PROVABLE ADVANTAGES OF MUON FROM DATA-INDUCED CURVATURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon, a matrix-orthogonalization-based optimizer, has shown strong empirical performance in large-scale language-model training, yet when and why spectral normalization yields a sustained optimization advantage remain incompletely understood. Since Muon acts on matrix-valued parameter blocks, we study a one-hidden-layer quadratic network with a single trainable matrix, which retains the matrix-gradient structure while linking input moments explicitly to parameter curvature. Motivated by empirical feature spectra, we consider Gaussian inputs with geometrically decaying centered variances and a nonzero mean whose direction is close to an active feature. In this model, data geometry induces anisotropic curvature and an imbalanced gradient spectrum, creating a regime favorable to spectral normalization. Under explicit conditions, we prove that exact-polar Muon achieves an optimization advantage over both gradient descent and SignGD, a special case of Adam's coordinatewise normalization, and that this advantage persists along the nonlinear trajectories on a fixed empirical objective. These results establish a provable mechanism linking data-induced curvature to the optimization advantage of spectral normalization.

1 INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

Muon has shown strong empirical performance in large-scale language-model training [22], motivating a growing theoretical effort to understand its optimization behavior [3, 16, 25, 35]. A central question is when and why Muon can achieve a sustained optimization advantage over competing methods in neural network training. We approach this question by studying how Muon's spectral normalization interacts with data-induced curvature in a nonlinear network model.

At the core of Muon is spectral normalization of a matrix update: it preserves the singular vectors while replacing the nonzero singular values by a common scale. When the gradient spectrum is imbalanced, this redistributes the update across its singular directions rather than preserving their relative magnitudes. To isolate this effect, we compare Muon with GD and SignGD [4], a memoryless special case of Adam's coordinatewise normalization [15, 2, 40]. Muon versus GD isolates the effect of spectral normalization, whereas Muon versus SignGD distinguishes spectral from coordinatewise normalization.

Since Muon acts blockwise on matrix-valued parameters, we isolate the geometry of a single trainable matrix and consider a one-hidden-layer network. This retains the matrix structure on which Muon operates while removing interactions across parameter blocks. For a trainable matrix $\boldsymbol{W} \in \mathbb{R}^{d\times r}$, we consider the quadratic response $f_{\boldsymbol{W}}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{x}$. This nonlinear parameterization retains a nontrivial matrix-gradient spectrum, while its curvature at an exact fit is determined explicitly by fourth moments of the input. It therefore provides a direct link from input statistics to parameter curvature and the resulting gradient spectrum. Quadratic models of this form are also closely related to nonconvex matrix factorization from rank-one measurements [20].

The input model is motivated by geometric patterns observed both in prior studies of language-model representations and in our experiments. Strong anisotropy and dominant spectral directions have been widely reported in neural language representations [27, 9, 10], while the mean representation has been observed to align closely with a leading principal direction [37]. Recent work

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.

has further studied the evolution and decay of representation eigenspectra during language-model training [33, 19]. In our experiments, centering substantially reduces a dominant spectral component, while the remaining feature spectrum remains anisotropic; separately, the input mean exhibits a concentrated projection toward a small number of low-index directions in fitted quadratic response models. Motivated by these observations, we model the centered covariance by a geometric spectral decay and study the regime in which the mean direction is close to an active feature. Through the input fourth moments, the mean contribution and covariance decay induce anisotropic curvature along the active parameter directions. When the mean contribution is sufficiently strong, one direction has larger curvature than the aggregate remaining directions. GD retains the resulting imbalance in gradient magnitudes, whereas Muon equalizes the corresponding singular-value amplitudes. This curvature imbalance induces the distinct optimization dynamics analyzed in this work.

Main results. Our main result shows that, under an explicit data-induced curvature-separation condition and for a suitable local initialization, Muon achieves a strict finite-step advantage over GD on a fixed nonlinear empirical objective. Our analysis makes this connection in two stages: an explicit population quadratic model isolates the effect of spectral normalization, and a perturbation argument transfers the resulting separation to nonlinear finite-sample trajectories from a local initialization. A complementary comparison with SignGD further highlights how spectral and coordinatewise normalization lead to distinct optimization dynamics.

1.2 Related Work

Muon spectrally normalizes momentum updates for hidden-layer weight matrices using Newton–Schulz iterations [13]. Early experiments demonstrated substantial training-speed improvements over AdamW [23, 13]. Liu et al. [22] subsequently established its scalability to large-scale LLM pretraining, reporting roughly $2\times$ compute efficiency over AdamW. Subsequent work has refined Muon's update rule and reduced its computational overhead. Li et al. [21] introduced NorMuon, which uses neuron-wise adaptive normalization to balance update magnitudes across neurons. Amsel et al. [1] developed Polar Express, a theoretically optimal method for computing the matrix sign, or polar factor, in low precision, thereby accelerating Muon's orthogonalization step.

In addition to empirical improvements, a growing line of work seeks to explain why Muon can outperform Adam. Bernstein & Newhouse [3] interpreted matrix-sign updates as steepest descent under the operator norm, while viewing Adam's element-wise normalization as steepest descent under the ℓ_∞ norm. Kovalev [16] further showed that gradient orthogonalization maximizes first-order descent under a spectral-norm trust region, yielding non-convex convergence guarantees and motivating momentum before orthogonalization for variance reduction. From a preconditioning perspective, Lau et al. [17] argued that Muon balances updates across singular directions rather than directly correcting Hessian anisotropy, while Chen et al. [8] showed that decoupled weight decay induces spectral-norm constraints on weight matrices and thereby spectral regularization. Complementary analyses establish convergence benefits under favorable structure: Shen et al. [31] studied low-rank and approximately block-diagonal Hessians, and Ma et al. [25] showed that spectral orthogonalization can remove condition-number dependence in matrix factorization and linear-transformer models. Related work on coordinatewise normalization studies Adam and SignGD from sign, variance, and ℓ_∞-geometry perspectives [15, 4, 2, 40, 39]; most closely related to our SignGD comparison, Paquette et al. [28] characterized covariance-dependent regimes separating spectral and entrywise normalization. Other work examines concrete learning mechanisms, linking Muon to more balanced learning of rare associations and favorable capacity and scaling properties [36, 14, 18], more uniform learning on imbalanced data [34], and more robust and transferable features associated with larger margins and higher-rank representations [30]. Finally, loss-landscape studies attribute Muon's gains to lower curvature penalties and directional sharpness [35], while Shumaylov et al. [32] questioned explanations based on precise spectral geometry and instead emphasized local alignment, descent potential, and step-size behavior. Anisotropy, dominant spectral directions, and decaying eigenspectra have also been widely observed in neural language representations [27, 9, 37, 10, 33, 19, 12]. The quadratic model studied here is closely related to nonconvex matrix factorization from rank-one measurements [20], while related provable analyses of nonlinear and structured matrix optimization include [24, 38, 7, 5, 6]. In contrast to prior quadratic and curvature-based analyses [25, 35], we derive the relevant curvature directly from the input distribution and establish a finite-step comparison along distinct nonlinear trajectories on one fixed empirical objective.

Organization and notation. Section 2 introduces the model and optimization methods. Section 3 studies the population quadratic dynamics. Section 4 presents the main theoretical results. Section 5 reports the numerical experiments. Section 6 concludes the paper. The proofs are deferred to Appendix B.

Vectors and matrices are bold. The norms $\|\cdot\|_2$, $\|\cdot\|_{\mathrm{F}}$, $\|\cdot\|_{\mathrm{op}}$, and $\|\cdot\|_*$ denote the Euclidean, Frobenius, spectral, and nuclear norms, respectively. Operator norms on matrices are induced by the Frobenius norm, and $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{Y})$. We write $[t]_+ = \max\{t, 0\}$, use $\mathbf{1}$ for an all-ones vector, and let $\boldsymbol{e}_j$ denote the j-th standard basis vector in the indicated space.

2 Model and Population Curvature

In this section, we specify the input distribution, response, and optimization methods. We then compute the population Hessian at the target factor. Its restriction to the singular parameter directions determines the quadratic comparison in Section 3.

2.1 Data and Response Model

Let $d \geq r \geq 3$, and let the orthogonal matrix $\boldsymbol{A}_0 \in \mathbb{R}^{d\times d}$ specify the input feature directions, ordered by decreasing centered variance. Let $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r)$ consist of its first r columns, and let $\bar{\boldsymbol{u}} \in \mathbb{R}^d$ be a unit vector. Motivated by the approximately geometric spectral decay and the mean-induced spikeness observed in Figure 2, we model the inputs as

$$\boldsymbol{x} = \mu\bar{\boldsymbol{u}} + \sqrt{\gamma}\,\boldsymbol{A}_0\,\mathrm{diag}(1, \rho^{1/2}, \ldots, \rho^{(d-1)/2})\boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{I}_d), \tag{2.1}$$

where $\mu > 0$, $\gamma > 0$, and $0 < \rho < 1$. Thus

$$\mathbb{E}\boldsymbol{x} = \mu\bar{\boldsymbol{u}}, \qquad \mathrm{Cov}(\boldsymbol{x}) = \gamma\boldsymbol{A}_0\,\mathrm{diag}(1, \rho, \ldots, \rho^{d-1})\boldsymbol{A}_0^\top,$$

and

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \mathrm{Cov}(\boldsymbol{x}) + \mu^2\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^\top. \tag{2.2}$$

Hence centering removes the rank-one mean contribution, while the centered covariance has eigenvalue ratio ρ. We observe N independent copies $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ of $\boldsymbol{x}$. The response uses the first r input features:

$$y_t = \boldsymbol{x}_t^\top \boldsymbol{X}_\star \boldsymbol{x}_t, \quad t = 1, \ldots, N, \qquad \boldsymbol{X}_\star = \boldsymbol{U}\,\mathrm{diag}(\sigma_j^2)\boldsymbol{U}^\top,$$

where $\sigma_1, \ldots, \sigma_r > 0$. Equivalently, $y_t = \sum_{j=1}^r \sigma_j^2(\boldsymbol{u}_j^\top \boldsymbol{x}_t)^2$. Thus the j-th active feature has centered variance $\gamma\rho^{j-1}$ and response weight σ_j^2. The coefficients σ_j are indexed by these input directions. Let $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r) \in \mathbb{R}^{r\times r}$ be orthogonal and set

$$\boldsymbol{W}_\star = \boldsymbol{U}\,\mathrm{diag}(\sigma_j)\boldsymbol{V}^\top, \qquad \boldsymbol{W}_\star\boldsymbol{W}_\star^\top = \boldsymbol{X}_\star. \tag{2.3}$$

2.2 Training Objective and Algorithms

To study the spectral mechanism behind Muon, we consider the following quadratic network model, which yields a nontrivial matrix-gradient spectrum and admits an explicit characterization of its population dynamics. For $\boldsymbol{W} \in \mathbb{R}^{d\times r}$, define the network and its empirical loss by

$$f_{\boldsymbol{W}}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{x} = \|\boldsymbol{W}^\top \boldsymbol{x}\|_2^2, \qquad \mathcal{L}_N(\boldsymbol{W}) = \frac{1}{2N}\sum_{t=1}^N (f_{\boldsymbol{W}}(\boldsymbol{x}_t) - y_t)^2. \tag{2.4}$$

Since $y_t = f_{\boldsymbol{W}_\star}(\boldsymbol{x}_t)$, we have $\mathcal{L}_N(\boldsymbol{W}_\star) = 0$ for every sample. We compare two methods that assign different magnitudes to the singular components of a matrix gradient. To isolate the effect of spectral normalization, we study the exact polar update without momentum and use line search to factor out stepsize selection. For a full-column-rank gradient with singular value decomposition $\boldsymbol{G} = \boldsymbol{U}_{\boldsymbol{G}}\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{V}_{\boldsymbol{G}}^\top$, define

$$\mathcal{D}_{\mathrm{M}}(\boldsymbol{G}) = \mathrm{polar}(\boldsymbol{G}) = \boldsymbol{U}_{\boldsymbol{G}}\boldsymbol{V}_{\boldsymbol{G}}^\top, \qquad \mathcal{D}_{\mathrm{G}}(\boldsymbol{G}) = \boldsymbol{G}. \tag{2.5}$$

Starting from $\boldsymbol{W}_0$, Muon and GD use

$$\boldsymbol{W}_{k+1}^{\mathrm{A}} = \boldsymbol{W}_k^{\mathrm{A}} - \alpha_k^{\mathrm{A}}\mathcal{D}_{\mathrm{A}}(\nabla\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{A}})), \qquad \mathrm{A} \in \{\mathrm{M}, \mathrm{G}\}. \tag{2.6}$$

The step α_k^{A} is the first positive strict local minimizer of $\alpha \mapsto \mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{A}} - \alpha\mathcal{D}_{\mathrm{A}}(\nabla\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{A}})))$.

2.3 POPULATION CURVATURE

To compare the mean direction with the active features, choose

$$j_\star \in \arg\max_{1\le j\le r} |\bar{\boldsymbol{u}}^\top \boldsymbol{u}_j|.$$

Orient $\boldsymbol{u}_{j_\star}$ and $\boldsymbol{v}_{j_\star}$ simultaneously so that $\bar{\boldsymbol{u}}^\top \boldsymbol{u}_{j_\star} \ge 0$, leaving $\boldsymbol{W}_\star$ unchanged, and define

$$\delta = \|\bar{\boldsymbol{u}} - \boldsymbol{u}_{j_\star}\|_2 = \min_{1\le j\le r} \min_{s\in\{\pm 1\}} \|\bar{\boldsymbol{u}} - s\boldsymbol{u}_j\|_2. \tag{2.7}$$

The index $j_\star$ selects the active direction nearest to the mean, whereas the index j in the data model retains the variance ordering. Let $\mathcal{H}_0$ be the Hessian at $\boldsymbol{W}_\star$ of the population loss when the Gaussian input has mean $\mu\boldsymbol{u}_{j_\star}$, with the same centered covariance and response rule $y = f_{\boldsymbol{W}_\star}(\boldsymbol{x})$. To identify the curvature along the parameter singular directions, note that $Df_{\boldsymbol{W}_\star}(\boldsymbol{x})[\boldsymbol{u}_j\boldsymbol{v}_j^\top] = 2\sigma_j(\boldsymbol{u}_j^\top\boldsymbol{x})^2$. The matrices $\boldsymbol{u}_j\boldsymbol{v}_j^\top$, $1 \le j \le r$, form an orthonormal basis of

$$\mathcal{S} = \{\boldsymbol{U}\,\mathrm{diag}(\boldsymbol{z})\boldsymbol{V}^\top : \boldsymbol{z} \in \mathbb{R}^r\}. \tag{2.8}$$

Define $K_{ij} = \langle \boldsymbol{u}_i\boldsymbol{v}_i^\top, \mathcal{H}_0[\boldsymbol{u}_j\boldsymbol{v}_j^\top]\rangle$ for $1 \le i, j \le r$. These entries are second moments of the linearized predictions, hence fourth moments of the input. For $1 \le i, j \le r$, the Gaussian moment formulas give

$$K_{j_\star j_\star} = 4\sigma_{j_\star}^2(\mu^4 + 6\mu^2\gamma\rho^{j_\star-1} + 3\gamma^2\rho^{2j_\star-2}), \tag{2.9a}$$

$$K_{j_\star j} = K_{jj_\star} = 4\gamma\sigma_{j_\star}\sigma_j(\mu^2 + \gamma\rho^{j_\star-1})\rho^{j-1}, \qquad j \ne j_\star, \tag{2.9b}$$

$$K_{ij} = 4\gamma^2\sigma_i\sigma_j\rho^{i+j-2}\big(1 + 2\mathbf{1}_{\{i=j\}}\big), \qquad i, j \ne j_\star. \tag{2.9c}$$

The following proposition shows that $\mathcal{S}$ is invariant under $\mathcal{H}_0$ and that the matrix $\boldsymbol{K}$ above represents its restriction to this subspace.

Proposition 2.1. *Suppose the input and response model in Section 2.1 holds. Then $\mathcal{H}_0(\mathcal{S}) \subseteq \mathcal{S}$, the restriction of $\mathcal{H}_0$ to $\mathcal{S}$ is represented by (2.9), and $\boldsymbol{K} \succ 0$.*

The proof is given in Appendix B.1. For a displacement $\boldsymbol{U}\,\mathrm{diag}(\boldsymbol{z})\boldsymbol{V}^\top$, the gradient of the reference quadratic loss has singular values $|[\boldsymbol{K}\boldsymbol{z}]_j|$. In (2.9), the mean contributes a term of order μ^4 to $K_{j_\star j_\star}$ and terms of order μ^2 to $K_{j_\star j}$, $j \ne j_\star$. The remaining entries depend on the centered variances and response weights. We next use this difference to compare the two update directions.

3 A WARM-UP: POPULATION QUADRATIC MUON AND GD DYNAMICS

We use the curvature matrix in Proposition 2.1 to compare Muon and GD on the aligned quadratic approximation

$$Q(\boldsymbol{z}) = \tfrac{1}{2}\boldsymbol{z}^\top\boldsymbol{K}\boldsymbol{z}.$$

Whenever $\boldsymbol{K}\boldsymbol{z}$ has no zero coordinates, matrix Muon and GD correspond to the directions $\mathrm{sign}(\boldsymbol{K}\boldsymbol{z})$ and $\boldsymbol{K}\boldsymbol{z}$, respectively. We compare the mean-associated coordinate with a common displacement in the remaining coordinates by writing

$$2Q(x\boldsymbol{e}_{j_\star} + y(\mathbf{1} - \boldsymbol{e}_{j_\star})) = Ax^2 + 2Bxy + Dy^2, \quad A = K_{j_\star j_\star}, \quad B = \sum_{j\ne j_\star} K_{j_\star j}, \quad D = \sum_{i,j\ne j_\star} K_{ij}.$$

Here A and D are the curvatures of the mean-associated coordinate and the common displacement in the remaining active coordinates, respectively, and B is their mixed term. The rank-one mean term in (2.2) strengthens the curvature along the distinguished direction $j_\star$. We impose the following curvature separation condition to ensure that this contribution dominates the anisotropic covariance.

Assumption 3.1. *We assume that the mean amplitude satisfies*

$$\frac{\mu^2}{\gamma} > \rho^{j_\star-1} + \frac{4\sum_{j\ne j_\star}\sigma_j^2\rho^{2j-2}}{\sigma_{j_\star}\min_{j\ne j_\star}\sigma_j\rho^{j-1}}.$$

Assumption 3.1 gives $A > D$, which provides the curvature separation underlying the Muon–GD comparison below. A corresponding reduction of the dominant spectral component after centering is also observed empirically; see Figure 2. Exact line search along $\mathbf{1}$ makes the new gradient orthogonal to $\mathbf{1}$. We select a reference displacement with this property by setting

$$a = \frac{B + D}{A + B}. \tag{3.1}$$

The following proposition relates the mean condition to the curvature and gradient at $\mathbf{1} - (1+a)\boldsymbol{e}_{j_\star}$.

