NORMALIZE-THEN-PRECONDITION: A HIERARCHICAL APPROACH TO MARGINAL SCALE AND INTERACTION GEOMETRY FOR LLM TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Matrix optimizers have emerged as a promising direction, with Muon standing out as a prominent design. Revisiting Muon through its full-Gram representation, we observe that it jointly processes marginal-scale and interaction information. This opens an alternative way to organize geometric information hierarchically, motivating the ***Normalize-Then-Precondition*** framework. Specifically, it first uses diagonal-Gram information to construct a marginally normalized update, then applies spectral preconditioning to its directional interaction geometry. Building on this framework, we develop **NormPre** with NormPre-G and NormPre-L adopting global and localized spectral preconditioning, grounded in spectral-norm steepest descent and a regularized formulation followed by leading mode selection, respectively. To enable large-scale training, NormPre-G uses Newton-Schulz iterations and NormPre-L employs randomized sketching to approximate the leading interaction eigenspace. Theoretically, we establish $\mathcal{O}(T^{-1/2})$ convergence guarantees for simplified versions of NormPre. Across extensive pretraining experiments on GPT-2 Small, LLaMA and Qwen3, both variants consistently outperform AdamW, Muon and MANO under matched training budgets. Further efficiency and spectral analyses reveal the complementary strengths of two variants and characterize their performance-efficiency trade-off. We open-source our code through an anonymous GitHub repository at https://anonymous.4open.science/r/NormPre.

1 INTRODUCTION

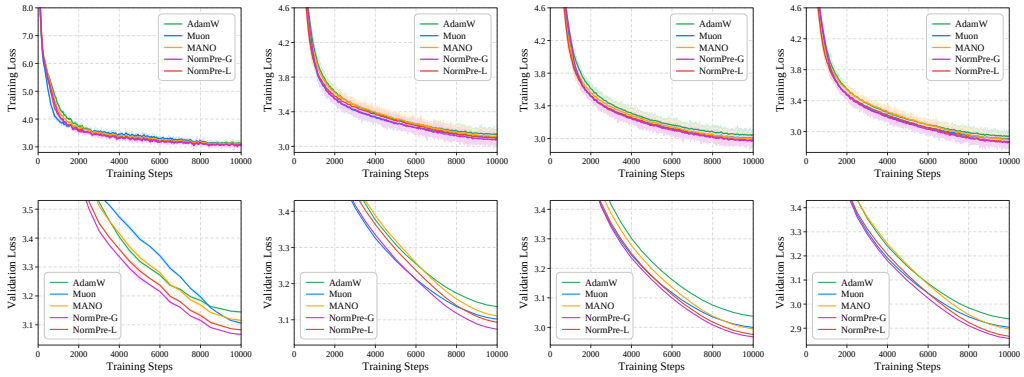
Large language model (LLM) training places increasing demands on optimizer performance, efficiency, and scalability. Recent optimizers address these demands by exploiting richer parameter structure, from individual coordinates (Kingma & Ba, 2015; Loshchilov & Hutter, 2019; Chen et al., 2023; Liu et al., 2024; Liang et al., 2026) and parameter blocks (Zhang et al., 2025; Wang et al., 2025) to full matrices (Gupta et al., 2018; Vyas et al., 2025; Jordan et al., 2024; Liu et al., 2025; Gu & Xie, 2026; Deng et al., 2026; Xu et al., 2026). More specifically, orthogonalization and normalization have emerged as prominent matrix-level techniques. On the orthogonalization front, Muon applies a matrix-sign transformation to momentum via Newton-Schulz iterations for rapid convergence (Jordan et al., 2024; Liu et al., 2025). Alternatively, other recent methods use row or column normalization to construct matrix updates. For example, MANO combines momentum projection with alternating normalization (Gu & Xie, 2026), and MOGA derives such normalized updates from mean-normalized operator norms under a steepest-descent perspective (Xu et al., 2026).

Orthogonalization and normalization may appear to follow different routes, yet their Gram representations reveal a shared structure. For a matrix update X, the transformations take the forms

$$\underbrace{\Phi = (\mathbf{X}\mathbf{X}^\top)^{-1/2}X}_{\text{Full-Gram Representation}}, \quad \underbrace{\Psi = \left[\mathbf{Diag(diag}(\mathbf{X}\mathbf{X}^\top)\mathbf{)}\right]^{-1/2}X}_{\text{Diagonal-Gram Representation}},$$

where $\mathrm{Diag}(\mathrm{diag}(\cdot))$ retains only the diagonal entries, and the inverse-square-root expressions require full row rank for Φ and nonzero rows for Ψ, respectively. Muon jointly processes marginal scales and cross-row interactions encoded in the full Gram matrix, while normalization draws solely on the Gram diagonal to equalize row norms without altering pairwise angles. As an extreme example,

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.



