS achieves the same logarithmic order.

Theorem B.19 (Robust polynomial multiplication complexity)**.** *Let $p \geq 2$, let ψ be an admissible polynomial, and write $D_\psi := \deg \psi$. Then*

$$\overline{e}_p(\psi) \geq \frac{1}{16} \min\left\{1, \frac{p-1}{D_\psi}\right\}. \tag{47}$$

Suppose that a permitted circuit using m matrix multiplications implements this map on every Frobenius-normalized input of rank at most p. If $\overline{e}_p(\psi) \leq \varepsilon < 1/16$, then

$$D_\psi \geq \frac{p-1}{16\varepsilon}, \qquad m \geq \log_2 \frac{p-1}{16\varepsilon}. \tag{48}$$

Let $m_*(p, \varepsilon)$ denote the minimum multiplication count among permitted circuits that implement an admissible map with $\overline{e}_p(\psi) \leq \varepsilon$. Then

$$m_*(p, \varepsilon) = \Theta\left(\log \frac{p-1}{\varepsilon}\right) \qquad (p \geq 2,\ \varepsilon \downarrow 0).$$

For $p = 1$, the identity map already has zero directional defect and requires no matrix multiplication.

Proof. Oddness and admissibility imply $\sup_{[-1,1]} |\psi| \leq 1$. The classical Bernstein polynomial inequality gives

$$|\psi'(s)| \leq \frac{D_\psi}{\sqrt{1 - s^2}}, \qquad -1 < s < 1.$$

Put $\ell := \min(p-1, D_\psi)$ and $s := 1/(4D_\psi)$. Choose the normalized spectrum consisting of ℓ copies of s and one value $s_0 := \sqrt{1 - \ell s^2}$. Then $s_0 \geq \sqrt{15}/4 > 3/4$ and $c := \psi(s_0) \geq s_0$. Since $\psi(0) = 0$,

$$\psi(s) \leq D_\psi \arcsin s \leq 2D_\psi s = 1/2.$$

The large singular value contributes zero to the directional-defect numerator. The remaining terms give

$$e_\psi = \frac{\ell s}{s_0 + \ell s}\left(1 - \frac{\psi(s)}{c}\right) \geq \frac{\ell s}{5/4}\frac{1}{3} = \frac{\ell}{15 D_\psi} \geq \frac{1}{16}\min\{1, (p-1)/D_\psi\}.$$

This proves (47) and the degree requirement in (48). Lemma C.2 gives $D_\psi \leq 2^m$ even on the Frobenius-normalized domain, proving the multiplication lower bound.

For the matching order, use $\psi = f_1^{\circ q}$, with $\Lambda = (3/2)^q$. Let $\chi = \sum_i s_i$ and $c = \psi(s_{\max})$. Since $c\chi \geq 1$, omitting one maximizing singular value and using Lemma C.1 gives

$$e_\psi \leq (r-1)\max_{0 \leq s \leq 1} s(1 - \psi(s)) \leq \frac{p-1}{\mathrm{e}\Lambda}.$$

Here, $(1-s)^\Lambda \leq e^{-\Lambda s}$ and $\max_{s \geq 0} s e^{-\Lambda s} = 1/(\mathrm{e}\Lambda)$. Therefore,

$$q = \left\lceil \frac{\log((p-1)/(\mathrm{e}\varepsilon))}{\log(3/2)} \right\rceil$$

suffices in the stated small-tolerance range. A cubic stage can be evaluated by forming $XX^\top$ and then $(3I/2 - XX^\top/2)X$, using two matrix multiplications. The upper count $2q$ and (48) have the same logarithmic order; their leading constants need not match. □

The computation theorem establishes the cost of uniform pointwise defect control over all normalized rank-at-most-p spectra, including arbitrarily small positive singular values. It complements the trajectory-level margin and time results. A joint training-complexity lower bound would require a construction that enforces both obstructions.

C Technical Lemmas

This appendix provides the scalar polynomial envelope used by the fixed-map construction and the degree estimate used in the computation lower bound. The optimization and margin arguments remain in Appendix B.

C.1 Polynomial and Bernstein Bounds

Recall the Taylor–NS maps f_r and their slopes a_r from (15). The next lemma controls a finite composition using only its slope at the origin. Its lower envelope is used in the fixed-map construction, while its upper envelope controls the response to small singular values.

Lemma C.1 (Taylor slope envelope)**.** *Let ψ be a finite composition of the polynomials*

$$f_r(s) = s\sum_{j=0}^{r} 4^{-j}\binom{2j}{j}(1-s^2)^j, \qquad r \geq 0,$$

where the identity is allowed as an empty composition. Write $\Lambda := \psi'(0)$. Then, for every $s \in [0,1]$,

$$1 - (1-s)^\Lambda \leq \psi(s) \leq \min\{1, \Lambda s\}. \tag{49}$$

Proof. For $r = 0$, the lower envelope is an equality. For $r \geq 1$, put $a_r = f_r'(0) = (2r+1)4^{-r}\binom{2r}{r}$. Then $1 < a_r < r+1$ and $f_r'(s) = a_r(1-s^2)^r$. For $g_{a_r}(s) := 1 - (1-s)^{a_r}$, the derivative ratio is

$$\frac{f_r'(s)}{g_{a_r}'(s)} = (1-s)^{r-a_r+1}(1+s)^r.$$

Its logarithmic derivative has numerator $a_r - 1 - (2r - a_r + 1)s$. The ratio increases from one, then decreases to zero, crossing one once in the interior. Since f_r and g_{a_r} agree at both endpoints, $f_r \geq g_{a_r}$. Composition uses $g_b \circ g_{b'} = g_{bb'}$. Each Taylor stage is increasing and concave on $[0,1]$, and the same holds for their composition. Its tangent at zero gives $\psi(s) \leq \Lambda s$; admissibility gives $\psi(s) \leq 1$. □

Bernstein control of small singular values. The classical inequality $|\psi'(s)| \le D_\psi/\sqrt{1-s^2}$ for $\sup_{[-1,1]} |\psi| \le 1$ is used in the proof of Theorem B.19 to build a spectrum with a directional defect of order $\min\{1, (p-1)/D_\psi\}$. That construction is retained with the computation theorem.

C.2 Matrix-Circuit Degree Bounds

Counting entrywise polynomial degree gives a simple bound for circuits on unconstrained inputs. To use that bound here, we must also show that restricting the circuit to Frobenius-normalized inputs cannot hide the degree of the implemented spectral polynomial. The following lemma establishes both statements.

Lemma C.2 (Polynomial degree on the normalized input domain)**.** *Consider the circuit model in Appendix B.8. Every output entry computed with m matrix multiplications is a polynomial of degree at most 2^m in the input entries. If the circuit implements an odd spectral polynomial ψ on every Frobenius-normalized input of rank at most $p \ge 2$, then $\deg \psi \le 2^m$.*

Proof. After m matrix multiplications, every circuit output entry has polynomial degree at most 2^m in the input entries, since one multiplication at most doubles the largest previous degree. To apply this fact on the normalized input domain, restrict to an embedded two-dimensional block $X(\theta) = \mathrm{diag}(\cos\theta, \sin\theta)$. Its Frobenius norm is one. A circuit output entry is then a trigonometric polynomial of frequency at most 2^m. The first diagonal entry of the required output is $\psi(\cos\theta)$, whose highest nonzero frequency is $D_\psi = \deg \psi$. Oddness ensures that this signed formula also holds when a diagonal entry is negative. Thus $D_\psi \le 2^m$, proving the degree bound on the normalized input domain and the multiplication lower bound used in Theorem B.19. □