Proposition 3.1. *Suppose Assumption 3.1 holds. Then*

$$A > D, \qquad 0 < a < 1, \qquad [\boldsymbol{K}(\mathbf{1}-(1+a)\boldsymbol{e}_{j_\star})]_{j_\star} < 0, \qquad [\boldsymbol{K}(\mathbf{1}-(1+a)\boldsymbol{e}_{j_\star})]_j > 0 \quad (j \neq j_\star).$$

The proof is given in Appendix B.2. When $j_\star = 1$ and $\sigma_1 = \cdots = \sigma_r$, the simpler condition $\mu^2/\gamma \geq \rho/(1-\rho)$ also suffices for Proposition 3.1. Let $\boldsymbol{g} = \boldsymbol{K}(\mathbf{1} - (1+a)\boldsymbol{e}_{j_\star})$. Since $\mathbf{1}^\top \boldsymbol{g} = 0$, Proposition 3.1 gives $|g_{j_\star}| = \sum_{j \neq j_\star} g_j$. We call this identity gradient balance and define $\boldsymbol{w} = (w_j)_{j \neq j_\star}$, ordered by increasing j, by

$$w_j = \frac{g_j}{|g_{j_\star}|}, \qquad j \neq j_\star, \qquad \boldsymbol{w} > 0, \quad \sum_{j \neq j_\star} w_j = 1. \tag{3.2}$$

After division by $|g_{j_\star}|$, the GD direction has coordinate -1 at $j_\star$ and coordinates w_j elsewhere. The Muon direction has the same signs and unit magnitudes. To compare their successive losses, set

$$q = \frac{(A-D)^2}{(A+D)^2 - 4B^2} \in (0,1), \quad \text{and} \quad g = \frac{1-q}{2}\frac{A-D}{A+B}(1 - \|\boldsymbol{w}\|_2^2) > 0. \tag{3.3}$$

To isolate the mechanism in a form that admits an exact dynamical analysis, we first consider a canonical balanced reference direction,

$$\boldsymbol{z}_0 = \frac{\mathbf{1} - (1+a)\boldsymbol{e}_{j_\star}}{\sqrt{a^2 + r - 1}}. \tag{3.4}$$

Starting from this direction, the Muon and GD dynamics can be compared exactly: each Muon update retains the fraction q of its current loss, whereas each GD update retains at least $q + g$.

Proposition 3.2. *Suppose Assumption 3.1 holds. Apply Muon and GD with exact line search to Q, starting at $\boldsymbol{z}_0$ in (3.4). Then $Q(\boldsymbol{z}_k^{\mathrm{M}})$ and $Q(\boldsymbol{z}_k^{\mathrm{G}})$ are positive for all $k \geq 0$,*

$$\frac{Q(\boldsymbol{z}_{k+1}^{\mathrm{M}})}{Q(\boldsymbol{z}_k^{\mathrm{M}})} = q, \quad \textit{and} \quad \frac{Q(\boldsymbol{z}_{k+1}^{\mathrm{G}})}{Q(\boldsymbol{z}_k^{\mathrm{G}})} - q \geq g, \quad \textit{for every} \quad k \geq 0. \tag{3.5}$$

The proof is given in Appendix B.3. Since $r \geq 3$, positivity and normalization of $\boldsymbol{w}$ imply $1 - \|\boldsymbol{w}\|_2^2 > 0$ and hence $g > 0$. The choice of a balances the distinguished component with the aggregate tail. Accordingly, define $\boldsymbol{Z}_{\mathrm{bal}} = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{z}_0)\boldsymbol{V}^\top$ as the corresponding reference direction. Section 4 further shows that every sufficiently nearby initial direction retains a loss-ratio gap of at least $g/2$.

4 Main Results

Proposition 3.2 gives a loss-ratio gap g for the reference quadratic loss. We now establish a gap of at least $g/2$ along the Muon and GD trajectories on one fixed empirical objective.

4.1 Finite-step Muon–GD advantage

We now transfer the quadratic separation in Proposition 3.2 to the empirical objective. Since the quadratic comparison is centered at $\boldsymbol{Z}_{\mathrm{bal}}$, we consider an initialization satisfying

$$0 < \|\boldsymbol{W}_0 - \boldsymbol{W}_\star\|_{\mathrm{F}} \leq \varepsilon \min_j \sigma_j, \qquad \left\| \frac{\boldsymbol{W}_0 - \boldsymbol{W}_\star}{\|\boldsymbol{W}_0 - \boldsymbol{W}_\star\|_{\mathrm{F}}} - \boldsymbol{Z}_{\mathrm{bal}} \right\|_{\mathrm{F}} \leq \varepsilon. \tag{4.1}$$

Indeed, a strict comparison between two optimization dynamics requires control of the direction of the initial perturbation. The reference direction $\boldsymbol{Z}_{\mathrm{bal}}$ isolates the effect of curvature anisotropy on the

two dynamics, while (4.1) permits perturbations around this direction. The two conditions control the local approximation error and the deviation from the balanced reference direction, respectively. We define data model induced conditioning as

$$\kappa_{\mathrm{act}} = \frac{1+\mu^2/(\gamma\rho^{j_\star-1})}{(1-\rho)^2}\,\frac{\max_{1\le j\le r}\sigma_j}{\rho^{r-1}\min_{1\le j\le r}\sigma_j}. \tag{4.2}$$

The following theorem extends the quadratic loss-ratio gap to the empirical trajectories over a finite number of iterations.

Theorem 4.1. *Suppose the observations follow the input and response model in Section 2.1 and Assumption 3.1 holds. There exist universal constants $C_0 > 0$ and $C_1 > 0$ satisfying the following property. Fix $0 < \varepsilon < 1/4$ and $0 < \zeta < 1$, and suppose $N \ge C_0 dr \log^3 \frac{dr}{\zeta}$. Then, with probability at least $1-\zeta$, the following holds simultaneously for every common initialization $\boldsymbol{W}_0$ satisfying (4.1) and every integer $T \ge 2$ satisfying*

$$T \le \frac{\log\bigl(1/(\varepsilon+\varepsilon_N+\delta)\bigr)}{C_1\log(\kappa_{\mathrm{act}}/\eta)}, \tag{4.3}$$

where $\varepsilon_N := \sqrt{\frac{dr\log^3(dr/\zeta)}{N}}$, $\eta := \frac{A-D}{A+D} \in (0,1)$. For $\mathrm{A} \in \{\mathrm{M},\mathrm{G}\}$, the steps α_k^{A} exist for $0 \le k < T$ and $\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{A}}) > 0$ for $0 \le k \le T$. Moreover,

$$\frac{\mathcal{L}_N(\boldsymbol{W}_{k+1}^{\mathrm{G}})}{\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{G}})} - \frac{\mathcal{L}_N(\boldsymbol{W}_{k+1}^{\mathrm{M}})}{\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{M}})} \ge \frac{g}{2} > 0, \qquad 0 \le k < T. \tag{4.4}$$

Consequently,

$$\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{M}}) \le (1-g/2)^k \mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{G}}), \qquad 0 \le k \le T. \tag{4.5}$$

The proof is given in Appendix B.6. In particular, the comparison interval is nonempty whenever $\varepsilon + \varepsilon_N + \delta \le (\eta/\kappa_{\mathrm{act}})^{2C_1}$. The theorem compares losses along the two methods' respective empirical trajectories. Its event is uniform over the stated initialization neighborhood, including initial matrices computed from the same observations. The next estimate quantifies the deviation from the quadratic trajectories.

Remark 4.2. Under the conditions of Theorem 4.1, let $\boldsymbol{z}_k^{\mathrm{A}}$ be the quadratic iterates in Proposition 3.2. For $\mathrm{A} \in \{\mathrm{M},\mathrm{G}\}$ and $0 \le k \le T$,

$$\frac{\|\boldsymbol{W}_k^{\mathrm{A}} - \boldsymbol{W}_\star - \|\boldsymbol{W}_0 - \boldsymbol{W}_\star\|_{\mathrm{F}}\boldsymbol{U}\,\mathrm{diag}(\boldsymbol{z}_k^{\mathrm{A}})\boldsymbol{V}^\top\|_{\mathrm{F}}}{\|\boldsymbol{W}_0 - \boldsymbol{W}_\star\|_{\mathrm{F}}\|\boldsymbol{z}_k^{\mathrm{A}}\|_2} \le \left(\frac{\kappa_{\mathrm{act}}}{\eta}\right)^{C(k+1)}(\varepsilon+\varepsilon_N+\delta),$$

where $C > 0$ is universal. This relative error includes the component orthogonal to $\mathcal{S}$. Within (4.3), the change in each successive loss ratio is at most $g/4$, which leaves the gap $g/2$ in (4.4).

4.2 Extension to SignGD

We next compare spectral normalization with the entrywise direction $\mathcal{D}_{\mathrm{S}}(\boldsymbol{G}) = \mathrm{sign}(\boldsymbol{G})$, using the same line-search rule. Let $m \ge 2$ be an integer, set $d = m + r - 1$, and suppose r admits an orthogonal matrix with entries of magnitude $r^{-1/2}$. Appendix B.7 constructs coordinates with $j_\star = 1$ in which SignGD assigns relative weight $\sqrt{m}$ to the mean-associated singular direction and weight one to each remaining direction. The following theorem gives the comparison in these coordinates.

Theorem 4.3. *Suppose the conditions of Theorem 4.1 hold with $j_\star = 1$, and use the coordinate realization (B.35) in Appendix B.7. On the same sample event of probability at least $1-\zeta$, for every integer $T \ge 2$ satisfying*

$$T \le \frac{\log\bigl(1/[mr(\varepsilon+\varepsilon_N+\delta)]\bigr)}{C_1\log(\kappa_{\mathrm{act}}/\eta)}, \tag{4.6}$$

the iterates satisfy

$$\frac{\mathcal{L}_N(\boldsymbol{W}_{k+1}^{\mathrm{M}})}{\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{M}})} < \frac{\mathcal{L}_N(\boldsymbol{W}_{k+1}^{\mathrm{S}})}{\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{S}})}, \qquad 0 \le k < T. \tag{4.7}$$

Consequently $\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{M}}) < \mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{S}})$ for $1 \le k \le T$.

The proof is given in Appendix B.7. In this construction, the additional weight increases the quadratic loss ratio. The factor mr in (4.6) enters through the lower bound for the smallest gradient entry, which controls preservation of the entrywise signs.

5 EXPERIMENTS

Section 4 links data-induced curvature to a sustained loss-ratio advantage through a local analysis of nonlinear trajectories. To connect this mechanism to language-model training, we examine the parameter and representation geometry of NanoGPT, then compare the optimization progress of Muon and AdamW.

5.1 EXPERIMENTAL SETUP

Following the common-initialization comparison in Section 4, we pair runs by initialization and token order. We train the 124M-parameter NanoGPT of Wang et al. [35] on FineWeb-10B [29] for 10,000 steps, with warmup and cosine decay for both optimizers. Muon updates attention and FFN matrices, with auxiliary AdamW elsewhere; the baseline uses AdamW throughout. Geometric measurements use one Muon trajectory and fixed validation data; optimizer comparisons use three paired seeds (Appendix A).

5.2 GEOMETRIC DIAGNOSTICS

Approach to a fixed parameter reference. Since our finite-step analysis is local, we first examine whether training exhibits a prolonged local phase near a fixed parameter reference. For each up and down projection, $\boldsymbol{W}_t$ denotes the weight after step t, with $\boldsymbol{W}_\star = \boldsymbol{W}_{10k}$ as an empirical reference. Across layers, Figure 1 shows the distance $\|\boldsymbol{W}_t - \boldsymbol{W}_\star\|_F$ decreasing rapidly at first and then more gradually. This trend gives a qualitative motivation for the local analysis, with the scale-normalized proximity test detailed in Appendix A.2.

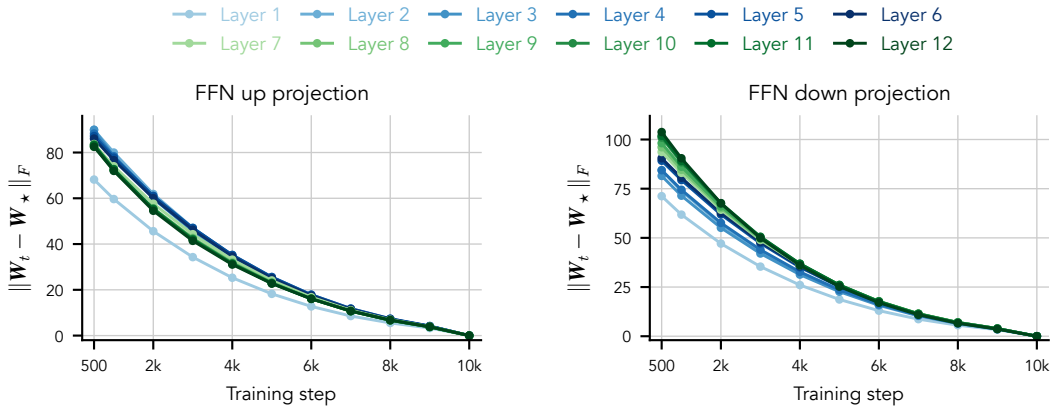

Figure 1: FFN weights approach a fixed reference during training. $\boldsymbol{W}_t$ is the weight after step t, and $\boldsymbol{W}_\star$ is its 10k value. Each curve represents one Transformer layer; the terminal distance is zero by construction.

Spectral concentration of attention outputs. The curvature separation in Theorem 4.1 depends on how the input mean interacts with anisotropic fluctuations. In the data model, these contributions are controlled by μ and the covariance spectrum $\gamma\rho^{j-1}$, which also enter the conditioning in (4.2). We therefore compare centered and uncentered spectra to distinguish the mean's rank-one contribution from variation around it. Figure 2 compares $\mathrm{Cov}(\boldsymbol{h})$ with $\mathbb{E}[\boldsymbol{h}\boldsymbol{h}^\top]$, where $\boldsymbol{h}$ is the attention output after projection and before residual addition and the next LayerNorm. Centering substantially reduces the dominant spectral component, while leaving an anisotropic spectrum. This combination provides an empirical motivation for the mean-plus-fluctuation structure underlying the curvature analysis. Appendix A.3 gives the full spectra and their relation to the model assumptions.

Mean-direction concentration in fitted quadratic responses. The mean's orientation enters the trajectory comparison through its relation to the active response directions; see Remark 4.2. To examine whether a corresponding directional concentration appears in trained networks, we fit a scalar FFN output at 10k by $\boldsymbol{x}^\top \widetilde{\boldsymbol{W}}_c \widetilde{\boldsymbol{W}}_c^\top \boldsymbol{x}$, where $\boldsymbol{x}$ is the FFN input after the second LayerNorm.

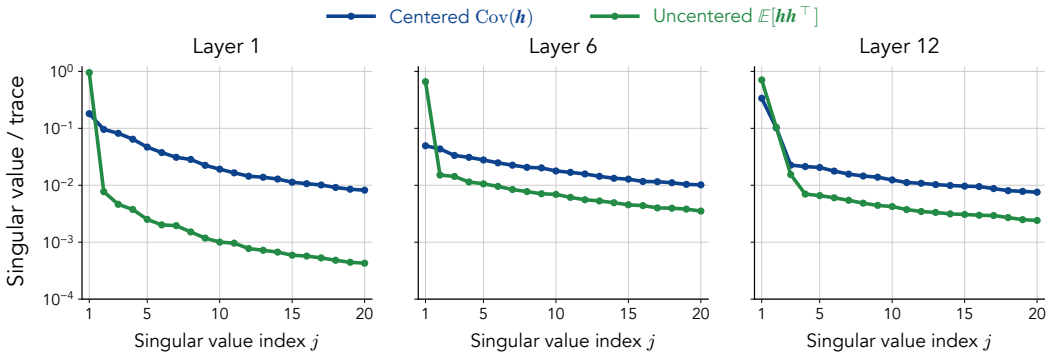

Figure 2: Centering reduces the dominant attention-output component. Spectra are measured in Layers 1, 6, and 12 at 10k with Muon and cosine decay, after the attention output projection and before residual addition. Blue denotes $\mathrm{Cov}(\boldsymbol{h})$ and green denotes $\mathbb{E}[\boldsymbol{h}\boldsymbol{h}^\top]$; the first 20 singular values are ordered decreasingly and normalized by the corresponding full trace.

The fitted left singular vectors provide empirical response directions for this comparison, with $\widetilde{\boldsymbol{W}}_c$ kept distinct from the actual FFN weights and the theoretical target.

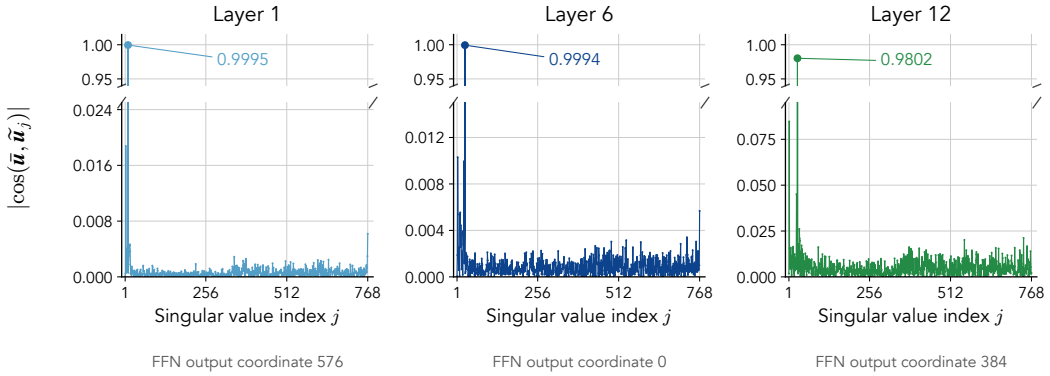

Figure 3: Alignment of the FFN-input mean $\bar{\boldsymbol{u}}$ with fitted directions $\widetilde{\boldsymbol{u}}_j$ at 10k. In these selected examples, the prominent low-index peaks lie in numerical nullspaces; indices follow decreasing fitted singular values. Broken axes retain both the peaks and smaller values, with separate lower-panel scales.

In the selected examples in Figure 3, the largest absolute cosines occur at small indices in the descending singular-value order. The associated singular values place these peaks in the numerical nullspace, which contains most of the mean's projection energy. Subspace concentration is the invariant observation here, since individual singular vectors can rotate within a degenerate subspace. Reading cosines together with response strengths is therefore essential to interpreting the active-direction alignment used in Section 4. Appendix A.4 quantifies this distinction and the relation between fitted and theoretical directions.

5.3 Validation Loss and Relative Loss Decrease

Section 4 relates a sustained advantage in successive loss ratios to a separation between optimization trajectories, as expressed in (4.4)–(4.5) and extended to SignGD in Theorem 4.3. This motivates examining both terminal loss and progress along the trajectory, using AdamW as a practical coordinatewise-adaptive baseline. Muon reaches a lower validation loss at 10k in all three paired runs. To resolve the intervening progress, we measure $r_t = \mathcal{L}_{\mathrm{val}}(t)/\mathcal{L}_{\mathrm{val}}(t-250)$ on a fixed validation prefix and average the ratios across seeds. A smaller ratio means a larger relative improvement over the preceding 250 updates.

Figure 4 shows comparable loss ratios over most intervals, with AdamW attaining a slightly smaller mean ratio more often. At the same time, Muon exhibits a pronounced advantage over an early finite

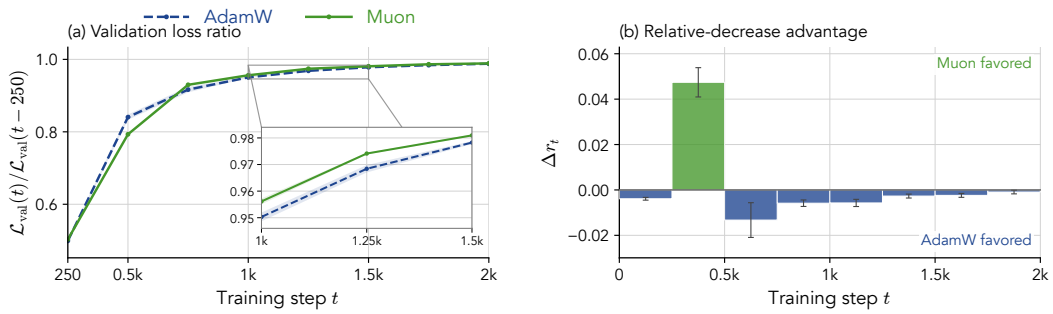

Figure 4: Relative validation loss decrease at 0–2k. (a) Mean loss ratios r_t at interval endpoints, with a lower-right inset enlarging 1–1.5k. (b) $\Delta r_t = r_t^{\text{AdamW}} - r_t^{\text{Muon}}$; each bar spans $[t-250, t]$, with green favoring Muon and blue favoring AdamW. Bands and error bars show one sample standard deviation across three paired seeds.

interval and reaches a lower validation loss at 10k in all paired runs. This behavior is consistent with the finite-step phenomenon highlighted by our theory. The experiment uses scheduled stochastic updates and AdamW, whereas the theoretical comparison concerns line-search updates with GD and SignGD on the quadratic-response training loss.