(a) GPT-2 Small on OWT (b) LLaMA-130M on C4 (c) LLaMA-350M on C4 (d) LLaMA-1.3B on C4

Figure 1: **GPT-2 Small on OpenWebText and LLaMA Models on C4.** Training and validation loss for GPT-2 Small and LLaMA-{130M, 350M, 1.3B} with AdamW, Muon, MANO, NormPre-G and NormPre-L. Dark and light curves denote 50-step moving averages and raw training trajectories.

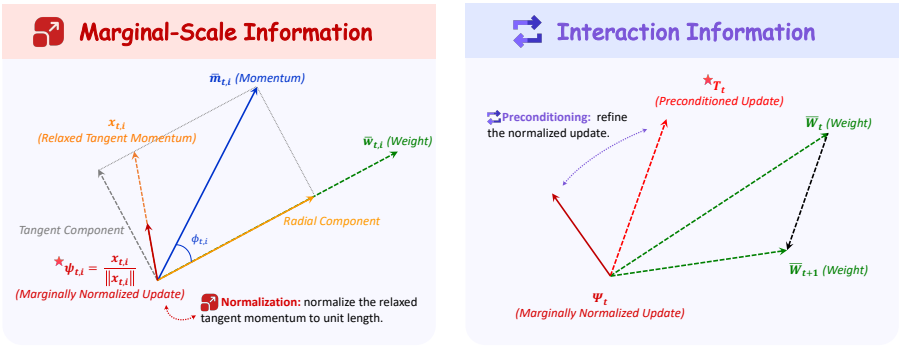
consider fixed row directions with one row norm of 10^4 and all others of 10^{-3}. As this scale gap increases further, the leading Gram eigenvector approaches the largest row's coordinate axis, even if the other rows are strongly correlated. The principal mode thus becomes dominated by marginal scales despite unchanged directional interactions. By removing this scale weighting, diagonal-Gram normalization provides a reference update for processing directional interactions. Together, these observations open an alternative way to organize the available geometric information:

Can we construct a normalized update from scales and refine it via directional interactions?

While recent methods apply normalization to gradients or momentum to improve the numerical conditioning of orthogonalization (Ma et al., 2024; Chang et al., 2026), *a systematic framework* for formalizing the hierarchical organization has yet to be established. In particular, it remains unclear *whether the normalized update should undergo a full-spectrum transformation or serve as a reference for more targeted spectral modifications*.

To address this gap, we formalize the hierarchical organization as the **Normalize-Then-Precondition** framework. The marginally normalized update retains a directional interaction spectrum encoded by $\Gamma = \Psi\Psi^\top$. To exploit this spectrum, one choice is *global spectral preconditioning*, which performs orthogonalization of Ψ and solves a steepest descent problem under the spectral norm constraint. Another choice is to treat the normalized update Ψ as a reference and minimize the changes needed to satisfy the spectral constraint, seeking the closest update within the unit spectral norm ball. This preference is formalized by regularized spectral steepest descent, whose solution clips singular values above one and leaves the others unchanged. A spectral budget r further restricts the transformation to at most r leading super-unit modes, giving *localized spectral preconditioning*. Here, r controls the selected subspace dimension for spectral transformation and the final update remains dense.

Building on this framework, we propose NormPre, a family of matrix optimizers comprising NormPre-G and NormPre-L, which realize marginal normalization followed by global and localized spectral preconditioning, respectively. Both variants combine relaxed tangent momentum processing with alternating row/column normalization and consistent update RMS scaling. Specifically, NormPre-G approximates the full-spectrum transformation via Newton-Schulz iterations. For NormPre-L, the leading interaction eigenspace is obtained through exact extraction or scalable randomized sketching. We also establish $\mathcal{O}(T^{-1/2})$ convergence guarantees for simplified NormPre-G and exact NormPre-L, extending them to sketch-based NormPre-L under controlled approximation error. Empirically, our extensive pretraining evaluations span GPT-2 Small on OpenWebText, LLaMA models up to 1.3B on C4, and Qwen3 models up to 1.7B on Pile. Under matched training budgets, both variants consistently outperform established baselines that represent distinct optimization paradigms: coordinate-wise adaptation (AdamW), matrix orthogonalization (Muon), and normalization (MANO). Furthermore, our spectral dynamics and efficiency analyses highlight that NormPre-L with localized precondi-



(a) Normalize: Diagonal-Gram Normalization

(b) Precondition: Spectral Interaction Preconditioning

Figure 2: Geometric Workflow of the NormPre Optimizer (Algorithm 1).

tioning delivers strong optimization gains at lower cost and NormPre-G with global preconditioning further reduces validation loss, providing practitioners with flexible choices to balance optimization performance and computational cost.