6 Conclusion

We identify a data-induced curvature regime in which spectral normalization yields a provable optimization advantage. The induced gradient spectral imbalance leads to a strict loss-ratio advantage of exact-polar Muon over GD and, under an explicit coordinate realization, SignGD. This separation persists along finite nonlinear trajectories on one fixed empirical objective, while corresponding geometric patterns are observed in trained language models. These results establish a provable link between data geometry, parameter curvature, and the advantage of spectral normalization.

AI use statement

In this work, we used generative AI tools for language polishing, literature discovery, drafting and revising portions of the manuscript, and for assisting with parts of the mathematical proof. We did not use generative AI tools to originate the research questions, model formulation, research motivation, or theoretical mechanisms, which were developed by the authors. We reviewed all AI-assisted work and independently verified the mathematical arguments included in the final manuscript. We take responsibility for the final content of this work, including text, claims, proofs, or other artifacts produced with the aid of generative AI.

References

[1] 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.

[2] Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 404–413, 2018.

[3] Jeremy Bernstein and Laker Newhouse. Old Optimizer, New Norm: An Anthology. In *OPT 2024: Optimization for Machine Learning*, 2024.

[4] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 560–569, 2018.

[5] Fengmiao Bian, Jian-Feng Cai, and Rui Zhang. A preconditioned riemannian gradient descent algorithm for low-rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 45(4):2075–2103, 2024. doi: 10.1137/23M1570442.

[6] Hanqin Cai, Jian-Feng Cai, and Juntao You. Structured gradient descent for fast robust low-rank hankel matrix completion. *SIAM Journal on Scientific Computing*, 45(3):A1172–A1198, 2023. doi: 10.1137/22M1491009.

[7] Jian-Feng Cai, Tianming Wang, and Ke Wei. Spectral compressed sensing via projected gradient descent. *SIAM Journal on Optimization*, 28(3):2625–2653, 2018. doi: 10.1137/17M1141394.

[8] Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

[9] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 55–65. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1006.

[10] Nathan Godey, Éric Clergerie, and Benoît Sagot. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35–48. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-long.3.

[11] Svante Janson. *Gaussian Hilbert Spaces*, volume 129 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1997. doi: 10.1017/CBO9780511526169.

[12] Nandan Kumar Jha and Brandon Reagen. NerVE: Nonlinear eigenspectrum dynamics in LLM feed-forward networks. In *International Conference on Learning Representations*, 2026.

[13] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

[14] Juno Kim, Eshaan Nichani, Denny Wu, Alberto Bietti, and Jason D Lee. Sharp capacity scaling of spectral optimizers in learning associative memory. *arXiv preprint arXiv:2603.26554*, 2026.

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

[16] Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

[17] Tim Tsz-Kit Lau, Qi Long, and Weijie Su. Polargrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective. *arXiv preprint arXiv:2505.21799*, 2025.

[18] Binghui Li, Kaifei Wang, Han Zhong, Pinyan Lu, and Liwei Wang. Muon in associative memory learning: Training dynamics and scaling laws. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

[19] Melody Li, Kumar Krishna Agrawal, Arna Ghosh, Komal Teru, Adam Santoro, Guillaume Lajoie, and Blake A. Richards. Tracing the representation geometry of language models from pretraining to post-training. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1829.

[20] Yuanxin Li, Cong Ma, Yuxin Chen, and Yuejie Chi. Nonconvex matrix factorization from rank-one measurements. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 1496–1505, 2019.

[21] Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. NorMuon: Making Muon more efficient and scalable. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

[22] 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.

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

[24] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 3345–3354, 2018.

[25] Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in Muon, 2026.

[26] Lester Mackey, Michael I. Jordan, Richard Y. Chen, Brendan Farrell, and Joel A. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42 (3):906–945, 2014. doi: 10.1214/13-AOP892.

[27] Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018.

[28] Elliot Paquette, Noah Marshall, Lucas Benigni, Guangyuan Wang, Atish Agarwala, and Courtney Paquette. Phases of Muon: When Muon eclipses SignSGD, 2026.

[29] Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. In *Advances in Neural Information Processing Systems*, volume 37, pp. 30811–30849, 2024. doi: 10.52202/079017-0970.

[30] Tianyu Ruan, Fengzhuo Zhang, Shuche Wang, and Shihua Zhang. Muon learns more robust and transferable features than adam. *arXiv preprint arXiv:2606.09658*, 2026.

[31] Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of muon. *arXiv preprint arXiv:2505.23737*, 2025.

[32] Zakhar Shumaylov, Nathaël Da Costa, Peter Zaika, Bálint Mucsányi, Alex Massucco, Yoav Gelberg, Carola-Bibiane Schönlieb, Yarin Gal, and Philipp Hennig. Muon is not that special: Random or inverted spectra work just as well. *arXiv preprint arXiv:2605.11181*, 2026.

[33] Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2230.

[34] Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How Muon's spectral design benefits generalization: A study on imbalanced data. In *International Conference on Learning Representations*, 2026.

[35] Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Dirk Bergemann, and Zhuoran Yang. Why Muon outperforms Adam: A curvature perspective, 2026.

[36] Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Y. F. Tan. Muon outperforms Adam in tail-end associative memory learning. In *International Conference on Learning Representations*, 2026.

[37] Zihan Wang, Chengyu Dong, and Jingbo Shang. "average" approximates "first principal component"? an empirical analysis on representations from neural language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5594–5603. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.453.

[38] Ke Wei, Jian-Feng Cai, Tony F. Chan, and Shingyu Leung. Guarantees of riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37 (3):1198–1222, 2016. doi: 10.1137/15M1050525.

[39] Ke Liang Xiao, Noah Marshall, Atish Agarwala, and Elliot Paquette. Exact risk curves of signSGD in high dimensions: Quantifying preconditioning and noise-compression effects. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 68391–68439, 2025.

[40] Shuo Xie and Zhiyuan Li. Implicit bias of AdamW: ℓ_∞-norm constrained optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 54488–54510, 2024.

APPENDIX

A Geometry and optimization during training **13**

A.1 Geometric diagnostics follow the same training trajectory 13

A.2 FFN weight distances decrease along a convex profile 13

A.3 Centering reduces the dominant attention-output component 14

A.4 Mean alignment is strongest in numerical nullspaces 16

A.5 Lower final loss need not imply larger relative reductions 17

B Proofs **17**

B.1 Proof of Proposition 2.1 17

B.2 Proof of Proposition 3.1 18

B.3 Proof of Proposition 3.2 19

B.4 Mean-direction perturbation 21

B.5 Empirical curvature estimates 21

B.6 Proof of Theorem 4.1 24

B.7 Proof of Theorem 4.3 30

A GEOMETRY AND OPTIMIZATION DURING TRAINING

A.1 GEOMETRIC DIAGNOSTICS FOLLOW THE SAME TRAINING TRAJECTORY

Figures 1–3 use the same Muon trajectory with cosine decay (seed 42). The NanoGPT model has 12 pre-LayerNorm Transformer layers, hidden width 768, FFN width 3,072, and 12 attention heads, with rotary positional embeddings and tied token/output embeddings. Training uses GPT-2-tokenized FineWeb-10B for 10,000 updates, sequence length 1,024, and an effective batch of 512 sequences across four GPUs. Linear and embedding weights use zero-mean Gaussian initialization with standard deviation 0.02; residual output projections use $0.02/\sqrt{24}$. Linear biases are zero, and LayerNorm scales and biases start at one and zero.

Muon updates attention and FFN matrices with momentum 0.95 and five BF16 Newton–Schulz iterations on the Frobenius-normalized Nesterov input. Its peak learning rate is 0.01, without shape-dependent scaling. Auxiliary AdamW updates the remaining parameters with peak rate 0.002, $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, and zero weight decay. Both rates increase linearly for 700 updates and then follow cosine decay to 10% of their peaks at 10k. Training uses FP32 parameter storage, BF16 autocast, and global gradient clipping at norm one. The geometric diagnostics describe this single trajectory; layers and token data groups are not independent training replications.

A.2 FFN WEIGHT DISTANCES DECREASE ALONG A CONVEX PROFILE

For each layer, we extract the up projection of size 3072×768 and down projection of size 768×3072. We use $\boldsymbol{W}_\star = \boldsymbol{W}_{10k}$ as the empirical reference and compute $\|\boldsymbol{W}_t - \boldsymbol{W}_\star\|_F$ in FP64 from the stored FP32 weights at $t \in \{500, 1000, 2000, \ldots, 10000\}$. All distances use the original parameter coordinates, without orthogonal alignment or a fitted quadratic factor. Transposing the wide down projection to the theory's tall-factor convention preserves these distances and the thin singular values.

For all 24 curves, secant slopes between successive nonterminal checkpoints are negative and become less negative. The resulting piecewise-linear interpolants are decreasing and convex on this saved grid. The terminal distance is zero by construction and breaks the grid's convexity when included.

The theoretical initialization requires both proximity on the smallest parameter scale and a normalized displacement close to the balanced direction. If the terminal weight is used as a candidate target, a necessary proximity test is

$$R(t, 10k) = \frac{\|\boldsymbol{W}_t - \boldsymbol{W}_{10k}\|_F}{\sigma_{\min}(\boldsymbol{W}_{10k})} < \frac{1}{4}, \qquad t < 10000, \tag{A.1}$$

where $\sigma_{\min}$ is the smallest of the 768 thin singular values. Over the 240 nonterminal layer–projection–checkpoint combinations in Figure 1, the minimum ratio is 3.570684, attained by Layer 1's up projection at 9k. Consequently, none of these candidates satisfies the necessary proximity bound under this reference identification, irrespective of the balanced-direction condition. The $\boldsymbol{W}_\star$ notation in Figure 1 denotes this empirical proxy and does not establish that the terminal weight is the exact theoretical target.

A.3 Centering reduces the dominant attention-output component

Activation and sampling. For an incoming residual $\boldsymbol{z}$, the pre-LayerNorm block computes

$$\boldsymbol{h} = \mathrm{Attn}(\mathrm{LN}_1(\boldsymbol{z})), \qquad \boldsymbol{z}' = \boldsymbol{z} + \boldsymbol{h}, \qquad \boldsymbol{x} = \mathrm{LN}_2(\boldsymbol{z}'), \qquad \boldsymbol{z}'' = \boldsymbol{z}' + \mathrm{FFN}(\boldsymbol{x}). \tag{A.2}$$

Figure 2 measures $\boldsymbol{h}$ in Layers 1, 6, and 12 at 10k, after the attention output projection and before residual addition and the second LayerNorm; Figure 3 uses the FFN input $\boldsymbol{x}$. We sample 512 uniformly spaced positions from each of the 64 fitting sequences in Appendix A.4, giving $N = 32768$ validation vectors, with identical positions across layers and moments.

Estimation. For the empirical distribution of full token vectors, $\bar{\boldsymbol{h}} = \mathbb{E}\boldsymbol{h} = N^{-1}\sum_i \boldsymbol{h}_i$; micro-batches accumulate all cross-coordinate products:

$$\mathrm{Cov}(\boldsymbol{h}) = \frac{1}{N}\sum_i (\boldsymbol{h}_i - \bar{\boldsymbol{h}})(\boldsymbol{h}_i - \bar{\boldsymbol{h}})^\top, \qquad \mathbb{E}[\boldsymbol{h}\boldsymbol{h}^\top] = \mathrm{Cov}(\boldsymbol{h}) + \bar{\boldsymbol{h}}\bar{\boldsymbol{h}}^\top. \tag{A.3}$$

Forward passes use FP32 on CPU without autocast; moment accumulation and eigendecomposition use FP64. The replay uses the implementation's explicit causal-attention path, and its FFN inputs are checked against the original activation cache. We decompose each full 768×768 matrix and plot its first 20 eigenvalues divided by the sum of all 768 eigenvalues. We do not renormalize by the sum of the displayed 20 values, take square roots, or smooth the spectra. The two descending spectra need not share eigenvectors; their separate trace normalization compares spectral concentration rather than absolute variance. Table 1 summarizes the leading fractions and the first-to-second ratios. These second-moment diagnostics describe anisotropy and the mean contribution. Gaussianity, exact geometric covariance decay, and the response-dependent fourth-moment curvature separation are additional requirements of the theoretical model; spectral concentration alone does not establish them.

Table 1: Attention-output spectral concentration at 10k for Figure 2. Eigenvalues refer to each column's moment matrix; sums use all 768 values.

Layer	$\mathrm{Cov}(\boldsymbol{h})$ $\lambda_1/\sum_j \lambda_j$	$\mathbb{E}[\boldsymbol{h}\boldsymbol{h}^\top]$ $\lambda_1/\sum_j \lambda_j$	$\mathrm{Cov}(\boldsymbol{h})$ λ_1/λ_2	$\mathbb{E}[\boldsymbol{h}\boldsymbol{h}^\top]$ λ_1/λ_2
1	0.1810	0.9530	1.89	123.48
6	0.0494	0.6602	1.14	43.65
12	0.3375	0.7077	3.30	6.77

Complete spectra across layers and training. Figure 5 shows all 768 singular values in every Transformer layer at 0k, 5k, and 10k. Each panel overlays six curves combining the centered covariance, uncentered second moment, and three checkpoints. We use the same 64 validation sequences and token positions at each checkpoint and normalize each matrix by its own full trace.

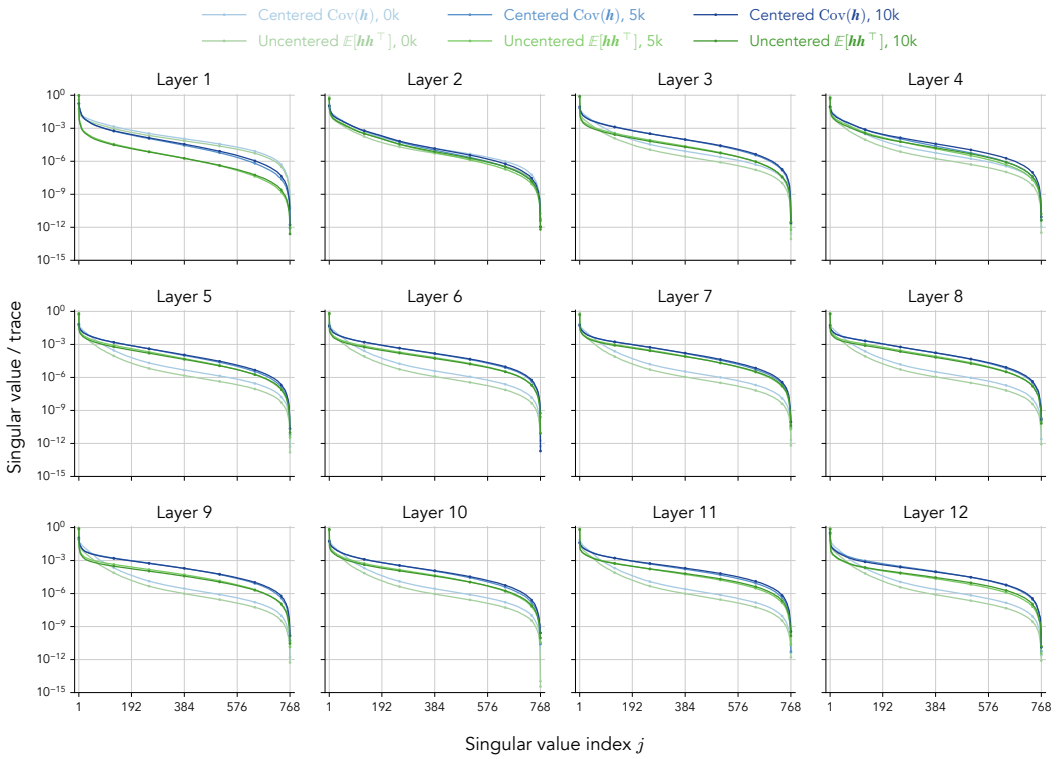

Figure 5: Full attention-output spectra at 0k, 5k, and 10k, supplementing Figure 2. Each panel shows one layer: $\mathrm{Cov}(\boldsymbol{h})$ (blue) and $\mathbb{E}[\boldsymbol{h}\boldsymbol{h}^\top]$ (green), with light, medium, and dark shades denoting 0k, 5k, and 10k, respectively. All 768 singular values are divided by the corresponding full trace. The horizontal axis is linear and the vertical axis is logarithmic.

A.4 MEAN ALIGNMENT IS STRONGEST IN NUMERICAL NULLSPACES

The 10k quadratic surrogate. For a scalar FFN output coordinate p, we fit the 10k response with

$$\widehat{\boldsymbol{Z}}_p = \arg\min_{\boldsymbol{Z}\succeq 0} \frac{1}{N_{\text{fit}}} \sum_i \left[(\text{FFN}_{10k}(\boldsymbol{x}_i))_p - \boldsymbol{x}_i^\top \boldsymbol{Z} \boldsymbol{x}_i\right]^2, \qquad \widetilde{\boldsymbol{W}}_c \widetilde{\boldsymbol{W}}_c^\top = \widehat{\boldsymbol{Z}}_p. \tag{A.4}$$

The fit uses all $d = 768$ input dimensions and allows every PSD quadratic form, without an additional rank constraint. The fitted factor can be rank deficient; its width is not identified with the theory's active rank r. The fitted factor $\widetilde{\boldsymbol{W}}_c$ is distinct from the trained reference $\boldsymbol{W}_\star = \boldsymbol{W}_{10k}$ used in Figure 1.

We use 128 disjoint validation sequences at 10k, retaining 512 positions per sequence. Sets of 32, 64, and 32 sequences respectively construct a complete orthonormal input basis, fit the response on 32,768 vectors, and evaluate predictions on 16,384 vectors. The basis comes from the uncentered input moment; the full-width fit is equivalent to the PSD problem above. FP64 projected optimization uses an invertible conditioning transform, relative projected-gradient tolerance 10^{-7}, and at most 2,000 iterations. Predictions are restored to original units, with no intercept, linear term, target shift, or positive spectral floor. Fitting errors for the three displayed surrogate factors appear in Table 2. A PSD quadratic predictor is nonnegative and cannot exactly reproduce signed FFN outputs; the fitted surrogate is therefore an approximation, even when its optimization converges.

Table 2: Approximation quality of the three 10k surrogate factors in Figure 3. All three fits meet the optimization stopping criterion. MSE and R^2 use 16,384 held-out token vectors that also participate in alignment-based selection.

Layer	Output p	Numerical rank	Held-out MSE	Held-out R^2
1	576	2	2.718	-0.756
6	0	8	0.887	0.108
12	384	22	4.771	0.451

Cosines and display selection. On the pooled 32,768-token fitting data, define the normalized mean $\bar{\boldsymbol{u}} = \bar{\boldsymbol{x}}/\left\|\bar{\boldsymbol{x}}\right\|_2$ and write

$$\widetilde{\boldsymbol{W}}_c = \widetilde{\boldsymbol{U}} \,\text{diag}(\widetilde{\sigma}_j) \widetilde{\boldsymbol{V}}^\top, \qquad c_j = \left|\bar{\boldsymbol{u}}^\top \widetilde{\boldsymbol{u}}_j\right|. \tag{A.5}$$

The horizontal order follows decreasing $\widetilde{\sigma}_j$, and c_j measures alignment of the mean, rather than an average of tokenwise cosines. The theory instead orders its shared input–response directions $\boldsymbol{u}_j$ by centered variance and selects $j_\star \in \arg\max_{1\le j\le r} |\bar{\boldsymbol{u}}^\top \boldsymbol{u}_j|$ among active directions. The tilde distinguishes the empirical surrogate basis from that assumed common basis.