Our main contributions are summarized as follows.

(a) The Normalize-Then-Precondition Framework. ***Conceptually***, departing from joint geometric transformations, we establish a systematic framework to hierarchically construct a normalized update from marginal scales and refine it via interaction information. Unifying full-spectrum and targeted schemes, we formulate global and localized spectral preconditioning as a solution to spectral-norm steepest descent and a regularized formulation with leading mode selection, respectively (Section 2).

(b) The NormPre Optimizer. ***Algorithmically***, we instantiate this framework as the NormPre optimizer, featuring alternating row-column normalization and spectral preconditioning. We provide scalable implementations using Newton-Schulz iterations for global preconditioning (NormPre-G) and randomized sketching for localized eigenspace extraction (NormPre-L), together with theoretical $\mathcal{O}(T^{-1/2})$ convergence guarantees for simplified variants (Theorem 1 and Corollary 1).

(c) Optimization Performance and Trade-offs. ***Empirically***, we demonstrate the superior optimization performance of NormPre across the evaluated architectures and model scales, consistently outperforming established baselines (AdamW, Muon, MANO). Furthermore, we characterize the performance-efficiency trade-off between its two variants, providing flexible choices to balance optimization gains and computational cost (Section 4).

2 Geometric Foundations of Normalize-Then-Precondition

This section formalizes the Normalize-Then-Precondition framework. We begin by establishing a full-Gram view of row-orthogonalized updates (Section 2.1) and derive the normalized update with the marginal-scale information encoded in the Gram diagonal (Section 2.2). We finally formulate different spectral preconditioning realizations (Section 2.3). Throughout this section, $X \in \mathbb{R}^{m\times n}$ denotes a matrix-valued update signal. We write its thin SVD as $X = U\Sigma V^\top$. We present the row-side analysis and the column-side counterpart follows by applying the same construction to $X^\top$.

2.1 Orthogonalized Updates and Full-Gram Transformation

Muon (Jordan et al., 2024) orthogonalizes a momentum matrix X by replacing it with the nearest semi-orthogonal matrix. For $X \in \mathbb{R}^{m\times n}$ with $m \le n$, row orthogonalization is defined by

$$\Phi \in \arg\min_{O}\{\|O - X\|_F : OO^\top = I_m\}. \tag{1}$$

One solution is given by the matrix sign function, $\Phi = \mathrm{msign}(X) = UV^\top$. Accordingly, the matrix-sign update transforms the singular-value matrix Σ into identity, yielding $\Phi\Phi^\top = I_m$. In practice, Muon approximates this orthogonalization efficiently using Newton-Schulz iterations. We next identify an equivalent full-Gram form of the same update, showing how Muon's orthogonalization depends on the full row Gram matrix.

Full-Gram Representation of Muon. *If $XX^\top \succ 0$, the matrix-sign update admits the full-Gram representation*

$$\Phi = \mathrm{msign}(X) = (XX^\top)^{-1/2}X. \tag{2}$$

Let $x_i^\top$ denote the i-th row of X. The entries of the full row Gram matrix satisfy $(XX^\top)_{ii} = \|x_i\|_2^2$, and $(XX^\top)_{ij} = \langle x_i, x_j\rangle$ for $i \neq j$. Its diagonal entries encode the marginal scales of individual rows, whereas its off-diagonal entries encode their pairwise interactions. Together, these two components determine the inverse-square-root operator in Equation 2, which acts on the entire singular spectrum of X. Thus, Muon jointly utilizes marginal-scale and cross-row interaction information through the full row Gram matrix.

2.2 Diagonal-Gram Normalization

Within the joint geometry, marginal-scale information is structurally simpler, since each row scale depends only on its corresponding row. This motivates a natural question: *is it possible to first use this simpler marginal-scale information to construct a valid matrix update?* To answer this question, we isolate the marginal-scale information encoded in the diagonal of $XX^\top$. Since $(XX^\top)_{ii} = \|x_i\|_2^2$, define the diagonal matrix of row scales as

$$D(X) := \mathrm{Diag}\left(\mathrm{diag}(XX^\top)\right)^{1/2} = \mathrm{Diag}\left(\|x_1\|_2, \ldots, \|x_m\|_2\right),$$

where $\mathrm{diag}(\cdot)$ extracts the diagonal entries as a vector and $\mathrm{Diag}(\cdot)$ constructs a diagonal matrix.

Diagonal-Gram Representation of Marginal Normalization. *If X has no zero rows, the marginally normalized update admits the diagonal-Gram representation*

$$\Psi = D(X)^{-1}X = \mathrm{Diag}\left(\mathrm{diag}(XX^\top)\right)^{-1/2}X. \tag{3}$$

Equation 3 provides the diagonal-Gram counterpart of the Muon transformation in Equation 2. Muon constructs its orthogonalized update from the inverse square root of the full row Gram matrix, while Ψ uses only its diagonal component to normalize the marginal scales. This transformation coincides with the core normalization used by normalization-based optimizers (Gu & Xie, 2026; Xu et al., 2026; Deng et al., 2026; Pethick et al., 2025a; Glentis et al., 2025). Define the normalized interaction matrix as $\Gamma(X) := \Psi\Psi^\top$. By construction, $\Gamma(X)_{ii} = 1$ and its off-diagonal entries $\Gamma(X)_{ij} = \langle x_i, x_j\rangle/(\|x_i\|_2\|x_j\|_2)$ encode the cosine similarities between rows.

2.3 Spectral Interaction Preconditioning

Hierarchical Geometric Construction. Beyond serving as an effective standalone update, the normalization Ψ retains directional correlations in $\Gamma(X)$, providing a natural foundation for further spectral preconditioning. Having processed the marginal scales, we now study two realizations of this preconditioning stage (Figure 3): *global spectral preconditioning*, which performs full-spectrum transformation, and *localized spectral preconditioning*, which retains Ψ as a reference and concentrates the transformation on selected interaction modes.

Let $\Psi = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top$ be a thin SVD with singular values $\widetilde{\sigma}_1 \geq \cdots \geq \widetilde{\sigma}_m \geq 0$. Accordingly, $\Gamma(X) = \Psi\Psi^\top = \widetilde{U}\Lambda\widetilde{U}^\top$, where $\Lambda = \widetilde{\Sigma}^2 = \mathrm{Diag}(\lambda_1, \ldots, \lambda_m)$ and $\lambda_i = \widetilde{\sigma}_i^2$. The columns $\widetilde{u}_i$ of $\widetilde{U}$ define the corresponding spectral interaction directions.

Global Spectral Preconditioning. A direct method is to solve the spectral steepest descent problem for the normalized update Ψ (Bernstein & Newhouse, 2024),

$$T^{\mathrm{G}} \in \arg\max_{\|T\|_{\mathrm{op}}\leq 1} \langle \Psi, T\rangle_F. \tag{4}$$

A solution is the row-orthogonalized update obtained by performing transformation in Equation 2, $T^{\mathrm{G}} = P^{\mathrm{G}}\Psi = \mathrm{msign}(\Psi)$, with the global preconditioner $P^{\mathrm{G}} = \Gamma(X)^{-1/2} = \widetilde{U}\Lambda^{-1/2}\widetilde{U}^\top$ when $\Gamma(X) \succ 0$. Along the i-th spectral interaction direction, P^{G} applies the scaling factor $\lambda_i^{-1/2} = \widetilde{\sigma}_i^{-1}$, mapping the corresponding singular value $\widetilde{\sigma}_i$ to one.

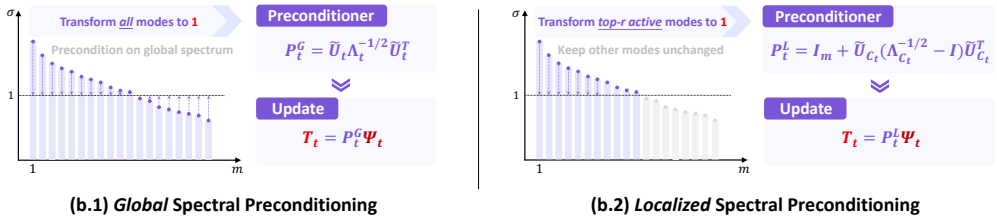


Figure 3: Spectral Interaction Preconditioning.

Localized Spectral Preconditioning. Another possible method starts from the observation that Ψ already serves as a marginally normalized dense update. This motivates retaining Ψ as the reference and considering the regularized spectral steepest descent problem

$$T_+ := \arg\min_{\|T\|_{\mathrm{op}} \le 1} \frac{1}{2}\|T - \Psi\|_F^2 = \arg\max_{\|T\|_{\mathrm{op}} \le 1} \left\{ \langle \Psi, T \rangle_F - \frac{1}{2}\|T\|_F^2 \right\}. \tag{5}$$