We inspect 36 fitted surrogate factors: output coordinates $0, 64, \ldots, 704$ in each of Layers 1, 6, and 12. Of these, 28 are converged and nonzero. For display, we require at least one cosine ≥ 0.1 and every such cosine within the first 76 indices, on both fitting and held-out data. Among the converged nonzero fits, this post hoc rule selects 2 of 4, 6 of 12, and 1 of 12 in the three layers, respectively. We retain Layer 6's coordinate 0 and select the smallest qualifying coordinate in the other layers, giving 576 and 384. The panels therefore use different output coordinates and are not a controlled depth comparison. Held-out data participate in selection, so this plot is not an independent validation of the selected pattern.

Resolved directions and numerical nullspaces. A factor direction is resolved when $\widetilde{\sigma}_j > 10^{-5}\widetilde{\sigma}_{\max}$. The selected factors have numerical ranks 2, 8, and 22. Their peak cosines are 0.9995, 0.9994, and 0.9802, but the corresponding relative singular values are 9.56×10^{-9}, 7.18×10^{-11}, and 5.69×10^{-10}. The largest resolved-direction cosines are only 0.000487, 0.01030, and 0.08473. Thus a low index in these plots does not imply a substantial response singular value.

The numerical nullspaces contain 99.99998%, 99.9845%, and 99.1118% of the normalized means' squared projection energy. Unlike individual singular vectors in a numerically degenerate subspace, this aggregate projection is invariant to rotations within that subspace. Figure 3 retains all 768 cosines per panel; separate lower-panel scales expose small values and no observations lie in the

omitted intervals. The large peaks demonstrate concentrated mean projections in these selected fits, but do not establish alignment with resolved active features or identify exact nullspaces of the original FFN.

A.5 Lower final loss need not imply larger relative reductions

Figure 4 uses the three complete paired runs with seeds 42, 43, and 44. Both optimizers use the architecture, batch size, and training budget described above, with the same warmup duration and cosine decay factor. The full-model AdamW baseline uses the auxiliary AdamW hyperparameters, including peak learning rate 0.002; Muon uses peak rate 0.01 for matrix parameters. Initializations and ordered token data are matched by seed, although the optimizer cohorts use different hardware. Every 250 updates, we evaluate cross entropy on the same 2,621,440-token validation prefix. Ratios are computed separately per seed and summarized by their mean and sample standard deviation, without smoothing or subtracting an estimated loss floor. Both panels display the eight intervals from 0→250 through 1750→2000; panel (a) positions ratios at their ending steps, while panel (b) uses bars spanning each interval. The inset uses interval endpoints 1000, 1250, and 1500. Muon has the smaller mean ratio in one of these eight intervals and 9 of 40 intervals over the full 10k trajectory. The final validation losses are 3.275895 ± 0.001739 for Muon and 3.303727 ± 0.001564 for AdamW (mean $\pm$ sample standard deviation). Muon has lower final loss in each pair; the mean difference is 0.027832. These are raw cross-entropy ratios, not ratios of excess loss above a known optimum.

B Proofs

We prove the results in the order in which they are used. The population calculations give the quadratic comparison, the Gaussian moment estimates control its empirical perturbation, and the final two proofs compare the corresponding iterates.

Throughout the appendix, $C > 0$ denotes a universal numerical constant whose value may change from line to line.

B.1 Proof of Proposition 2.1

The population loss $\mathcal{L}(\boldsymbol{W}) = \frac{1}{2}\mathbb{E}(f_{\boldsymbol{W}}(\boldsymbol{x}) - \boldsymbol{x}^\top \boldsymbol{X}_\star \boldsymbol{x})^2$ has Hessian at $\boldsymbol{W}_\star$

$$\mathcal{H}_{\mathrm{pop}}[\boldsymbol{D}] = 4\mathbb{E}\big[(\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{W}_\star^\top \boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{W}_\star\big]. \tag{B.1}$$

For the proofs, couple the actual input with the mean-aligned reference

$$\boldsymbol{x}^{(0)} = \boldsymbol{x} + \mu(\boldsymbol{u}_{j_\star} - \bar{\boldsymbol{u}}), \qquad \|\boldsymbol{x} - \boldsymbol{x}^{(0)}\|_2 = \mu\delta. \tag{B.2}$$

Then $\boldsymbol{x}^{(0)}$ has the same centered covariance as $\boldsymbol{x}$ and mean $\mu\boldsymbol{u}_{j_\star}$, so

$$\mathbb{E}[\boldsymbol{x}^{(0)}\boldsymbol{x}^{(0)\top}] = \mathrm{Cov}(\boldsymbol{x}) + \mu^2 \boldsymbol{u}_{j_\star}\boldsymbol{u}_{j_\star}^\top. \tag{B.3}$$

We first record two bounds for the aligned Hessian that will also be used in the empirical comparison.

Lemma B.1. *Suppose that the input model in Section 2.1 holds. Then*

$$\|\mathcal{H}_0\|_{\mathrm{op}} \le 96\gamma^2(\max_j \sigma_j)^2 \left(\frac{\mu^2}{\gamma} + \frac{1}{1-\rho}\right)^2. \tag{B.4}$$

Moreover,

$$\lambda_{\min}(\boldsymbol{K}) \ge 8\gamma^2(\min_j \sigma_j)^2 \rho^{2r-2}. \tag{B.5}$$

The second bound implies $\boldsymbol{K} \succ 0$, while the first controls the full Hessian on $\mathbb{R}^{d\times r}$.

Proof. Write $\boldsymbol{x}^{(0)} = \mu\boldsymbol{u}_{j_\star} + \boldsymbol{y}$, where $\boldsymbol{y}$ is centered Gaussian with covariance $\gamma\boldsymbol{A}_0 \,\mathrm{diag}(1, \rho, \ldots, \rho^{d-1})\boldsymbol{A}_0^\top$. For $\|\boldsymbol{D}\|_{\mathrm{F}} = 1$,

$$\|\mathcal{H}_0[\boldsymbol{D}]\|_{\mathrm{F}} \le 4(\max_j \sigma_j)^2 \mathbb{E}\|\boldsymbol{x}^{(0)}\|_2^4.$$

Moreover,

$$\mathbb{E}\|\boldsymbol{y}\|_2^4 = \gamma^2\left[\left(\sum_{i=1}^{d}\rho^{i-1}\right)^2 + 2\sum_{i=1}^{d}\rho^{2i-2}\right] \le \frac{3\gamma^2}{(1-\rho)^2}.$$

Minkowski's inequality therefore gives

$$\mathbb{E}\|\boldsymbol{x}^{(0)}\|_2^4 \le 24\left(\mu^2 + \frac{\gamma}{1-\rho}\right)^2,$$

which proves (B.4). For $\boldsymbol{z} \in \mathbb{R}^r$, the formula (2.9) gives

$$\begin{aligned}\tfrac{1}{4}\boldsymbol{z}^\top \boldsymbol{K}\boldsymbol{z} = &\left[\sigma_{j_\star}(\mu^2 + \gamma\rho^{j_\star-1})z_{j_\star} + \sum_{j\neq j_\star}\sigma_j\gamma\rho^{j-1}z_j\right]^2 \\ &+ 2\sigma_{j_\star}^2\gamma\rho^{j_\star-1}(2\mu^2 + \gamma\rho^{j_\star-1})z_{j_\star}^2 + 2\sum_{j\neq j_\star}\gamma^2\sigma_j^2\rho^{2j-2}z_j^2.\end{aligned}$$

Dropping the first term and using $j_\star \le r$ proves (B.5). □

Proof of Proposition 2.1. For the aligned input, the coordinates $\boldsymbol{u}_j^\top \boldsymbol{x}^{(0)}$ are independent Gaussian variables, with mean μ for $j = j_\star$ and zero mean otherwise, and variance $\gamma\rho^{j-1}$. Hence

$$\begin{aligned}&\mathbb{E}(\boldsymbol{u}_{j_\star}^\top \boldsymbol{x}^{(0)})^2 = \mu^2 + \gamma\rho^{j_\star-1}, \quad \mathbb{E}(\boldsymbol{u}_{j_\star}^\top \boldsymbol{x}^{(0)})^4 = \mu^4 + 6\mu^2\gamma\rho^{j_\star-1} + 3\gamma^2\rho^{2j_\star-2},\\ &\mathbb{E}(\boldsymbol{u}_j^\top \boldsymbol{x}^{(0)})^2 = \gamma\rho^{j-1}, \qquad\quad \mathbb{E}(\boldsymbol{u}_j^\top \boldsymbol{x}^{(0)})^4 = 3\gamma^2\rho^{2j-2} \quad (j \neq j_\star).\end{aligned}$$

Independence gives (2.9). To verify invariance, complete $\boldsymbol{u}_j\boldsymbol{v}_j^\top$ to the orthonormal matrix basis $(\boldsymbol{A}_0\boldsymbol{e}_a)\boldsymbol{v}_j^\top$, $1 \le a \le d$, $1 \le j \le r$. The linearized predictions of the added directions are products of two distinct input coordinates. Their inner products with squared coordinates vanish: at least one centered Gaussian coordinate has an odd exponent, and only coordinate $j_\star$ has nonzero mean. Thus the Hessian has zero entries between $\mathcal{S}$ and its orthogonal complement. This proves $\mathcal{H}_0(\mathcal{S}) \subseteq \mathcal{S}$, and (B.5) gives $\boldsymbol{K} \succ 0$. □

B.2 Proof of Proposition 3.1

For the remaining proofs, set

$$S_1 = \sum_{j\neq j_\star}\sigma_j\rho^{j-1}, \qquad S_2 = \sum_{j\neq j_\star}\sigma_j^2\rho^{2j-2}, \qquad p = \sigma_{j_\star}\rho^{j_\star-1}, \qquad c = \frac{\mu^2}{\gamma\rho^{j_\star-1}}.$$

Proposition 2.1 gives $\boldsymbol{K} \succ 0$. We now use the mean condition to compare A with D and to bound the coordinates of $\boldsymbol{K}(\mathbf{1} - (1+a)\boldsymbol{e}_{j_\star})$.

Proof. From (2.9), we have

$$A = 4\gamma^2p^2(c^2+6c+3), \qquad B = 4\gamma^2p(c+1)S_1, \qquad D = 4\gamma^2(S_1^2 + 2S_2).$$

Since $S_2 \ge (\min_{j\neq j_\star}\sigma_j\rho^{j-1})S_1$ and $S_2 \le S_1^2$, Assumption 3.1 gives $pc > 4S_1$. Therefore $p^2c^2 > 16S_1^2 > S_1^2 + 2S_2$, and hence $A > D$. Since $\boldsymbol{K} \succ 0$, we also have $AD > B^2$. It follows that $0 < a < 1$ and

$$[\boldsymbol{K}(\mathbf{1} - (1+a)\boldsymbol{e}_{j_\star})]_{j_\star} = B - Aa = -\frac{AD - B^2}{A+B} < 0.$$

It remains to prove positivity of the remaining coordinates. Substituting $a = (B+D)/(A+B)$ and taking positive parts, we have

$$\begin{aligned}[ap(c+1) - S_1]_+ &= \left[\frac{2p[(c+1)S_2 - p(1+2c)S_1]}{p^2(c^2+6c+3) + p(c+1)S_1}\right]_+ \\ &\le \frac{2(c+1)S_2}{p(c^2+6c+3)} \le \frac{4S_2}{pc} < \min_{j\neq j_\star}\sigma_j\rho^{j-1}.\end{aligned}$$

Here the second inequality uses $c > 1$, and the last uses Assumption 3.1. Hence, for every $j \neq j_\star$,

$$[\boldsymbol{K}(\mathbf{1} - (1+a)\boldsymbol{e}_{j_\star})]_j = 4\gamma^2\sigma_j\rho^{j-1}\left\{S_1 - ap(c+1) + 2\sigma_j\rho^{j-1}\right\} \ge 4\gamma^2\sigma_j^2\rho^{2j-2} > 0.$$

This proves Proposition 3.1. □

B.3 Proof of Proposition 3.2

We first record a monotonicity property of GD with exact line search. Once the Muon loss ratio is computed, this property reduces the comparison of all GD iterates to the comparison at the initial point.

Lemma B.2. *Suppose $\boldsymbol{K} \succ 0$ and GD with exact line search for $\boldsymbol{z}^\top \boldsymbol{K}\boldsymbol{z}/2$ has a positive initial loss ratio. Then its subsequent loss ratios are positive and nondecreasing.*

Proof. Let $\boldsymbol{z}_+$ and $\boldsymbol{z}_{++}$ be the next two GD iterates from $\boldsymbol{z}$. We prove $Q(\boldsymbol{z}_{++})/Q(\boldsymbol{z}_+) \geq Q(\boldsymbol{z}_+)/Q(\boldsymbol{z})$. By homogeneity, assume $\|\boldsymbol{K}^{1/2}\boldsymbol{z}\|_2 = 1$ and set $\boldsymbol{e} = \boldsymbol{K}^{1/2}\boldsymbol{z}$ and $m_j = \boldsymbol{e}^\top \boldsymbol{K}^j \boldsymbol{e}$ for $1 \leq j \leq 4$. Then $Q(\boldsymbol{z}_+)/Q(\boldsymbol{z}) = 1 - m_1^2/m_2$ and $\boldsymbol{e}_+ := \boldsymbol{K}^{1/2}\boldsymbol{z}_+ = (\boldsymbol{I} - m_1\boldsymbol{K}/m_2)\boldsymbol{e}$. Expanding gives

$$\begin{aligned} \|\boldsymbol{e}_+\|_2^2 &= 1 - m_1^2/m_2, \\ \boldsymbol{e}_+^\top \boldsymbol{K}\boldsymbol{e}_+ &= -m_1 + m_1^2 m_3/m_2^2, \\ \boldsymbol{e}_+^\top \boldsymbol{K}^2\boldsymbol{e}_+ &= m_2 - 2m_1m_3/m_2 + m_1^2 m_4/m_2^2. \end{aligned}$$

Moreover,

$$1 - \frac{Q(\boldsymbol{z}_{++})}{Q(\boldsymbol{z}_+)} = \frac{(\boldsymbol{e}_+^\top \boldsymbol{K}\boldsymbol{e}_+)^2}{\|\boldsymbol{e}_+\|_2^2 \boldsymbol{e}_+^\top \boldsymbol{K}^2 \boldsymbol{e}_+}.$$

The assumption $Q(\boldsymbol{z}_+)/Q(\boldsymbol{z}) > 0$ gives $\boldsymbol{e}_+ \neq 0$, so the denominator is positive. Substituting the three identities above and multiplying by the positive denominators, the desired inequality follows from

$$\begin{aligned} &(m_2 - m_1^2)(m_2^3 - 2m_1m_2m_3 + m_1^2m_4) - (m_1m_3 - m_2^2)^2 \\ &= m_1^2 \det \begin{pmatrix} 1 & m_1 & m_2 \\ m_1 & m_2 & m_3 \\ m_2 & m_3 & m_4 \end{pmatrix} \geq 0. \end{aligned}$$

The determinant is nonnegative because the matrix is the Gram matrix of $\boldsymbol{e}, \boldsymbol{K}\boldsymbol{e}, \boldsymbol{K}^2\boldsymbol{e}$. Induction now proves the claim. □

Proof of Proposition 3.2. Permute the active coordinates so that $j_\star$ is first. This simultaneous permutation leaves Q, the exact line-search updates, and the quantities A, B, D, a, q, g unchanged. In the remainder of the proof, we use the first coordinate for the mean-associated mode and write $\boldsymbol{w}$ for the reordered tail weights. Set $\boldsymbol{d}_- = (-1, \mathbf{1})^\top$ and $\boldsymbol{d}_+ = (1, \mathbf{1})^\top$. Write

$$\boldsymbol{K} = \begin{pmatrix} A & \boldsymbol{b}^\top \\ \boldsymbol{b} & \boldsymbol{K}_{\mathrm{T}} \end{pmatrix}, \qquad B = \boldsymbol{b}^\top \mathbf{1}, \quad D = \mathbf{1}^\top \boldsymbol{K}_{\mathrm{T}} \mathbf{1}.$$

Positive definiteness gives $AD > B^2$. Since $A > D$ and $\boldsymbol{b} \geq 0$, we have $0 \leq B < A$ and

$$a - \frac{B}{A} = \frac{AD - B^2}{A(A+B)} > 0, \qquad 1 - a = \frac{A - D}{A + B} > 0.$$

Since both methods with exact line search commute with multiplication of the initial vector by a positive scalar, it suffices to start at $(-a, \mathbf{1})^\top$. Define

$$\boldsymbol{z}_- = (-D - B, (A+B)\mathbf{1})^\top, \qquad \boldsymbol{z}_+ = (D - B, (A-B)\mathbf{1})^\top.$$

We have $\boldsymbol{z}_- = (A+B)(-a, \mathbf{1})^\top$ and

$$[\boldsymbol{K}\boldsymbol{z}_-]_1 = -(AD - B^2), \qquad [\boldsymbol{K}\boldsymbol{z}_+]_1 = AD - B^2.$$

Furthermore, $A \pm B > 0$ and

$$\frac{D - B}{A - B} > -\frac{B}{A} > -a.$$

For $y > 0$ and $x + ay \geq 0$, Proposition 3.1 gives

$$[\boldsymbol{K}(x, y\mathbf{1})^\top]_j = y[\boldsymbol{K}(-a, \mathbf{1})^\top]_j + (x + ay)K_{j1} > 0, \qquad 2 \leq j \leq r. \tag{B.6}$$

Applying this identity to $\boldsymbol{z}_-$ and $\boldsymbol{z}_+$ proves that the Muon directions are $\boldsymbol{d}_-$ and $\boldsymbol{d}_+$, respectively. The formula $\boldsymbol{z} \mapsto \boldsymbol{z} - (\boldsymbol{d}^\top \boldsymbol{K}\boldsymbol{z})/(\boldsymbol{d}^\top \boldsymbol{K}\boldsymbol{d})\,\boldsymbol{d}$ then gives

$$\boldsymbol{z}_- \longmapsto \frac{A-D}{A-2B+D}\boldsymbol{z}_+, \qquad \boldsymbol{z}_+ \longmapsto \frac{A-D}{A+2B+D}\boldsymbol{z}_-. \tag{B.7}$$

Both multiplying factors are positive. Since

$$2Q(\boldsymbol{z}_\pm) = (AD - B^2)(A + D \mp 2B),$$

substitution into (B.7) yields

$$\frac{Q(\boldsymbol{z}_{k+1}^{\mathrm{M}})}{Q(\boldsymbol{z}_k^{\mathrm{M}})} = q, \qquad \boldsymbol{z}_{k+2}^{\mathrm{M}} = q\boldsymbol{z}_k^{\mathrm{M}}, \qquad k \geq 0.$$

In particular, $Q(\boldsymbol{z}_k^{\mathrm{M}}) > 0$ for every $k \geq 0$. To compare GD, set $\boldsymbol{g} = \boldsymbol{K}(-a, \boldsymbol{1})^\top$. Gradient balance gives

$$\boldsymbol{g} = (-g_1)(-1, \boldsymbol{w})^\top, \qquad -g_1 = Aa - B = D - aB = \frac{AD - B^2}{A + B} > 0.$$

By (3.2), $\boldsymbol{w} > 0$ and $\boldsymbol{1}^\top \boldsymbol{w} = 1$. Thus $u = \|\boldsymbol{w}\|_2^2 < 1$ because $r \geq 3$. Multiplying a direction by a positive scalar leaves its line-search update unchanged. We therefore compare $\boldsymbol{d}_-$ with the GD direction $\boldsymbol{d}_{\mathrm{G}} = (-1, \boldsymbol{w})^\top$. The exact line-search formula gives

$$1 - q = \frac{4(-g_1)}{(1+a)\boldsymbol{d}_-^\top \boldsymbol{K}\boldsymbol{d}_-}, \qquad 1 - \frac{Q(\boldsymbol{z}_1^{\mathrm{G}})}{Q(\boldsymbol{z}_0)} = \frac{(-g_1)(1+u)^2}{(1+a)\boldsymbol{d}_{\mathrm{G}}^\top \boldsymbol{K}\boldsymbol{d}_{\mathrm{G}}}.$$

Since $-g_1 = Aa - B = A(a - B/A)$, the Schur complement satisfies $(\boldsymbol{K}_{\mathrm{T}} - \boldsymbol{b}\boldsymbol{b}^\top/A)\boldsymbol{1} = (-g_1)(\boldsymbol{w} + \boldsymbol{b}/A)$. Set $t = (u + \boldsymbol{b}^\top \boldsymbol{w}/A)/(1+u)$. Since $\boldsymbol{b} \geq 0$ and $\boldsymbol{1}^\top \boldsymbol{w} = 1$, we have

$$\frac{1 + B/A}{2} - t = \frac{(1 - B/A)(1-u) + 2(B - \boldsymbol{b}^\top \boldsymbol{w})/A}{2(1+u)} \geq \frac{(1 - B/A)(1-u)}{2(1+u)} > 0.$$

For $t \leq (1 + B/A)/2$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[A(1-t)^2 + \frac{A(a - B/A)}{1 + B/A}t^2\right] = -2A + \frac{2A(1+a)}{1 + B/A}t \leq A(a-1) < 0.$$

At $t = (1 + B/A)/2$, the expression in brackets equals $\boldsymbol{d}_-^\top \boldsymbol{K}\boldsymbol{d}_-/4$. Hence

$$\begin{aligned}
\frac{\boldsymbol{d}_{\mathrm{G}}^\top \boldsymbol{K}\boldsymbol{d}_{\mathrm{G}}}{(1+u)^2} &= A(1-t)^2 + \frac{\boldsymbol{w}^\top(\boldsymbol{K}_{\mathrm{T}} - \boldsymbol{b}\boldsymbol{b}^\top/A)\boldsymbol{w}}{(1+u)^2} \\
&\geq A(1-t)^2 + \frac{[\boldsymbol{1}^\top(\boldsymbol{K}_{\mathrm{T}} - \boldsymbol{b}\boldsymbol{b}^\top/A)\boldsymbol{w}]^2}{(1+u)^2 \boldsymbol{1}^\top(\boldsymbol{K}_{\mathrm{T}} - \boldsymbol{b}\boldsymbol{b}^\top/A)\boldsymbol{1}} \\
&= A(1-t)^2 + \frac{A(a - B/A)}{1 + B/A}t^2 \\
&\geq \frac{\boldsymbol{d}_-^\top \boldsymbol{K}\boldsymbol{d}_-}{4} + A(1-a)\left(\frac{1 + B/A}{2} - t\right) \\
&\geq \frac{\boldsymbol{d}_-^\top \boldsymbol{K}\boldsymbol{d}_-}{4} + \frac{(A-B)(1-a)(1-u)}{2(1+u)}.
\end{aligned} \tag{B.8}$$

Here the first inequality follows from Cauchy–Schwarz in the Schur-complement inner product, the second from integrating the derivative bound, and the last from the displayed lower bound for $(1 + B/A)/2 - t$. Substituting this estimate into the formula for $1 - Q(\boldsymbol{z}_1^{\mathrm{G}})/Q(\boldsymbol{z}_0)$ gives

$$\begin{aligned}
\frac{Q(\boldsymbol{z}_1^{\mathrm{G}})}{Q(\boldsymbol{z}_0)} - q &\geq (1-q)\frac{2(A-B)(1-a)(1-u)}{(A - 2B + D)(1+u) + 2(A-B)(1-a)(1-u)} \\
&\geq \frac{1-q}{2}(1-a)(1-u) = \frac{1-q}{2}\frac{A-D}{A+B}(1-u).
\end{aligned}$$

For the second inequality, $A - 2B + D \leq 2(A - B)$ and $(1+u) + (1-a)(1-u) \leq 2$, so the denominator is at most $4(A-B)$. Since $Q(\boldsymbol{z}_1^{\mathrm{G}})/Q(\boldsymbol{z}_0) \geq q + g > 0$, Lemma B.2 gives $Q(\boldsymbol{z}_{k+1}^{\mathrm{G}})/Q(\boldsymbol{z}_k^{\mathrm{G}}) \geq q + g$ for every $k \geq 0$. This proves (3.5). □

B.4 Mean-direction perturbation

The inputs $\boldsymbol{x}$ and $\boldsymbol{x}^{(0)}$ have the same centered Gaussian coordinates. We first record the resulting population estimate.

Lemma B.3. *Under the input model in Section 2.1,*

$$\|\mathcal{H}_{\mathrm{pop}} - \mathcal{H}_0\|_{\mathrm{op}} \le 16\big(\max_j \sigma_j\big)^2 \mu \left(\mu + 3^{1/4}\sqrt{\frac{\gamma}{1-\rho}}\right)^3 \delta. \tag{B.9}$$

Proof. Put $\boldsymbol{y} = \boldsymbol{x} - \mu\bar{\boldsymbol{u}} = \boldsymbol{x}^{(0)} - \mu\boldsymbol{u}_{j_\star}$ and $R = \mu + \|\boldsymbol{y}\|_2$. Since $\|\bar{\boldsymbol{u}}\|_2 = \|\boldsymbol{u}_{j_\star}\|_2 = 1$, we have

$$\|\boldsymbol{x}\|_2 \le R, \qquad \|\boldsymbol{x}^{(0)}\|_2 \le R, \qquad \|\boldsymbol{x} - \boldsymbol{x}^{(0)}\|_2 = \mu\delta.$$

For $\boldsymbol{B} = \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{W}_\star$ and $\boldsymbol{B}_0 = \boldsymbol{x}^{(0)}\boldsymbol{x}^{(0)\top}\boldsymbol{W}_\star$, we have

$$\begin{aligned}
\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}^{(0)}\boldsymbol{x}^{(0)\top} &= (\boldsymbol{x} - \boldsymbol{x}^{(0)})\boldsymbol{x}^\top + \boldsymbol{x}^{(0)}(\boldsymbol{x} - \boldsymbol{x}^{(0)})^\top, \\
\|\boldsymbol{B}\|_{\mathrm{F}} \le (\max_j \sigma_j)R^2, &\qquad \|\boldsymbol{B}_0\|_{\mathrm{F}} \le (\max_j \sigma_j)R^2, \\
\|\boldsymbol{B} - \boldsymbol{B}_0\|_{\mathrm{F}} &\le 2(\max_j \sigma_j)R\mu\delta.
\end{aligned}$$

For matrices $\boldsymbol{C}, \boldsymbol{D}$, let $(\boldsymbol{C} \otimes \boldsymbol{D})[\boldsymbol{Z}] = \boldsymbol{C}\langle \boldsymbol{D}, \boldsymbol{Z}\rangle$. Then

$$\|\boldsymbol{B} \otimes \boldsymbol{B} - \boldsymbol{B}_0 \otimes \boldsymbol{B}_0\|_{\mathrm{op}} \le (\|\boldsymbol{B}\|_{\mathrm{F}} + \|\boldsymbol{B}_0\|_{\mathrm{F}})\|\boldsymbol{B} - \boldsymbol{B}_0\|_{\mathrm{F}}.$$

By (B.1), we have

$$\mathcal{H}_{\mathrm{pop}} = 4\mathbb{E}(\boldsymbol{B} \otimes \boldsymbol{B}), \qquad \mathcal{H}_0 = 4\mathbb{E}(\boldsymbol{B}_0 \otimes \boldsymbol{B}_0).$$

Since $\boldsymbol{y}$ is centered Gaussian with covariance $\gamma\boldsymbol{A}_0\,\mathrm{diag}(1, \rho, \ldots, \rho^{d-1})\boldsymbol{A}_0^\top$, we have

$$\mathbb{E}\|\boldsymbol{y}\|_2^4 = \gamma^2\left[\left(\sum_{i=1}^d \rho^{i-1}\right)^2 + 2\sum_{i=1}^d \rho^{2i-2}\right] \le \frac{3\gamma^2}{(1-\rho)^2}.$$

Consequently,

$$\begin{aligned}
\|\mathcal{H}_{\mathrm{pop}} - \mathcal{H}_0\|_{\mathrm{op}} &\le 16(\max_j \sigma_j)^2\mu\delta\,\mathbb{E}R^3 \\
&\le 16(\max_j \sigma_j)^2\mu\delta\left(\mu + 3^{1/4}\sqrt{\frac{\gamma}{1-\rho}}\right)^3.
\end{aligned} \tag{B.10}$$

The second inequality follows from Minkowski's and Hölder's inequalities. This proves (B.9).

For a fixed sample the same pointwise calculation, followed by Hölder's inequality, gives

$$\|\mathcal{H}_N - \mathcal{H}_{N,0}\|_{\mathrm{op}} \le 16(\max_j \sigma_j)^2\mu\delta\left[\frac{1}{N}\sum_{t=1}^N (\mu + \|\boldsymbol{x}_t - \mu\bar{\boldsymbol{u}}\|_2)^4\right]^{3/4}, \tag{B.11}$$

where $\mathcal{H}_{N,0}$ and $\mathcal{H}_N$ are the empirical operators at $\boldsymbol{W}_\star$ computed from $\boldsymbol{x}_t^{(0)}$ and $\boldsymbol{x}_t$, respectively. □

B.5 Empirical curvature estimates

The following lemma controls the empirical Hessian at $\boldsymbol{W}_\star$ on the full matrix space. Its proof combines the Gaussian moment estimates with the mean-direction estimate in Appendix B.4.

Lemma B.4. *Suppose the observations follow the input and response model in Section 2.1. There is a universal constant $C_0 > 1$ such that, for $0 < \zeta < 1$, if*

$$N \ge C_0 dr \log^3\frac{dr}{\zeta},$$

then, with probability at least $1 - \zeta$,

$$\|\nabla^2\mathcal{L}_N(\boldsymbol{W}_\star) - \mathcal{H}_0\|_{\mathrm{op}} \le C_0\|\mathcal{H}_0\|_{\mathrm{op}}\left(\varepsilon_N + \kappa_{\mathrm{act}}^2\delta\right). \tag{B.12}$$

Proof. At $\boldsymbol{W}_\star$, the empirical Gauss–Newton operator is

$$\mathcal{H}_N[\boldsymbol{D}] = \frac{4}{N}\sum_{t=1}^{N}(\boldsymbol{x}_t^\top \boldsymbol{D}\boldsymbol{W}_\star^\top \boldsymbol{x}_t)\boldsymbol{x}_t\boldsymbol{x}_t^\top \boldsymbol{W}_\star. \tag{B.13}$$

Let $\boldsymbol{\psi}(\boldsymbol{\xi})$ collect the normalized Gaussian Hermite coordinates

$$1, \qquad \xi_i\ (1\le i\le d), \qquad \frac{\xi_i^2-1}{\sqrt{2}}\ (i\le r), \qquad \xi_i\xi_j\ (1\le i<j\le d,\ i\le r). \tag{B.14}$$

Its dimension is $p = 1 + d + dr - r(r-1)/2 \le 2dr$, and orthogonality of the Hermite coordinates gives

$$\mathbb{E}\boldsymbol{\psi}\boldsymbol{\psi}^\top = \boldsymbol{I}_p. \tag{B.15}$$

We prove that its empirical Gram matrix is close to $\boldsymbol{I}_p$. Set $\ell = \max\{2, \lceil \log(4p/\zeta)\rceil\}$ and $\boldsymbol{Y}_t = \boldsymbol{\psi}_t\boldsymbol{\psi}_t^\top - \boldsymbol{I}_p$. For a random scalar X, write $\|X\|_{L^s} = (\mathbb{E}|X|^s)^{1/s}$. Gaussian hypercontractivity [11] gives

$$\|\boldsymbol{a}^\top\boldsymbol{\psi}\|_{L^s} \le (s-1)\|\boldsymbol{a}\|_2, \qquad s\ge 2.$$

In particular, for every unit vector $\boldsymbol{a}$,

$$\boldsymbol{a}^\top\mathbb{E}\boldsymbol{Y}_t^2\boldsymbol{a} = \sum_{j=1}^{p}\mathbb{E}\big[\psi_j^2(\boldsymbol{a}^\top\boldsymbol{\psi})^2\big] - 1 \le \sum_{j=1}^{p}\|\psi_j\|_{L^4}^2\|\boldsymbol{a}^\top\boldsymbol{\psi}\|_{L^4}^2 \le 81p.$$

Moreover, Minkowski's inequality gives

$$\big\|\|\boldsymbol{Y}_t\|_{\mathrm{op}}\big\|_{L^{4\ell}} \le 1 + \sum_{j=1}^{p}\|\psi_j\|_{L^{8\ell}}^2 \le Cp\ell^2.$$

Applying the matrix Rosenthal inequality [26, Corollary 7.4] and comparing the operator norm with the Schatten 4ℓ-norm yields

$$\left\|\left\|\frac{1}{N}\sum_{t=1}^{N}\boldsymbol{Y}_t\right\|_{\mathrm{op}}\right\|_{L^{4\ell}} \le C\left\{\sqrt{\frac{p\ell}{N}} + \frac{p\ell^3 N^{1/(4\ell)}}{N}\right\} \le C\sqrt{\frac{p\ell^3}{N}} \le \frac{C_2\varepsilon_N}{e}.$$

The second inequality uses $N/(p\ell^3) \ge 1$ and the boundedness of $p^{1/(4\ell)}$ and $\ell^{3/(4\ell)}$. Since $p \le 2dr$ and $dr \ge 9$, we have $\ell \le \log(8dr/\zeta) + 1 \le 3\log(dr/\zeta)$. The sample-size condition gives $N/(p\ell^3) \ge C_0/54 \ge 1$ for $C_0 \ge 54$. The last inequality then holds for a fixed universal $C_2 > 1$. Choosing $C_0 \ge 4C_2^2$ also gives $C_2\varepsilon_N \le 1/2$. Markov's inequality yields

$$\mathbb{P}\left\{\left\|\frac{1}{N}\sum_{t=1}^{N}\boldsymbol{\psi}_t\boldsymbol{\psi}_t^\top - \boldsymbol{I}_p\right\|_{\mathrm{op}} > C_2\varepsilon_N\right\} \le e^{-4\ell} \le \frac{\zeta}{2}. \tag{B.16}$$

On the complementary event, for every $\boldsymbol{a}\in\mathbb{R}^p$,

$$(1 - C_2\varepsilon_N)\|\boldsymbol{a}\|_2^2 \le \frac{1}{N}\sum_{t=1}^{N}(\boldsymbol{a}^\top\boldsymbol{\psi}_t)^2 \le (1 + C_2\varepsilon_N)\|\boldsymbol{a}\|_2^2. \tag{B.17}$$

Thus the empirical Gram matrix is positive definite on this event.

We next bound the empirical fourth moment. Put $R_t = \mu + \|\boldsymbol{x}_t - \mu\bar{\boldsymbol{u}}\|_2$. Minkowski's inequality and the Gaussian fourth moment give

$$\mathbb{E}R_t^4 \le \left(\mu + 3^{1/4}\sqrt{\frac{\gamma}{1-\rho}}\right)^4 \le 8\left(\mu^2 + \frac{\gamma}{1-\rho}\right)^2.$$

For every integer $k\ge 2$, Minkowski's inequality applied to $\|\boldsymbol{x}_t - \mu\bar{\boldsymbol{u}}\|_2^2 = \gamma\sum_{i=1}^{d}\rho^{i-1}\xi_{ti}^2$ gives

$$\big\|\|\boldsymbol{x}_t - \mu\bar{\boldsymbol{u}}\|_2^2\big\|_{L^{2k}} \le \frac{4k\gamma}{1-\rho}, \qquad \|R_t^4\|_{L^k} \le Ck^2\left(\mu^2 + \frac{\gamma}{1-\rho}\right)^2.$$

Applying the scalar case of the same Rosenthal inequality to $R_t^4 - \mathbb{E}R_t^4$ yields

$$\left\| \frac{1}{N}\sum_{t=1}^{N}(R_t^4 - \mathbb{E}R_t^4) \right\|_{L^{4\ell}} \le C\left\{ \sqrt{\frac{\ell}{N}} + \frac{\ell^3 N^{1/(4\ell)}}{N} \right\} \left(\mu^2 + \frac{\gamma}{1-\rho}\right)^2 \le \frac{8}{e}\left(\mu^2 + \frac{\gamma}{1-\rho}\right)^2,$$

where the last inequality follows from the sample-size condition by fixing C_0 sufficiently large. Markov's inequality therefore gives

$$\mathbb{P}\left\{ \frac{1}{N}\sum_{t=1}^{N} R_t^4 > 16\left(\mu^2 + \frac{\gamma}{1-\rho}\right)^2 \right\} \le e^{-4\ell} \le \frac{\zeta}{4}. \tag{B.18}$$

In fact,

$$\frac{1}{N}\sum_{t=1}^{N}\|\boldsymbol{x}_t\|_2^4 \le \frac{1}{N}\sum_{t=1}^{N} R_t^4 \le 16\left(\mu^2 + \frac{\gamma}{1-\rho}\right)^2. \tag{B.19}$$

In the remainder of the proof, work on the event where (B.17) holds and $N^{-1}\sum_t R_t^4 \le 16(\mu^2 + \gamma/(1-\rho))^2$. By (B.16) and (B.18), its probability is at least $1-\zeta$.