Its solution is $T_+ = P_+ \Psi$. Defining the active set $\mathcal{A} := \{i : \lambda_i > 1\}$, the corresponding preconditioner is $P_+ = I_m + \widetilde{U}_{\mathcal{A}} (\Lambda_{\mathcal{A}}^{-1/2} - I) \widetilde{U}_{\mathcal{A}}^\top$ (see Proposition 1 in Appendix D.1). Thus, the solution modifies only the active spectral modes (clipping to one) while leaving the remaining modes unchanged. To further localize the transformation, we impose a spectral budget r and let $\mathcal{C} \subseteq \mathcal{A}$ index the $\min\{r, |\mathcal{A}|\}$ largest active eigenvalues (see Proposition 2 in Appendix D.1). The resulting update is $T^{\mathrm{L}} = P^{\mathrm{L}} \Psi$, with the localized preconditioner $P^{\mathrm{L}} = I_m + \widetilde{U}_{\mathcal{C}} (\Lambda_{\mathcal{C}}^{-1/2} - I) \widetilde{U}_{\mathcal{C}}^\top$.

3 NormPre

This section instantiates the Normalize-Then-Precondition framework as the NormPre optimizer, with variants NormPre-G and NormPre-L using global and localized spectral preconditioning, respectively.

3.1 The NormPre Optimizer

Let $W_t \in \mathbb{R}^{m \times n}$ denote a matrix-valued parameter with stochastic gradient $G_t = \nabla_{W_t} \mathcal{L}_t$, and let $M_t = \mu M_{t-1} + G_t$ denote its first-order momentum. Following Sections 2.2 and 2.3, Algorithm 1 summarizes the NormPre optimization procedure.

Algorithm 1 The NormPre Optimizer

Require: Layer weight $W_t \in \mathbb{R}^{m \times n}$, learning rate η_t, momentum coefficient μ, weight decay coefficient λ_{wd}, optimizer variant (NormPre-G or NormPre-L).

Initialize $M_0 \leftarrow \mathbf{0}$
for each step **do**
  $G_t \leftarrow \nabla_{W_t} \mathcal{L}_t$
  $M_t \leftarrow \mu M_{t-1} + G_t$
  $k_t \leftarrow t \bmod 2$ ▷ Alternating Row/Column Orientation
  $X_t \leftarrow \mathcal{T}_{W_t, k_t}(M_t)$ ▷ Relaxed Tangent Momentum
  $\Psi_t \leftarrow D(X_t)^{-1} X_t$ ▷ Diagonal-Gram Normalization
  if NormPre-G **then**
    $T_t \leftarrow P_t^{\mathrm{G}} \Psi_t = \mathrm{msign}(\Psi_t)$
  else if NormPre-L **then**
    $\left(\widetilde{U}_{\mathcal{C}_t}, \Lambda_{\mathcal{C}_t}\right) \leftarrow \mathrm{Eig}_{\mathcal{C}_t}(\Psi_t \Psi_t^\top)$ ▷ Eigenspace Extraction
    $T_t \leftarrow P_t^{\mathrm{L}} \Psi_t = (I + \widetilde{U}_{\mathcal{C}_t} (\Lambda_{\mathcal{C}_t}^{-1/2} - I) \widetilde{U}_{\mathcal{C}_t}^\top) \Psi_t$
  end if ▷ Spectral Interaction Preconditioning
  $W_{t+1} \leftarrow W_t - \eta_t \left(\mathcal{R}_{k_t}(T_t) + \lambda_{\mathrm{wd}} W_t\right)$ ▷ Consistent Update RMS
end for

The remaining optimizer-level components in Algorithm 1 specify how the update signal is constructed across matrix orientations and how the final matrix update is rescaled to a consistent RMS.

Alternating Row/Column Orientation. We alternate the active orientation between rows and columns across optimization steps, allowing the row-side construction in Section 2 to apply symmetrically to both matrix axes. Specifically, let $k_t = t \bmod 2$ with $\mathcal{O}_0(A) = A$ and $\mathcal{O}_1(A) = A^\top$, and define $\overline{W}_t := \mathcal{O}_{k_t}(W_t)$ and $\overline{M}_t := \mathcal{O}_{k_t}(M_t)$.

Relaxed Tangent Momentum. We distinguish the row-wise momentum components orthogonal and parallel to the current parameter direction (*i.e.*, tangent and radial components). These components primarily govern changes in parameter direction and scale, respectively. Let $\bar{w}_{t,i}$ and $\bar{m}_{t,i}$ denote the i-th rows of $\overline{W}_t$ and $\overline{M}_t$, respectively. For each row, we define $x_{t,i} := \bar{m}_{t,i} - \langle \bar{m}_{t,i}, \bar{w}_{t,i} \rangle \bar{w}_{t,i}$. This formulation recovers strict tangent projection when $\|\bar{w}_{t,i}\|_2 = 1$. For general weights, this relaxed form preserves the tangent component while retaining a norm-dependent radial contribution. Collecting these rows yields the input signal $X_t := \mathcal{T}_{W_t,k_t}(M_t)$ for our framework.

Consistent Update RMS. Following the Muon convention (Liu et al., 2025), we set the target RMS of matrix updates to 0.2, facilitating shared hyperparameters and direct comparison with AdamW and Muon. For update $T \in \mathbb{R}^{m\times n}$ with $\mathrm{RMS}(T) := \|T\|_F/\sqrt{mn}$, we define $\mathcal{R}_k(T) := 0.2 \cdot \mathcal{O}_k^{-1}(T)/\,\mathrm{RMS}(T)$. Here, $\mathcal{O}_k^{-1}$ maps the update from the active orientation back to the original parameter layout and the rescaling ensures an update RMS of 0.2 across matrix parameters.

3.2 Implementation of Spectral Preconditioning

We now detail the spectral preconditioning step in Algorithm 1, implemented via Newton-Schulz iteration for NormPre-G and eigenspace extraction for NormPre-L.

NormPre-G: Newton-Schulz Implementation. For the global realization, we approximate $T_t = \mathrm{msign}(\Psi_t)$ using the Newton-Schulz iteration adopted by Muon (Jordan et al., 2024; Liu et al., 2025). The iteration acts on the normalized update Ψ_t and maps its singular values toward one, realizing full-spectrum preconditioning. We use five Newton-Schulz iterations throughout our experiments.

NormPre-L: Exact and Sketch Implementations. For the localized realization, the preconditioner depends on a selected eigenspace of the interaction matrix $\Gamma_t = \Psi_t \Psi_t^\top$.

- *Exact Eigenspace Extraction.* The exact implementation computes the full eigendecomposition of Γ_t to identify the up to r largest active eigenvalues ($\lambda_{t,i} > 1$). The selected eigenpairs $(\widetilde{U}_{\mathcal{C}_t}, \Lambda_{\mathcal{C}_t})$ are then used in the localized preconditioner.
- *Sketch-Based Eigenspace Extraction.* For scalable training, we approximate the leading interaction eigenspace with a randomized sketch (Halko et al., 2011). The sketch builds a low-dimensional subspace from Ψ_t and applies Rayleigh–Ritz extraction to estimate the leading eigenpairs. Modes with $\widehat{\lambda}_{t,i} > 1$ are retained within the rank budget r and used in the localized preconditioner. The complete procedure is given in Algorithm 2, with details in Appendix B.

Remark 1 (Computational Complexity)**.** Muon and NormPre-G have the same computational complexity $\mathcal{O}(mn + qmns)$, where $s = \min\{m, n\}$. Both use the same q-step Newton-Schulz transformation ($\mathcal{O}(qmns)$), while the additional diagonal-Gram normalization in NormPre-G contributes only $\mathcal{O}(mn)$. For NormPre-L, the Exact implementation requires $\mathcal{O}(m^2n+m^3)$ computation to form and decompose the full interaction matrix. The Sketch-based implementation reduces this cost to $\mathcal{O}\big((p+1)mn\ell + (m+n)\ell^2 + \ell^3\big)$ by restricting eigenspace extraction to a subspace of dimension $\ell = \min\{m, r+o\}$, ensuring scalability for $\ell \ll m$ (see Appendix C for detailed analysis).

In the following, we provide the convergence guarantee for NormPre.

Theorem 1 (Convergence of NormPre without Momentum)**.** *Suppose $\mathcal{L}$ is L-smooth and lower bounded by $\mathcal{L}_{\inf}$ and let $\Delta_0 := \mathcal{L}(W_0) - \mathcal{L}_{\inf}$. Under a fixed orientation, let $\phi_{t,i}$ denote the angle between gradient $\bar{g}_{t,i}$ and weight $\bar{w}_{t,i}$, and $\phi'_{t,j,i}$ the angle between normalized row $\psi_{t,j}$ and weight $\bar{w}_{t,i}$. Suppose $\sin\phi_{t,i} \geq \gamma > 0$ and $(\sum_j \cos^2\phi'_{t,j,i})^{1/2} \leq \gamma'$ for all t, i. Define the maximum radial ratio $\nu_t := \max_i \|\bar{g}_{t,i} - x_{t,i}\|_2/\|x_{t,i}\|_2$ and the spectral factor $\chi_t := \widetilde{\sigma}_{t,1}/\widetilde{\sigma}_{t,m}$ for* NormPre-G *(assuming $\Gamma_t \succ 0$ for all t), or $\chi_t := \widetilde{\sigma}_{t,1}$ for* NormPre-L *(Exact). Assume there exists $\epsilon > 0$ such that $\max_{0\leq t\leq T} \nu_t\gamma'\chi_t \leq 1-\epsilon$. For any constant $C > 0$, choosing $\eta = C/\sqrt{T+1}$ yields*

$$\min_{0\leq t\leq T} \|\nabla\mathcal{L}(W_t)\|_F \leq \frac{1}{\sqrt{T+1}} \left(\frac{\Delta_0 \max\{m,n\}^{1/2}}{\epsilon\gamma C} + \frac{LC\max\{m,n\}^{3/2}}{2\epsilon\gamma} \right).$$