Write $\boldsymbol{D} = \boldsymbol{A}_0\boldsymbol{M}\boldsymbol{V}^\top$ and $\widetilde{x}_i = \mu(\boldsymbol{A}_0^\top \bar{\boldsymbol{u}})_i + \sqrt{\gamma\rho^{i-1}}\xi_i$. Then

$$2\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{W}_\star^\top \boldsymbol{x} = 2\sum_{a=1}^{d}\sum_{j=1}^{r}\sigma_j M_{aj}\widetilde{x}_a\widetilde{x}_j.$$

The product has the Hermite expansion

$$\widetilde{x}_a\widetilde{x}_j = \mathbb{E}(\widetilde{x}_a\widetilde{x}_j) + (\mathbb{E}\widetilde{x}_a)\sqrt{\gamma\rho^{j-1}}\,\xi_j + (\mathbb{E}\widetilde{x}_j)\sqrt{\gamma\rho^{a-1}}\,\xi_a + \gamma\rho^{(a+j-2)/2}\begin{cases}\xi_j^2 - 1, & a = j,\\ \xi_a\xi_j, & a \ne j.\end{cases}$$

Since $j \le r$, the displayed response belongs to the span of (B.14). We also have

$$\langle \boldsymbol{D}, \mathcal{H}_N[\boldsymbol{D}]\rangle = \frac{1}{N}\sum_{t=1}^{N}(2\boldsymbol{x}_t^\top \boldsymbol{D}\boldsymbol{W}_\star^\top \boldsymbol{x}_t)^2, \qquad \langle \boldsymbol{D}, \mathcal{H}_{\mathrm{pop}}[\boldsymbol{D}]\rangle = \mathbb{E}(2\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{W}_\star^\top \boldsymbol{x})^2,$$

Thus (B.17) gives

$$(1 - C_2\varepsilon_N)\mathcal{H}_{\mathrm{pop}} \preceq \mathcal{H}_N \preceq (1 + C_2\varepsilon_N)\mathcal{H}_{\mathrm{pop}}.$$

Changing $\boldsymbol{V}$ or $\bar{\boldsymbol{u}}$ changes only the coefficients of these linear combinations. The event is therefore simultaneous over all orthogonal $\boldsymbol{V}$ and all unit $\bar{\boldsymbol{u}}$. In particular, $\|\mathcal{H}_{N,0} - \mathcal{H}_0\|_{\mathrm{op}} \le C_2\varepsilon_N\|\mathcal{H}_0\|_{\mathrm{op}}$ for the coupled aligned input. The relative operator inequalities imply

$$\|\mathcal{H}_N - \mathcal{H}_{\mathrm{pop}}\|_{\mathrm{op}} \le C_2\varepsilon_N\|\mathcal{H}_{\mathrm{pop}}\|_{\mathrm{op}}. \tag{B.20}$$

To compare directly with $\mathcal{H}_0$, use (B.11) and the fourth-moment estimate. Since $\|\mathcal{H}_0\|_{\mathrm{op}} \ge K_{j_\star j_\star} \ge 4\sigma_{j_\star}^2(\mu^2 + \gamma\rho^{j_\star - 1})^2$, we have

$$\frac{\|\mathcal{H}_N - \mathcal{H}_{N,0}\|_{\mathrm{op}}}{\|\mathcal{H}_0\|_{\mathrm{op}}} \le 32\left(\frac{\max_j \sigma_j}{\sigma_{j_\star}}\right)^2 \frac{\mu\left(\mu^2 + \gamma/(1-\rho)\right)^{3/2}}{(\mu^2 + \gamma\rho^{j_\star-1})^2}\delta \le \frac{32}{\rho^{2j_\star-2}(1-\rho)^2}\left(\frac{\max_j \sigma_j}{\min_j \sigma_j}\right)^2 \delta \le 32\kappa_{\mathrm{act}}^2\delta.$$

The second inequality follows from $\mu \le (\mu^2 + \gamma/(1-\rho))^{1/2}$ and

$$\mu^2 + \frac{\gamma}{1-\rho} \le \frac{\mu^2 + \gamma\rho^{j_\star - 1}}{\rho^{j_\star-1}(1-\rho)}.$$

The last inequality uses $j_\star \le r$ and (4.2). Combining this with the estimate for $\mathcal{H}_{N,0} - \mathcal{H}_0$ and taking $C_0 \ge \max\{32, C_2\}$ gives (B.12), since $\nabla^2\mathcal{L}_N(\boldsymbol{W}_\star) = \mathcal{H}_N$. □

Recovery of the response weights. The same event also identifies the target factor when the model basis $\boldsymbol{U}$ is given. Define

$$\widehat{\boldsymbol{\lambda}} = \operatorname*{arg\,min}_{\boldsymbol{\lambda}\in\mathbb{R}^r} \sum_{t=1}^{N} \left[y_t - \sum_{j=1}^{r} \lambda_j (\boldsymbol{u}_j^\top \boldsymbol{x}_t)^2 \right]^2 . \tag{B.21}$$

Lemma B.5. *On the event constructed in the proof of Lemma B.4, the minimizer in (B.21) is unique and satisfies $\widehat{\lambda}_j = \sigma_j^2$ for $1 \le j \le r$. Consequently,*

$$\boldsymbol{W}_c = \boldsymbol{U}\,\mathrm{diag}(\sqrt{\widehat{\lambda}_j})\boldsymbol{V}^\top = \boldsymbol{W}_\star.$$

Proof. For every unit mean direction, $\boldsymbol{u}_i^\top \boldsymbol{x} = \mu \boldsymbol{u}_i^\top \bar{\boldsymbol{u}} + \sqrt{\gamma\rho^{i-1}}\xi_i$ for $i \le r$. For any $\boldsymbol{a} \in \mathbb{R}^r$, orthogonality of the Hermite coordinates gives

$$\mathbb{E}\left[\left(\sum_{j=1}^{r} a_j (\boldsymbol{u}_j^\top \boldsymbol{x})^2 \right) \frac{\xi_i^2 - 1}{\sqrt{2}} \right] = \sqrt{2}\, a_i \gamma \rho^{i-1}.$$

On the event (B.17), Bessel's inequality yields

$$\begin{aligned}
\frac{1}{N}\sum_{t=1}^{N} \left(\sum_{j=1}^{r} a_j (\boldsymbol{u}_j^\top \boldsymbol{x}_t)^2 \right)^2 &\ge \tfrac{1}{2}\mathbb{E}\left(\sum_{j=1}^{r} a_j (\boldsymbol{u}_j^\top \boldsymbol{x})^2 \right)^2 \\
&\ge \tfrac{1}{2}\sum_{i=1}^{r} \left| \mathbb{E}\left[\left(\sum_{j=1}^{r} a_j (\boldsymbol{u}_j^\top \boldsymbol{x})^2 \right) \frac{\xi_i^2 - 1}{\sqrt{2}} \right] \right|^2 \\
&= \gamma^2 \sum_{i=1}^{r} a_i^2 \rho^{2i-2} > 0 \qquad (\boldsymbol{a} \neq 0).
\end{aligned}$$

The coefficients $\lambda_j = \sigma_j^2$ give zero residual. Applying the strict lower bound to $a_j = \widehat{\lambda}_j - \sigma_j^2$ proves uniqueness and the stated identities. □

B.6 Proof of Theorem 4.1

The curvature estimates control the empirical loss near $\boldsymbol{W}_\star$. We first show how errors in the loss and the current iterate affect a single line-search update, and then apply this estimate successively to both methods.

Lemma B.6. *Suppose $0 \preceq \mathcal{H} \preceq \mathcal{I}$, $\mathcal{H}(\mathcal{S}) \subseteq \mathcal{S}$, and the eigenvalues of $\mathcal{H}|_{\mathcal{S}}$ lie in $[\kappa^{-1}, 1]$, where $\kappa > r - 1 \ge 2$. Let $0 < \nu \le 1$ and $\boldsymbol{Z}_* \in \mathcal{S}$ with $\|\boldsymbol{Z}_*\|_{\mathrm{F}} = 1$. For Muon, set $\boldsymbol{D}_* = \mathrm{polar}(\mathcal{H}[\boldsymbol{Z}_*])/\sqrt{r}$ and suppose $\sigma_r(\mathcal{H}[\boldsymbol{Z}_*]) \ge \nu/(2\kappa\sqrt{r})$ and $\langle \mathcal{H}[\boldsymbol{Z}_*], \boldsymbol{D}_* \rangle \ge 2/(3\kappa)$. For GD, set $\boldsymbol{D}_* = \mathcal{H}[\boldsymbol{Z}_*]/\|\mathcal{H}[\boldsymbol{Z}_*]\|_{\mathrm{F}}$. Suppose F is twice continuously differentiable, set $Q(\boldsymbol{Z}) = \langle \boldsymbol{Z}, \mathcal{H}[\boldsymbol{Z}] \rangle/2$, and assume*

$$\sup_{\|\boldsymbol{Z}\|_{\mathrm{F}} \le 4\sqrt{\kappa}} \max\left\{ |F(\boldsymbol{Z}) - Q(\boldsymbol{Z})|,\ \|\nabla F(\boldsymbol{Z}) - \mathcal{H}[\boldsymbol{Z}]\|_{\mathrm{F}},\ \|\nabla^2 F(\boldsymbol{Z}) - \mathcal{H}\|_{\mathrm{op}} \right\} \le v.$$

Let $\boldsymbol{Z}_{*,+}$ denote the exact line-search update of Q from $\boldsymbol{Z}_*$ along $-\boldsymbol{D}_*$. For $\boldsymbol{Z}$ satisfying $\|\boldsymbol{Z} - \boldsymbol{Z}_*\|_{\mathrm{F}} \le e$, put $u = e + v$. There exists a universal constant $c_1 > 0$ such that, if

$$u \le c_1 \frac{\nu}{\kappa^3}, \tag{B.22}$$

then the chosen method applied to F at $\boldsymbol{Z}$ has a first positive strict local line minimizer. If $\boldsymbol{Z}_+$ is the resulting iterate, then

$$\|\boldsymbol{Z}_+ - \boldsymbol{Z}_{*,+}\|_{\mathrm{F}} \le \frac{32\kappa^2}{\nu} u, \qquad \left| \frac{F(\boldsymbol{Z}_+)}{F(\boldsymbol{Z})} - \frac{Q(\boldsymbol{Z}_{*,+})}{Q(\boldsymbol{Z}_*)} \right| \le \frac{160\kappa^3}{\nu} u. \tag{B.23}$$

Here $F(\boldsymbol{Z}) > 0$.

The first estimate controls the next iterate, while the second controls the quantity compared in Theorem 4.1. Both estimates depend on the same sum of the state and objective errors.

Proof. For full-column-rank $\boldsymbol{G}$ and $\boldsymbol{G}_*$, write their polar decompositions as $\boldsymbol{G} = \boldsymbol{P}\boldsymbol{S}$ and $\boldsymbol{G}_* = \boldsymbol{P}_*\boldsymbol{S}_*$, where $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{P}_*^\top\boldsymbol{P}_* = \boldsymbol{I}_r$ and $\boldsymbol{S}, \boldsymbol{S}_* \succ 0$. Expanding the trace and applying Cauchy–Schwarz gives

$$\begin{aligned}\frac{\sigma_r(\boldsymbol{G}) + \sigma_r(\boldsymbol{G}_*)}{2}\|\boldsymbol{P} - \boldsymbol{P}_*\|_{\mathrm{F}}^2 &\le \operatorname{tr}\left[(\boldsymbol{S} + \boldsymbol{S}_*)\left(\boldsymbol{I}_r - \frac{\boldsymbol{P}^\top\boldsymbol{P}_* + \boldsymbol{P}_*^\top\boldsymbol{P}}{2}\right)\right] \\ &= \langle \boldsymbol{G} - \boldsymbol{G}_*, \boldsymbol{P} - \boldsymbol{P}_*\rangle \le \|\boldsymbol{G} - \boldsymbol{G}_*\|_{\mathrm{F}}\|\boldsymbol{P} - \boldsymbol{P}_*\|_{\mathrm{F}}, \\ \|\boldsymbol{P} - \boldsymbol{P}_*\|_{\mathrm{F}} &\le \frac{2\|\boldsymbol{G} - \boldsymbol{G}_*\|_{\mathrm{F}}}{\sigma_r(\boldsymbol{G}) + \sigma_r(\boldsymbol{G}_*)}. \end{aligned} \tag{B.24}$$

Here the first inequality uses $\boldsymbol{I}_r - (\boldsymbol{P}^\top\boldsymbol{P}_* + \boldsymbol{P}_*^\top\boldsymbol{P})/2 \succeq 0$, and the last follows by division. Let $\boldsymbol{D}$ be the unit search direction for F at $\boldsymbol{Z}$. For Muon, Weyl's inequality gives

$$\sigma_r(\nabla F(\boldsymbol{Z})) \ge \frac{\nu}{2\kappa\sqrt{r}} - u > \frac{\nu}{4\kappa\sqrt{r}} > 0.$$

Equation (B.24) and $\boldsymbol{D} = \operatorname{polar}(\nabla F(\boldsymbol{Z}))/\sqrt{r}$ give the first bound below. For GD, it follows from $\|\mathcal{H}[\boldsymbol{Z}_*]\|_{\mathrm{F}} \ge \kappa^{-1}$ and $\|X/\|X\|_{\mathrm{F}} - Y/\|Y\|_{\mathrm{F}}\|_{\mathrm{F}} \le 2\|X - Y\|_{\mathrm{F}}/\|Y\|_{\mathrm{F}}$. Hence

$$\|\boldsymbol{D} - \boldsymbol{D}_*\|_{\mathrm{F}} \le \frac{4\kappa}{\nu}u \le \frac{1}{128\kappa^2},$$

where the last inequality follows from (B.22) by taking c_1 sufficiently small. Let $h_* = \langle \boldsymbol{D}_*, \mathcal{H}[\boldsymbol{D}_*]\rangle$ and $h = \langle \boldsymbol{D}, \mathcal{H}[\boldsymbol{D}]\rangle$. Then $h \ge h_* - 2\|\boldsymbol{D} - \boldsymbol{D}_*\|_{\mathrm{F}} \ge (2\kappa)^{-1}$. The exact step $\alpha_* = \langle \mathcal{H}[\boldsymbol{Z}_*], \boldsymbol{D}_*\rangle/h_*$ satisfies $2/(3\kappa) \le \alpha_* \le \sqrt{\kappa}$. The lower bound follows from $\langle \mathcal{H}[\boldsymbol{Z}_*], \boldsymbol{D}_*\rangle \ge 2/(3\kappa)$ and $h_* \le 1$; the upper bound follows from $\langle \mathcal{H}[\boldsymbol{Z}_*], \boldsymbol{D}_*\rangle^2 \le \langle \boldsymbol{Z}_*, \mathcal{H}[\boldsymbol{Z}_*]\rangle h_* \le h_*$ and $h_* \ge \kappa^{-1}$. Since $\mathcal{H}^2 \preceq \mathcal{H}$, we also have $\alpha_*\|\mathcal{H}[\boldsymbol{D}_*]\|_{\mathrm{F}} \le 1$. Since $Q(\boldsymbol{Z}_{*,+}) \le Q(\boldsymbol{Z}_*)$, $\|\mathcal{H}[\boldsymbol{Z}_{*,+}]\|_{\mathrm{F}}^2 \le 2Q(\boldsymbol{Z}_{*,+}) \le 2Q(\boldsymbol{Z}_*) \le 1$.

For a unit $\boldsymbol{D}$, the quadratic line-minimization map is $P_{\boldsymbol{D}} = \mathcal{I} - \boldsymbol{D} \otimes \mathcal{H}[\boldsymbol{D}]/h$, where $(\boldsymbol{D} \otimes \mathcal{H}[\boldsymbol{D}])[\boldsymbol{Z}] = \boldsymbol{D}\langle \mathcal{H}[\boldsymbol{D}], \boldsymbol{Z}\rangle$. Its norm satisfies

$$\|P_{\boldsymbol{D}}\|_{\mathrm{op}}^2 = 1 + \frac{\|\mathcal{H}[\boldsymbol{D}] - h\boldsymbol{D}\|_{\mathrm{F}}^2}{h^2} = \frac{\|\mathcal{H}[\boldsymbol{D}]\|_{\mathrm{F}}^2}{h^2} \le h^{-1} \le 2\kappa,$$

where $\|\mathcal{H}[\boldsymbol{D}]\|_{\mathrm{F}}^2 \le h$ gives the first inequality. Let $\alpha(\boldsymbol{D})$ be the quadratic line minimizer from $\boldsymbol{Z}_*$ along $-\boldsymbol{D}$. Orthogonality at $\boldsymbol{Z}_{*,+}$ gives

$$|\alpha_* - \alpha(\boldsymbol{D})| = \left|\frac{\alpha_*\langle \boldsymbol{D}, \mathcal{H}[\boldsymbol{D} - \boldsymbol{D}_*]\rangle - \langle \mathcal{H}[\boldsymbol{Z}_{*,+}], \boldsymbol{D} - \boldsymbol{D}_*\rangle}{h}\right| \le 5\kappa\|\boldsymbol{D} - \boldsymbol{D}_*\|_{\mathrm{F}}.$$

The inequality uses $\alpha_*\|\mathcal{H}[\boldsymbol{D}]\|_{\mathrm{F}} \le 1 + \sqrt{\kappa}\|\boldsymbol{D} - \boldsymbol{D}_*\|_{\mathrm{F}}$. It follows that

$$\|P_{\boldsymbol{D}}\boldsymbol{Z} - \boldsymbol{Z}_{*,+}\|_{\mathrm{F}} \le \sqrt{2\kappa}\, e + 6\kappa\|\boldsymbol{D} - \boldsymbol{D}_*\|_{\mathrm{F}}.$$

The quadratic step $\widehat{\alpha}$ from $\boldsymbol{Z}$ also satisfies $|\widehat{\alpha} - \alpha_*| \le \sqrt{2\kappa}\, e + 5\kappa\|\boldsymbol{D} - \boldsymbol{D}_*\|_{\mathrm{F}} \le 24\kappa^2 u/\nu \le 3/(64\kappa)$.

For $0 \le \alpha \le 2\alpha_*$, $\|\boldsymbol{Z} - \alpha\boldsymbol{D}\|_{\mathrm{F}} \le 1 + e + 2\sqrt{\kappa} < 4\sqrt{\kappa}$. The assumptions on F therefore imply

$$\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}F(\boldsymbol{Z} - \alpha\boldsymbol{D}) \ge h - v \ge (4\kappa)^{-1}, \qquad 0 \le \alpha \le 2\alpha_*.$$

At $\alpha = \widehat{\alpha}$, the derivative has absolute value at most v, since the derivative of $Q(\boldsymbol{Z} - \alpha\boldsymbol{D})$ vanishes. By (B.22), after decreasing c_1 if necessary, $|\widehat{\alpha} - \alpha_*| + 8\kappa v < \alpha_*/2$. Thus $\widehat{\alpha} \pm 8\kappa v$ lie in $[\alpha_*/2, 3\alpha_*/2]$. Since the line curvature is at least $(4\kappa)^{-1}$ and its derivative at $\widehat{\alpha}$ has absolute value at most v,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}F(\boldsymbol{Z} - \alpha\boldsymbol{D})\bigg|_{\alpha = \widehat{\alpha} - 8\kappa v} \le -v \le 0 \le v \le \frac{\mathrm{d}}{\mathrm{d}\alpha}F(\boldsymbol{Z} - \alpha\boldsymbol{D})\bigg|_{\alpha = \widehat{\alpha} + 8\kappa v}.$$

The derivative is strictly increasing on $[0, 2\alpha_*]$. Hence its first positive zero, denoted by α, satisfies $|\alpha - \widehat{\alpha}| \le 4\kappa v$, including when $v = 0$. Consequently,

$$\|\boldsymbol{Z}_+ - \boldsymbol{Z}_{*,+}\|_{\mathrm{F}} \le \sqrt{2\kappa}\, e + \frac{24\kappa^2}{\nu} u + 4\kappa v \le \frac{32\kappa^2}{\nu} u.$$

Since $\|\mathcal{H}[\boldsymbol{Z}_{*,+}]\|_{\mathrm{F}} \le 1$, we have

$$|F(\boldsymbol{Z}_+) - Q(\boldsymbol{Z}_{*,+})| \le v + \|\boldsymbol{Z}_+ - \boldsymbol{Z}_{*,+}\|_{\mathrm{F}} + \tfrac{1}{2}\|\boldsymbol{Z}_+ - \boldsymbol{Z}_{*,+}\|_{\mathrm{F}}^2 \le \left(1 + \frac{33\kappa^2}{\nu}\right) u,$$

$$\left|\frac{F(\boldsymbol{Z}_+)}{F(\boldsymbol{Z})} - \frac{Q(\boldsymbol{Z}_{*,+})}{Q(\boldsymbol{Z}_*)}\right| \le 4\kappa \left(3 + \frac{33\kappa^2}{\nu}\right) u \le \frac{160\kappa^3}{\nu} u.$$

The second line uses $|F(\boldsymbol{Z}) - Q(\boldsymbol{Z}_*)| \le 2u$, $F(\boldsymbol{Z}) \ge (4\kappa)^{-1}$, and $Q(\boldsymbol{Z}_{*,+}) \le Q(\boldsymbol{Z}_*)$.

□

For the reference Hessian $\mathcal{H}_0$ in Proposition 2.1, set

$$L = \|\mathcal{H}_0\|_{\mathrm{op}}, \qquad \kappa = L/\lambda_{\min}(\boldsymbol{K}).$$

Define

$$\nu = (r-1)\min_j w_j, \qquad \Lambda = \frac{32\kappa^2}{\nu\sqrt{q}}. \tag{B.25}$$

Here g is defined in (3.3). For the parameters $\varepsilon, \varepsilon_N, \delta$ in Theorem 4.1, set

$$\begin{aligned} v =& 8C_0\kappa\left(\varepsilon_N + \kappa_{\mathrm{act}}^2\delta\right) + 8192\kappa^2 \frac{\left(\mu^2 + \gamma/(1-\rho)\right)^2 \min_j \sigma_j}{L}\varepsilon(\max_j \sigma_j + \kappa\varepsilon \min_j \sigma_j), \\ u =& \varepsilon + 2v. \end{aligned} \tag{B.26}$$

The next lemma expresses the quantities in the one-update estimate in terms of the model parameters. It also verifies the smallness condition needed to apply that estimate throughout the interval in (4.3).