Remark 2. Theorem 1 analyzes deterministic NormPre-G and NormPre-L without momentum, RMS scaling, or weight decay. The condition $\max_{0\le t\le T} \nu_t\gamma'\chi_t \le 1-\epsilon$ is sufficient but not necessary for convergence. By bounding the product of the radial ratio (ν_t), geometric non-orthogonality (γ'), and spectral factor (χ_t), this condition controls the worst-case reduction in the first-order gradient signal arising from the radial interaction in the refined update. For any constant $C > 0$, choosing $\eta = C/\sqrt{T+1}$ yields an $\mathcal{O}(T^{-1/2})$ convergence rate. This bound is minimized at $\eta^\star = \sqrt{\frac{2\Delta_0}{L\max\{m,n\}(T+1)}}$, yielding $\min_{0\le t\le T}\|\nabla\mathcal{L}(W_t)\|_F \le \frac{\max\{m,n\}\sqrt{2L\Delta_0}}{\epsilon\gamma\sqrt{T+1}}$. We defer the proof of Theorem 1 and the Sketch analysis (Corollary 1) to Appendices D.2.2 and D.2.3, respectively.

4 Experiments

This section evaluates NormPre across diverse pretraining settings (Section 4.1), analyzes its efficiency and spectral dynamics (Section 4.2) and characterizes the performance-efficiency trade-off between NormPre-G and NormPre-L (Section 4.3).

4.1 Scaling Experiments

Experimental Setup. We evaluate NormPre across model scales, architectures and datasets. Our experiments cover GPT-2 Small on OpenWebText (Radford et al., 2019; Gokaslan & Cohen, 2019), LLaMA-{130M, 350M, 1.3B} on C4 (Touvron et al., 2023; Raffel et al., 2020), and Qwen3-{0.6B, 1.7B} on Pile (Yang et al., 2025; Gao et al., 2020). We compare NormPre-G and NormPre-L with AdamW (Loshchilov & Hutter, 2019), Muon (Jordan et al., 2024) and MANO (Gu & Xie, 2026). NormPre-G uses five Newton-Schulz iterations and NormPre-L uses the Sketch implementation with rank $r = 32$. All models are trained for $10{,}000$ steps with a sequence length of 1024 and an effective batch size of 512. Within each setting, all optimizers use matched model initialization, data order, training budget and evaluation protocol. Detailed configurations are provided in Appendix A.1.

GPT-2 Small Pretraining on OpenWebText. Figure 1a shows training and validation loss trajectories averaged over three matched seeds. Both variants achieve lower training and validation loss than the baselines during later training, demonstrating the effectiveness of NormPre. Table 1 confirms that both variants outperform all three baselines in mean best validation loss, with NormPre-G achieving the lowest value. Compared with the strongest baseline Muon, NormPre-G and NormPre-L reduce validation loss by 0.0397 and 0.0238, respectively.

LLaMA Pretraining on C4 and Qwen3 Pretraining on Pile. We extend the comparison to LLaMA-{130M, 350M, 1.3B} on C4 and Qwen3-{0.6B, 1.7B} on Pile. Table 1 and Figures 1b–1d show that both variants consistently outperform all three baselines across these diverse architectures, with NormPre-G achieving the lowest loss. For LLaMA, the advantages over the strongest baseline are most pronounced at the largest evaluated scale (1.3B), where NormPre-G and NormPre-L reduce validation loss by 0.0392 and 0.0301, respectively. These optimization gains extend similarly to Qwen3. Notably, NormPre-L performs closely to NormPre-G at the 1.7B scale with a minimal loss difference of 0.0029 (Qwen3 loss curves are deferred to Appendix A.2).