Lemma B.7. *Under the input model and scale conditions in Theorem 4.1, let $\mathfrak{R} = \kappa_{\mathrm{act}}/\eta$. Then $\mathfrak{R} \ge 32$ and*

$$\kappa \le \mathfrak{R}^3, \qquad \nu \ge \mathfrak{R}^{-2}, \qquad q \ge \mathfrak{R}^{-2}, \qquad 1-q \ge \mathfrak{R}^{-4}, \qquad g \ge \mathfrak{R}^{-10}, \qquad \Lambda \le \mathfrak{R}^{10}. \tag{B.27}$$

For any sufficiently small universal constant $c_2 > 0$, there is a sufficiently large universal constant $C_1 > 0$ such that every integer $T \ge 2$ satisfying (4.3) obeys

$$u \le c_2 \frac{\nu \min\{g, q\}}{\kappa^{7/2}} \Lambda^{-(T-1)}. \tag{B.28}$$

In particular, (B.28) combines the initialization, mean, and sampling bounds into the condition used in the induction below.

Proof. Assumption 3.1 gives $c > 1$. Since $r \ge 3$ and $\rho^{r-1}(1-\rho)^2 \le \rho^2(1-\rho)^2 \le 1/16$, definition (4.2) gives $\kappa_{\mathrm{act}} > 32$ and hence $\mathfrak{R} \ge 32$. Lemma B.1 gives

$$\lambda_{\min}(\boldsymbol{K}) \ge 8\gamma^2(\min_j \sigma_j)^2\rho^{2r-2}, \qquad L \le 96\gamma^2(\max_j \sigma_j)^2\left(\frac{\mu^2}{\gamma} + \frac{1}{1-\rho}\right)^2.$$

Since $c \ge \mu^2/\gamma$, substituting (4.2) gives $\kappa \le 12\kappa_{\mathrm{act}}^2 \le \mathfrak{R}^3$. We also need $\kappa > r-1$. The smallest eigenvalue of $\boldsymbol{K}$ is at most the average of the $r-1$ diagonal entries indexed by $j \ne j_\star$, so

$$\lambda_{\min}(\boldsymbol{K}) \le \frac{12\gamma^2 S_2}{r-1}, \qquad L \ge A > D = 4\gamma^2(S_1^2 + 2S_2) \ge 12\gamma^2 S_2.$$

These inequalities give $\kappa > r - 1 \ge 2$.

The bound in Appendix B.2 and $Aa - B = D - aB \le D$ give

$$\min_j w_j \ge \frac{(\min_{\ell \ne j_\star} \sigma_\ell \rho^{\ell-1})^2}{S_1^2 + 2S_2}.$$

Since $S_1^2 + 2S_2 \le 3(\max_j \sigma_j)^2/(1-\rho)^2$,

$$\nu \ge \frac{(\min_j \sigma_j)^2 \rho^{2r-2}(1-\rho)^2}{3(\max_j \sigma_j)^2} \ge \kappa_{\mathrm{act}}^{-2} \ge \mathfrak{R}^{-2}.$$

Equation (3.3) gives $q \ge \eta^2 \ge \mathfrak{R}^{-2}$. Using the expressions for A, B, D in Appendix B.2,

$$\begin{aligned} AD - B^2 &= 16\gamma^4 p^2 \left[(4c+2)S_1^2 + 2(c^2+6c+3)S_2\right] \\ &\ge 32\gamma^4 p^2 (c+1)^2 S_2. \end{aligned}$$

Since $A > D$ and $A \le 16\gamma^2 p^2 (c+1)^2$, it follows that

$$\begin{aligned} 1 - q = \frac{4(AD - B^2)}{(A+D)^2 - 4B^2} &\ge \frac{AD - B^2}{A^2} \\ &\ge \frac{S_2}{8p^2(c+1)^2} \ge \frac{1}{8\kappa_{\mathrm{act}}^2} \ge \mathfrak{R}^{-4}. \end{aligned}$$

Moreover, $(A-D)/(A+B) \ge \eta/2$ and

$$1 - \|\boldsymbol{w}\|_2^2 \ge 1 - \max_j w_j \ge (r-2) \min_j w_j \ge \nu/2.$$

Substitution into (3.3) gives $g \ge (8\mathfrak{R}^7)^{-1} \ge \mathfrak{R}^{-10}$. Finally, $\Lambda = 32\kappa^2/(\nu\sqrt{q}) \le 32\mathfrak{R}^9 \le \mathfrak{R}^{10}$, since $\mathfrak{R} \ge 32$.

The curvature estimate (B.12) is supplied by Lemma B.4. For the nonlinear term, $L \ge \lambda_{\min}(\boldsymbol{K})$ and Lemma B.1 give

$$\begin{aligned} 16\left(\mu^2 + \frac{\gamma}{1-\rho}\right)^2 \max\left\{\frac{\min_j \sigma_j \max_j \sigma_j}{L}, \frac{(\min_j \sigma_j)^2}{L}\right\} &\le \frac{2\max_j \sigma_j}{\rho^{2r-2}\min_j \sigma_j}\left(\frac{\mu^2}{\gamma} + \frac{1}{1-\rho}\right)^2 \\ &\le 2\kappa_{\mathrm{act}}^2. \end{aligned}$$

Substituting into (B.26), with $\kappa \le \mathfrak{R}^3$ and $\varepsilon < 1/4$, yields

$$u \le (1 + 2048\mathfrak{R}^8 + 512\mathfrak{R}^{11})\varepsilon + 16C_0\mathfrak{R}^3\varepsilon_N + 16C_0\mathfrak{R}^5\delta.$$

Consequently,

$$u \le (16C_0 + 4096)\mathfrak{R}^{18}(\varepsilon + \varepsilon_N + \delta). \tag{B.29}$$

By (4.3) and (B.27), we have

$$(16C_0 + 4096)\mathfrak{R}^{18 - C_1 T} \le c_2 \mathfrak{R}^{-10T-13} \le c_2 \frac{\nu \min\{g, q\}}{\kappa^{7/2}} \Lambda^{-(T-1)}$$

for $T \ge 2$, provided $C_1 \ge 26 + \log((16C_0 + 4096)/c_2)/(2\log 32)$. This proves (B.28). □

Proof of Theorem 4.1. Choose C_0 as in Lemma B.4. Fix $c_2 > 0$ sufficiently small depending only on c_1 in Lemma B.6, and choose C_1 as in Lemma B.7. Fix the event in Lemma B.4 and any common initial matrix satisfying (4.1). Use $L, \kappa, \nu, g, \Lambda, v, u$ from Appendix B.6. Within this proof normalize the reference loss by L, writing $\mathcal{H} = \mathcal{H}_0/L$ and $Q(\boldsymbol{Z}) = \langle \boldsymbol{Z}, \mathcal{H}[\boldsymbol{Z}]\rangle/2$; $\boldsymbol{K}, A, B, D$ denote their normalized values. Thus $0 \preceq \mathcal{H} \preceq \mathcal{I}$ and $\lambda_{\min}(\boldsymbol{K}) = \kappa^{-1}$, while $a, q, \boldsymbol{w}, \nu, g, \Lambda$ are unchanged. Permute the active coordinates so that $j_\star$ is first; this leaves all norms, singular values, and line-search iterates unchanged and gives the block notation used below.

Let $\boldsymbol{Z}_k^{\mathrm{M}}$ and $\boldsymbol{Z}_k^{\mathrm{G}}$ be the iterates for Q from $\boldsymbol{Z}_0 = \boldsymbol{Z}_{\mathrm{bal}}$. By Proposition 3.2, their successive loss ratios are q and at least $q+g$, respectively. For either sequence, write $t_k = \|\boldsymbol{Z}_k\|_{\mathrm{F}}$. Since $Q(\boldsymbol{Z}_k) \le Q(\boldsymbol{Z}_0) \le 1/2$, $Q(\boldsymbol{Z}_k) \ge t_k^2/(2\kappa)$, and $Q(\boldsymbol{Z}_k) \ge q^{k-j} Q(\boldsymbol{Z}_j)$ for $j \le k$, we obtain

$$t_k \le \sqrt{\kappa}, \qquad \frac{t_j}{t_k} \le \sqrt{\kappa}\, q^{-(k-j)/2}, \qquad 0 \le j \le k. \tag{B.30}$$

The identity $2Q((-a, \mathbf{1})^\top) = (Aa - B)(1+a)$ and $\lambda_{\min}(\boldsymbol{K}) = \kappa^{-1}$ give

$$(Aa - B)(1+a) \ge \frac{a^2 + r - 1}{\kappa}, \qquad Aa - B \ge \frac{r-1}{2\kappa}.$$

For each $\boldsymbol{z} \in \{(-a, \mathbf{1})^\top, ((D-B)/(A-B), \mathbf{1})^\top\}$, we have

$$\|\boldsymbol{z}\|_2 \le \sqrt{r}, \qquad |[\boldsymbol{K}\boldsymbol{z}]_1| = \sum_{j=2}^{r} [\boldsymbol{K}\boldsymbol{z}]_j \ge Aa - B,$$
$$[\boldsymbol{K}\boldsymbol{z}]_j \ge (Aa - B) w_{j-1} > 0, \qquad 2 \le j \le r.$$

The first-coordinate bound for the second vector uses $(AD - B^2)/(A - B) \ge Aa - B$; the tail bound follows from (B.6). Consequently, for each Muon iterate and its unit direction $\boldsymbol{D}_k$,

$$\sigma_r(\mathcal{H}[\boldsymbol{Z}_k/t_k]) \ge \frac{(Aa - B)\min_j w_j}{\sqrt{r}} \ge \frac{\nu}{2\kappa\sqrt{r}},$$
$$\langle \mathcal{H}[\boldsymbol{Z}_k/t_k], \boldsymbol{D}_k \rangle \ge \frac{2(Aa - B)}{r} \ge \frac{r-1}{\kappa r} \ge \frac{2}{3\kappa}. \tag{B.31}$$

For GD, $\|\mathcal{H}[\boldsymbol{Z}_k/t_k]\|_{\mathrm{F}} \ge \kappa^{-1}$. Thus both quadratic sequences satisfy the assumptions on $\boldsymbol{Z}_*$ and $\boldsymbol{D}_*$ in Lemma B.6.

Let $t = \|\boldsymbol{W}_0 - \boldsymbol{W}_\star\|_{\mathrm{F}}$. On the event of Lemma B.4, the target factor interpolates the data and (B.12) controls the full Hessian relative to $\mathcal{H}_0$. Moreover, (B.19) gives

$$m_4 := \frac{1}{N}\sum_t \|\boldsymbol{x}_t\|_2^4 \le 16\left(\mu^2 + \frac{\gamma}{1-\rho}\right)^2.$$

Differentiation gives

$$\nabla^2 \mathcal{L}_N(\boldsymbol{W})[\boldsymbol{D}] = \frac{1}{N}\sum_t \left\{4\langle \boldsymbol{x}_t\boldsymbol{x}_t^\top \boldsymbol{W}, \boldsymbol{D}\rangle \boldsymbol{x}_t\boldsymbol{x}_t^\top \boldsymbol{W} + 2(f_{\boldsymbol{W}}(\boldsymbol{x}_t) - y_t)\boldsymbol{x}_t\boldsymbol{x}_t^\top \boldsymbol{D}\right\}.$$

For any $\boldsymbol{W}_\star$, expanding the preceding identity at $\boldsymbol{W}_\star + \boldsymbol{Z}$ and applying Cauchy–Schwarz, we have

$$\begin{aligned}\|\nabla^2 \mathcal{L}_N(\boldsymbol{W}_\star + \boldsymbol{Z}) - \nabla^2 \mathcal{L}_N(\boldsymbol{W}_\star)\|_{\mathrm{op}} &\le \frac{1}{N}\sum_{t=1}^{N} \|\boldsymbol{x}_t\|_2^4 \big(12\|\boldsymbol{W}_\star\|_{\mathrm{op}}\|\boldsymbol{Z}\|_{\mathrm{F}} + 6\|\boldsymbol{Z}\|_{\mathrm{F}}^2\big) \\ &= 12 m_4 \|\boldsymbol{W}_\star\|_{\mathrm{op}}\|\boldsymbol{Z}\|_{\mathrm{F}} + 6 m_4 \|\boldsymbol{Z}\|_{\mathrm{F}}^2. \end{aligned} \tag{B.32}$$

At an exact fit, $\nabla^2 \mathcal{L}_N(\boldsymbol{W}_\star) = \mathcal{H}_N$.

Integrating (B.32) once and twice around the exact fit gives

$$\|\nabla \mathcal{L}_N(\boldsymbol{W}_\star + \boldsymbol{Z}) - \mathcal{H}_N[\boldsymbol{Z}]\|_{\mathrm{F}} \le 6 m_4 \|\boldsymbol{W}_\star\|_{\mathrm{op}}\|\boldsymbol{Z}\|_{\mathrm{F}}^2 + 2 m_4 \|\boldsymbol{Z}\|_{\mathrm{F}}^3,$$
$$\big|\mathcal{L}_N(\boldsymbol{W}_\star + \boldsymbol{Z}) - \tfrac{1}{2}\langle \boldsymbol{Z}, \mathcal{H}_N[\boldsymbol{Z}]\rangle\big| \le 2 m_4 \|\boldsymbol{W}_\star\|_{\mathrm{op}}\|\boldsymbol{Z}\|_{\mathrm{F}}^3 + \tfrac{1}{2} m_4 \|\boldsymbol{Z}\|_{\mathrm{F}}^4.$$

For a fixed method, define

$$F_k(\boldsymbol{Z}) = \frac{\mathcal{L}_N(\boldsymbol{W}_\star + t t_k \boldsymbol{Z})}{L t^2 t_k^2}.$$

For $\|\boldsymbol{Z}\|_{\mathrm{F}} \le 4\sqrt{\kappa}$, the preceding bounds give

$$|F_k(\boldsymbol{Z}) - Q(\boldsymbol{Z})| \le \frac{\|\mathcal{H}_N - \mathcal{H}_0\|_{\mathrm{op}}}{2L}\|\boldsymbol{Z}\|_{\mathrm{F}}^2 + \frac{m_4 t t_k}{L}\left(2\|\boldsymbol{W}_\star\|_{\mathrm{op}}\|\boldsymbol{Z}\|_{\mathrm{F}}^3 + \frac{t t_k}{2}\|\boldsymbol{Z}\|_{\mathrm{F}}^4\right),$$
$$\|\nabla F_k(\boldsymbol{Z}) - \mathcal{H}[\boldsymbol{Z}]\|_{\mathrm{F}} \le \frac{\|\mathcal{H}_N - \mathcal{H}_0\|_{\mathrm{op}}}{L}\|\boldsymbol{Z}\|_{\mathrm{F}} + \frac{m_4 t t_k}{L}\left(6\|\boldsymbol{W}_\star\|_{\mathrm{op}}\|\boldsymbol{Z}\|_{\mathrm{F}}^2 + 2 t t_k \|\boldsymbol{Z}\|_{\mathrm{F}}^3\right),$$
$$\|\nabla^2 F_k(\boldsymbol{Z}) - \mathcal{H}\|_{\mathrm{op}} \le \frac{\|\mathcal{H}_N - \mathcal{H}_0\|_{\mathrm{op}}}{L} + \frac{m_4 t t_k}{L}\left(12\|\boldsymbol{W}_\star\|_{\mathrm{op}}\|\boldsymbol{Z}\|_{\mathrm{F}} + 6 t t_k \|\boldsymbol{Z}\|_{\mathrm{F}}^2\right).$$

Using (B.12), $t \le \varepsilon \min_j \sigma_j$, $t_k \le \sqrt{\kappa}$, the bound for m_4, and the definition of v in (B.26), we obtain

$$\max\left\{|F_k(\boldsymbol{Z}) - Q(\boldsymbol{Z})|,\ \|\nabla F_k(\boldsymbol{Z}) - \mathcal{H}[\boldsymbol{Z}]\|_{\mathrm{F}},\ \|\nabla^2 F_k(\boldsymbol{Z}) - \mathcal{H}\|_{\mathrm{op}}\right\} \le v,$$

For either method, let

$$e_k = \frac{\|(\boldsymbol{W}_k - \boldsymbol{W}_\star)/t - \boldsymbol{Z}_k\|_{\mathrm{F}}}{t_k}.$$

Condition (4.1) gives $e_0 \le \varepsilon$. Applying the first inequality in (B.23) to F_k at $(\boldsymbol{W}_k - \boldsymbol{W}_\star)/(tt_k)$ gives

$$e_{k+1} \le \frac{32\kappa^2}{\nu}\frac{t_k}{t_{k+1}}(e_k + v).$$

Iterating this inequality and using (B.30) with $\Lambda = 32\kappa^2/(\nu\sqrt{q}) \ge 128$ gives

$$\begin{aligned} e_k + v &\le \left(\frac{32\kappa^2}{\nu}\right)^k \frac{t_0}{t_k}\varepsilon + v\sum_{j=0}^{k-1}\left(\frac{32\kappa^2}{\nu}\right)^{k-j}\frac{t_j}{t_k} + v \\ &\le \sqrt{\kappa}\Lambda^k\varepsilon + \sqrt{\kappa}v\sum_{j=1}^{k}\Lambda^j + v \\ &\le \sqrt{\kappa}\Lambda^k(\varepsilon + 2v) = \sqrt{\kappa}\Lambda^k u, \qquad k \ge 0. \end{aligned} \tag{B.33}$$

provided $e_j + v \le c_1\nu/\kappa^3$ for $0 \le j < k$. For $k < T$, (B.28) and (B.33) give

$$e_k + v \le c_2\frac{\nu\min\{g,q\}}{\kappa^3}.$$

Choose $c_2 \le c_1$. Then $e_k + v \le c_1\nu/\kappa^3$, so Lemma B.6 applies at iteration k. Starting with $e_0 \le \varepsilon$, induction proves that all iterates through T exist and that (B.33) holds. The second inequality in (B.23) gives, for $\mathrm{A} \in \{\mathrm{M}, \mathrm{G}\}$ and $0 \le k < T$,

$$\left|\frac{\mathcal{L}_N(\boldsymbol{W}_{k+1}^{\mathrm{A}})}{\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{A}})} - \frac{Q(\boldsymbol{Z}_{k+1}^{\mathrm{A}})}{Q(\boldsymbol{Z}_k^{\mathrm{A}})}\right| \le 160c_2\min\{g,q\} \le \frac{\min\{g,q\}}{4}, \tag{B.34}$$

where the last inequality follows by choosing $c_2 \le 1/640$. Therefore

$$\frac{\mathcal{L}_N(\boldsymbol{W}_{k+1}^{\mathrm{G}})}{\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{G}})} - \frac{\mathcal{L}_N(\boldsymbol{W}_{k+1}^{\mathrm{M}})}{\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{M}})} \ge (q+g) - q - \frac{g}{4} - \frac{g}{4} = \frac{g}{2}.$$

Moreover, each empirical quotient is at least $q - q/4 > 0$. This proves (4.4) and positivity of the losses through iteration T. At every iterate, the line-search derivative satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{A}} - \alpha\mathcal{D}_{\mathrm{A}}(\nabla\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{A}})))\Big|_{\alpha=0} = -\langle\nabla\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{A}}), \mathcal{D}_{\mathrm{A}}(\nabla\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{A}}))\rangle < 0.$$

The positive line curvature in Lemma B.6 then gives $\mathcal{L}_N(\boldsymbol{W}_{k+1}^{\mathrm{A}}) < \mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{A}})$. Therefore each GD ratio is at most one, and

$$\frac{\mathcal{L}_N(\boldsymbol{W}_{k+1}^{\mathrm{M}})/\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{M}})}{\mathcal{L}_N(\boldsymbol{W}_{k+1}^{\mathrm{G}})/\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{G}})} \le 1 - \frac{g}{2}\frac{\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{G}})}{\mathcal{L}_N(\boldsymbol{W}_{k+1}^{\mathrm{G}})} \le 1 - g/2.$$

The initial losses agree. Multiplication proves (4.5).

For Remark 4.2, combine (B.33), (B.29), and (B.27). With $\mathfrak{R} = \kappa_{\mathrm{act}}/\eta \ge 32$, for either method and $0 \le k \le T$ we obtain

$$e_k \le \sqrt{\kappa}\Lambda^k u \le (16C_0 + 4096)\mathfrak{R}^{10k+39/2}(\varepsilon + \varepsilon_N + \delta) \le \mathfrak{R}^{C(k+1)}(\varepsilon + \varepsilon_N + \delta).$$

The last inequality holds for $C \ge 20 + \log(16C_0 + 4096)/\log 32$. Since $\boldsymbol{Z}_k = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{z}_k)\boldsymbol{V}^\top$ and $t_k = \|\boldsymbol{z}_k\|_2$, the definition of e_k gives the displayed estimate in the remark. This completes the proof. □

The dependence on the sample size can also be read directly from (4.3). If $\varepsilon + \delta \le \varepsilon_N$, the condition

$$N \ge C_0 dr\log^3\frac{dr}{\zeta}\left(\frac{\kappa_{\mathrm{act}}}{\eta}\right)^{2C_1T}$$

gives $\varepsilon + \varepsilon_N + \delta \le 2\varepsilon_N \le (\eta/\kappa_{\mathrm{act}})^{C_1T}$ and hence (4.3). For fixed model parameters and ζ, this yields a comparison interval logarithmic in N while $\varepsilon + \delta \le \varepsilon_N$.

B.7 Proof of Theorem 4.3

We give the coordinate realization used in Theorem 4.3. Let $m \geq 2$, $d = m + r - 1$, and let r admit an orthogonal matrix whose entries have magnitude $r^{-1/2}$. With $\boldsymbol{e}_i$ denoting the ith standard basis vector of $\mathbb{R}^d$, take

$$\boldsymbol{u}_1 = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \boldsymbol{e}_i, \qquad \boldsymbol{u}_j = \boldsymbol{e}_{m+j-1} \ (2 \leq j \leq r), \qquad |V_{ij}| = \frac{1}{\sqrt{r}}, \quad \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_r. \tag{B.35}$$

Complete $\boldsymbol{U}$ to $\boldsymbol{A}_0$ and choose the mean direction so that $j_\star = 1$. For $\boldsymbol{g} \in \mathbb{R}^r$ with $g_j \neq 0$ for all j, these coordinates give

$$\operatorname{sign}(\boldsymbol{U} \operatorname{diag}(\boldsymbol{g}) \boldsymbol{V}^\top) = \sqrt{r}\, \boldsymbol{U} \operatorname{diag}\big(\sqrt{m} \operatorname{sign}(g_1), \operatorname{sign}(g_2), \ldots, \operatorname{sign}(g_r)\big)\, \boldsymbol{V}^\top. \tag{B.36}$$

We first compute the quadratic SignGD iterates, then bound the smallest gradient entry to compare the empirical updates.

Proof. For the coordinates in (B.35), the first m rows of $\boldsymbol{U} \operatorname{diag}(\boldsymbol{g}) \boldsymbol{V}^\top$ have entry magnitude $|g_1|/\sqrt{mr}$; each remaining row has entry magnitude $|g_j|/\sqrt{r}$ for one $j \geq 2$. Thus (B.36) applies when all components of $\boldsymbol{g}$ are nonzero. Consequently, in the coordinates $\boldsymbol{z}$, SignGD uses $(\sqrt{m} \operatorname{sign}([\boldsymbol{K}\boldsymbol{z}]_1), \operatorname{sign}([\boldsymbol{K}\boldsymbol{z}]_2), \ldots, \operatorname{sign}([\boldsymbol{K}\boldsymbol{z}]_r))^\top$. Put $h = \sqrt{m}$ and

$$q_h = \frac{(Ah^2 - D)^2}{(Ah^2 + D)^2 - 4h^2 B^2}.$$

For $h \geq 1$, set

$$\boldsymbol{d}_-^{(h)} = (-h, \mathbf{1})^\top, \qquad \boldsymbol{d}_+^{(h)} = (h, \mathbf{1})^\top.$$

Define

$$\boldsymbol{z}_+^{(h)} = (D - hB, (hA - B)\mathbf{1})^\top, \qquad \boldsymbol{z}_-^{(h)} = (-D - hB, (hA + B)\mathbf{1})^\top.$$

Starting at $(-a, \mathbf{1})^\top$, exact line minimization along $-\boldsymbol{d}_-^{(h)}$ gives

$$(-a, \mathbf{1})^\top \longmapsto \frac{h - a}{Ah^2 - 2hB + D} \boldsymbol{z}_+^{(h)}.$$

Since $0 < a < 1$, the multiplying factor is positive. Direct multiplication gives $[\boldsymbol{K}\boldsymbol{z}_+^{(h)}]_1 = AD - B^2$ and $[\boldsymbol{K}\boldsymbol{z}_-^{(h)}]_1 = -(AD - B^2)$. Moreover,

$$\frac{D - hB}{hA - B} + \frac{B}{A} = \frac{AD - B^2}{A(hA - B)} > 0, \qquad \frac{\mathrm{d}}{\mathrm{d}h} \frac{D + hB}{hA + B} = \frac{B^2 - AD}{(hA + B)^2} < 0.$$

The first inequality and $a > B/A$ give $(D - hB)/(hA - B) > -a$. The derivative identity and $(D + B)/(A + B) = a$ give $-(D + hB)/(hA + B) \geq -a$. Since $hA \pm B > 0$, equation (B.6) yields

$$[\boldsymbol{K}\boldsymbol{z}_\pm^{(h)}]_j \geq (hA \mp B)[\boldsymbol{K}(-a, \mathbf{1})^\top]_j > 0, \qquad 2 \leq j \leq r.$$

The directions are therefore $\boldsymbol{d}_+^{(h)}$ and $\boldsymbol{d}_-^{(h)}$ at $\boldsymbol{z}_+^{(h)}$ and $\boldsymbol{z}_-^{(h)}$, respectively. The exact line-search formula gives

$$\boldsymbol{z}_+^{(h)} \longmapsto \frac{Ah^2 - D}{Ah^2 + 2hB + D} \boldsymbol{z}_-^{(h)}, \qquad \boldsymbol{z}_-^{(h)} \longmapsto \frac{Ah^2 - D}{Ah^2 - 2hB + D} \boldsymbol{z}_+^{(h)}. \tag{B.37}$$

Since $2Q(\boldsymbol{z}_\pm^{(h)}) = (AD - B^2)(Ah^2 + D \mp 2hB)$, $Q(\boldsymbol{z}_{k+1}^{\mathrm{S}})/Q(\boldsymbol{z}_k^{\mathrm{S}}) = q_h$ for every $k \geq 1$. At $k = 0$, the preceding first-iterate formula gives

$$\frac{Q(\boldsymbol{z}_1^{\mathrm{S}})}{Q(\boldsymbol{z}_0)} = \frac{(A + B)(h - a)^2}{(1 + a)(Ah^2 - 2Bh + D)}. \tag{B.38}$$

Here the initial normalization cancels in the ratio.

To bound the difference from q, the formulas above give

$$\frac{Q(\boldsymbol{z}_1^{\mathrm{S}})/Q(\boldsymbol{z}_0)-q}{1-q}=1-\frac{(h+1)^2(A-2B+D)}{4(Ah^2-2Bh+D)}\geq\frac{(h-1)^2}{2(h^2+1)},$$
$$\frac{q_h-q}{1-q}=\frac{(h^2-1)(A^2h^2-D^2)}{(Ah^2+D)^2-4h^2B^2}\geq\frac{(h^2-1)^2}{(h^2+1)^2}.$$

For the first inequality, the numerator after subtraction is $(h-1)[A(3h+1)+2B(h-1)-D(h+3)]\geq 2A(h-1)^2$, and the denominator is at most $4A(h^2+1)$. For the second inequality,

$$A^2h^2-D^2\geq A^2(h^2-1),\qquad (Ah^2+D)^2-4h^2B^2\leq A^2(h^2+1)^2.$$

Both lower bounds increase for $h\geq 1$. Since $h=\sqrt{m}\geq\sqrt{2}$ and $\sqrt{2}-1>1/3$, we obtain

$$g_{\mathrm{S}}:=\min\left\{\frac{Q(\boldsymbol{z}_1^{\mathrm{S}})}{Q(\boldsymbol{z}_0)}-q,\ q_{\sqrt{m}}-q\right\}\geq\frac{1-q}{54}>0.$$

Use L,κ,ν,Λ,v,u from Appendix B.6. Normalize the reference loss by L as in the proof of Theorem 4.1. Since $Q(\boldsymbol{Z}_{k+1}^{\mathrm{S}})/Q(\boldsymbol{Z}_k^{\mathrm{S}})\geq q$, (B.30) holds for the SignGD sequence as well as for the Muon sequence.

Divide $\boldsymbol{z}_-^{(h)}$ and $\boldsymbol{z}_+^{(h)}$ by $hA+B$ and $hA-B$, respectively. The resulting vectors are

$$\left(-\frac{D+hB}{hA+B},\mathbf{1}\right)^\top,\qquad\left(\frac{D-hB}{hA-B},\mathbf{1}\right)^\top.$$

The first coordinates satisfy

$$-a\leq-\frac{D+hB}{hA+B}<1,\qquad -a<\frac{D-hB}{hA-B}<1,$$
$$\left\|\frac{\boldsymbol{z}_-^{(h)}}{hA+B}\right\|_2\leq\sqrt{r},\qquad\left\|\frac{\boldsymbol{z}_+^{(h)}}{hA-B}\right\|_2\leq\sqrt{r}.$$

The first row follows from the inequalities preceding (B.37). For either $\boldsymbol{z}\in\{\boldsymbol{z}_-^{(h)}/(hA+B),\ \boldsymbol{z}_+^{(h)}/(hA-B)\}$,

$$|[\boldsymbol{K}\boldsymbol{z}]_1|\geq\frac{Aa-B}{h},\qquad[\boldsymbol{K}\boldsymbol{z}]_j\geq(Aa-B)w_{j-1},\qquad 2\leq j\leq r.$$

The first inequality follows from $(AD-B^2)/(hA\pm B)\geq(AD-B^2)/(h(A+B))$; the second follows from (B.6). Using $Aa-B\geq(r-1)/(2\kappa)$ and (B.36), we obtain, both at $\boldsymbol{Z}_*=\boldsymbol{Z}_{\mathrm{bal}}$ and at every normalized quadratic SignGD iterate $\boldsymbol{Z}_*$,

$$\min_{1\leq i\leq d,\,1\leq j\leq r}|[\mathcal{H}[\boldsymbol{Z}_*]]_{ij}|\geq\frac{Aa-B}{r}\min\left\{\frac{1}{m},\min_j w_j\right\}\geq\frac{\nu}{2\kappa mr}.$$

Writing $\boldsymbol{G}=\mathcal{H}[\boldsymbol{Z}_*]$ and $\boldsymbol{D}=\mathrm{sign}(\boldsymbol{G})/\sqrt{dr}$, it follows that

$$\langle\boldsymbol{G},\boldsymbol{D}\rangle=\frac{\sum_{i=1}^d\sum_{j=1}^r|G_{ij}|}{\sqrt{dr}}\geq\min_{i,j}|G_{ij}|\geq\frac{\nu}{2\kappa mr}.$$

Let $F,\boldsymbol{Z},\boldsymbol{Z}_*,e,v$ satisfy the value, gradient, Hessian, and distance bounds in Lemma B.6. Suppose $e+v\leq\nu/(16\kappa^2mr)$. Since $\boldsymbol{G}=\nabla Q(\boldsymbol{Z}_*)$ and $\|\mathcal{H}\|_{\mathrm{op}}\leq 1$,

$$\max_{1\leq i\leq d,\,1\leq j\leq r}|[\nabla F(\boldsymbol{Z})-\boldsymbol{G}]_{ij}|\leq\|\nabla F(\boldsymbol{Z})-\boldsymbol{G}\|_{\mathrm{F}}\leq e+v<\frac{\nu}{2\kappa mr}\leq\min_{i,j}|G_{ij}|.$$

Hence $\mathrm{sign}(\nabla F(\boldsymbol{Z}))=\mathrm{sign}(\boldsymbol{G})$, so the two unit SignGD directions both equal $\boldsymbol{D}$. Define $\alpha_*=\langle\boldsymbol{G},\boldsymbol{D}\rangle/\langle\boldsymbol{D},\mathcal{H}[\boldsymbol{D}]\rangle$. Since $\boldsymbol{D}\in\mathcal{S}$, $\nu/(2\kappa mr)\leq\alpha_*\leq\sqrt{\kappa}$. Therefore, for $0\leq\alpha\leq 2\alpha_*$ we have $\|\boldsymbol{Z}-\alpha\boldsymbol{D}\|_{\mathrm{F}}\leq 1+e+2\sqrt{\kappa}<4\sqrt{\kappa}$. Thus the hypotheses on $F-Q$ hold at every point of this

segment. Let $\widehat{\alpha} = \langle \mathcal{H}[\boldsymbol{Z}], \boldsymbol{D}\rangle / \langle \boldsymbol{D}, \mathcal{H}[\boldsymbol{D}]\rangle$. The interval of radius $2\kappa v$ about $\widehat{\alpha}$ lies in $[\alpha_*/2, 3\alpha_*/2]$, since

$$|\widehat{\alpha} - \alpha_*| + 2\kappa v \le \frac{\|\mathcal{H}[\boldsymbol{D}]\|_{\mathrm{F}}}{\langle \boldsymbol{D}, \mathcal{H}[\boldsymbol{D}]\rangle} e + 2\kappa v \le \sqrt{\kappa}\, e + 2\kappa v \le 2\kappa(e+v) \le \frac{\nu}{8\kappa m r} \le \frac{\alpha_*}{4}.$$

Here the second inequality uses $\|\mathcal{H}[\boldsymbol{D}]\|_{\mathrm{F}}^2 \le \langle \boldsymbol{D}, \mathcal{H}[\boldsymbol{D}]\rangle$ and $\langle \boldsymbol{D}, \mathcal{H}[\boldsymbol{D}]\rangle \ge \kappa^{-1}$. For $0 \le \alpha \le 2\alpha_*$, $\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2} F(\boldsymbol{Z} - \alpha \boldsymbol{D}) \ge (2\kappa)^{-1}$. At $\widehat{\alpha}$, the first derivative has absolute value at most v. Therefore

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} F(\boldsymbol{Z} - \alpha \boldsymbol{D})\Big|_{\alpha = \widehat{\alpha} - 2\kappa v} \le 0 \le \frac{\mathrm{d}}{\mathrm{d}\alpha} F(\boldsymbol{Z} - \alpha \boldsymbol{D})\Big|_{\alpha = \widehat{\alpha} + 2\kappa v}.$$

The derivative is strictly increasing on $[0, 2\alpha_*]$. Its first positive zero α therefore satisfies $|\alpha - \widehat{\alpha}| \le 2\kappa v$, and

$$\begin{aligned}
\|\boldsymbol{Z}_+ - \boldsymbol{Z}_{*,+}\|_{\mathrm{F}} &\le e + |\alpha_* - \widehat{\alpha}| + |\widehat{\alpha} - \alpha| \le 3\kappa(e+v) \le \frac{1}{32\kappa}, \\
|F(\boldsymbol{Z}_+) - Q(\boldsymbol{Z}_{*,+})| &\le v + \|\boldsymbol{Z}_+ - \boldsymbol{Z}_{*,+}\|_{\mathrm{F}} + \tfrac{1}{2}\|\boldsymbol{Z}_+ - \boldsymbol{Z}_{*,+}\|_{\mathrm{F}}^2 \le 5\kappa(e+v), \\
\left| \frac{F(\boldsymbol{Z}_+)}{F(\boldsymbol{Z})} - \frac{Q(\boldsymbol{Z}_{*,+})}{Q(\boldsymbol{Z}_*)} \right| &\le 4\kappa[5\kappa(e+v) + 2(e+v)] \le 32\kappa^2(e+v).
\end{aligned}$$

The last line uses $|F(\boldsymbol{Z}) - Q(\boldsymbol{Z}_*)| \le 2(e+v)$, $F(\boldsymbol{Z}) \ge (4\kappa)^{-1}$, and $Q(\boldsymbol{Z}_{*,+}) \le Q(\boldsymbol{Z}_*)$.

Since $3\kappa \le 32\kappa^2/\nu$, the recurrence used in (B.33) applies to SignGD as well. It suffices to have

$$u \le c_2 \frac{\nu}{\kappa^{7/2}} \min\left\{ g_{\mathrm{S}}, q, \frac{\kappa}{mr} \right\} \Lambda^{-(T-1)}, \tag{B.39}$$

with the universal constant c_2 chosen sufficiently small as in the proof of Theorem 4.1. In fact, for $k < T$ the recurrence yields

$$e_k + v \le c_2 \frac{\nu}{\kappa^3} \min\left\{ g_{\mathrm{S}}, q, \frac{\kappa}{mr} \right\}.$$

Choosing $c_2 \le \min\{c_1, 1/16\}$ gives both $e_k + v \le c_1 \nu/\kappa^3$ and $e_k + v \le \nu/(16\kappa^2 mr)$. Thus the preceding estimates apply at every iteration by induction. Write e_k^{M} and e_k^{S} for the corresponding quantities for the two methods. For $0 \le k < T$,

$$\begin{aligned}
\frac{\mathcal{L}_N(\boldsymbol{W}_{k+1}^{\mathrm{S}})}{\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{S}})} - \frac{\mathcal{L}_N(\boldsymbol{W}_{k+1}^{\mathrm{M}})}{\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{M}})} &\ge \frac{Q(\boldsymbol{Z}_{k+1}^{\mathrm{S}})}{Q(\boldsymbol{Z}_k^{\mathrm{S}})} - q - 32\kappa^2(e_k^{\mathrm{S}} + v) - \frac{160\kappa^3}{\nu}(e_k^{\mathrm{M}} + v) \\
&\ge g_{\mathrm{S}} - 176 c_2 \min\{g_{\mathrm{S}}, q\} \ge \frac{g_{\mathrm{S}}}{2} > 0,
\end{aligned}$$

where the last inequality holds when $c_2 \le 1/352$. Each empirical quotient is also at least $q - 176 c_2 q \ge q/2 > 0$. Hence all losses remain positive through T, and (4.7) follows.

It remains to verify (B.39). With $\mathfrak{R} = \kappa_{\mathrm{act}}/\eta$, Lemma B.7 gives $g_{\mathrm{S}} \ge \mathfrak{R}^{-6}$, $q \ge \mathfrak{R}^{-2}$, and $\Lambda \le \mathfrak{R}^{10}$. Since $\kappa > 2$ and $mr \ge 1$, we have $\min\{g_{\mathrm{S}}, q, \kappa/(mr)\} \ge \mathfrak{R}^{-6}/(mr)$. Hence

$$u \le \frac{(16C_0 + 4096)\mathfrak{R}^{18 - C_1 T}}{mr} \le \frac{c_2 \mathfrak{R}^{-10T-9}}{mr} \le c_2 \frac{\nu}{\kappa^{7/2}} \min\left\{ g_{\mathrm{S}}, q, \frac{\kappa}{mr} \right\} \Lambda^{-(T-1)}.$$

The first inequality follows from (B.29) and (4.6); the second holds by increasing the universal C_1; the last follows from Lemma B.7. Since the initial losses are equal, multiplying (4.7) over $0 \le j < k$ gives $\mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{M}}) < \mathcal{L}_N(\boldsymbol{W}_k^{\mathrm{S}})$ for $1 \le k \le T$. Finally, $mr \ge 1$ shows that (4.6) implies (4.3); Theorem 4.1 therefore applies on the same iteration interval. □