4.2 Training Efficiency and Spectral Dynamics

Training Efficiency. Table 2 reports training efficiency on GPT-2 Small and LLaMA-1.3B under the same configurations as the main experiments. As summarized in Table 9 and Remark 1 (Appendix C), NormPre-G shares Muon's asymptotic complexity, but its additional normalization increases optimizer latency by 5.76% and 4.37% relative to Muon on the two models, respectively. NormPre-L approximates the interaction subspace via sketching, reducing optimizer latency by 38.12% and 67.00%. These optimizer-level differences have a smaller effect on end-to-end training efficiency because optimizer updates account for only a small fraction of step time, which is largely spent on forward and backward computation. Relative to Muon, NormPre-G increases step time by about 0.44% and 0.08% on both models, while NormPre-L reduces it by 1.16% and 3.13%, with corresponding improvements in throughput. Both variants remain slower per step than AdamW and MANO due to the additional spectral processing, but match the whole-training peak memory of Muon and MANO. Further discussions and efficiency results on Qwen3 are provided in Appendix A.2.

Table 1: **Validation loss across model scales, architectures and pretraining datasets.** GPT-2 Small results show mean ± standard deviation over three matched seeds, and larger models report single-run results under matched training budgets. Shaded columns indicate our variants. The best baseline and our outperforming variants are shown in underline and **bold**, respectively.

Setting		Validation Loss ↓				
Model	Dataset	AdamW	Muon	MANO	NormPre-G	NormPre-L
GPT-2 Small	OpenWebText	3.1444 ±0.0054	3.1064 ±0.0054	3.1156 ±0.0057	**3.0667 ±0.0049**	**3.0826 ±0.0034**
LLaMA-130M	C4	3.1363	3.1019	3.1102	**3.0737**	**3.0930**
LLaMA-350M	C4	3.0378	2.9999	2.9946	**2.9678**	**2.9758**
LLaMA-1.3B	C4	2.9385	2.9037	2.8963	**2.8571**	**2.8662**
Qwen3-0.6B	Pile	2.8956	2.8335	2.8382	**2.7967**	**2.8066**
Qwen3-1.7B	Pile	2.6758	2.6408	2.6205	**2.5868**	**2.5897**

Table 2: **Training Efficiency Comparisons.** Measurements use the same training configurations as the main experiments and are averaged over 100 optimizer steps after 20 warmup steps.

Model	Dataset	Optimizer	Optimizer Latency ↓ (ms/step)	E2E Step Time ↓ (ms/step)	Throughput ↑ (k tokens/s)	Peak Memory ↓ (GiB)
GPT-2 Small	OpenWebText	AdamW	2.7	3364.8	155.82	14.93
		Muon	112.8	3479.8	150.67	14.61
		MANO	8.8	3365.6	155.78	14.61
		NormPre-G	119.3 ↑5.76%	3495.0 ↑0.44%	150.01 ↓0.44%	14.62
		NormPre-L	69.8 ↓38.12%	3439.4 ↓1.16%	152.44 ↑1.17%	14.61
LLaMA-1.3B	C4	AdamW	25.1	22121.0	23.70	42.34
		Muon	1010.1	23124.2	22.67	37.84
		MANO	87.1	22210.3	23.61	37.84
		NormPre-G	1054.2 ↑4.37%	23141.8 ↑0.08%	22.66 ↓0.04%	37.83
		NormPre-L	333.3 ↓67.00%	22399.7 ↓3.13%	23.41 ↑3.26%	37.84

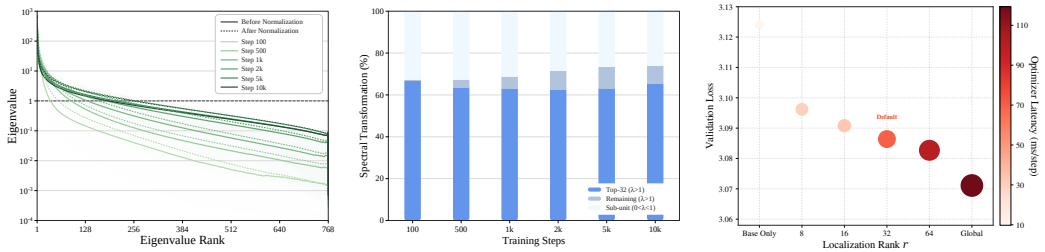
Spectral Dynamics. Figures 4a–4b validate our hierarchical Normalize-Then-Precondition design, revealing two key patterns. *(a) Normalization yields an anisotropic base update.* Equalizing row norms removes scale weighting in the Gram matrix, giving smaller-norm rows greater relative influence. This redistributes the spectral mass and might raise eigenvalues across multiple ranks (dotted curves over solid curves). Crucially, this anisotropy persists throughout training, necessitating further refinement. Table 7 confirms that constructing the preconditioner from Ψ rather than X yields lower losses. *(b) Localized preconditioning captures the dominant transformation energy.* Building on the normalized base, global preconditioning contracts modes above one and amplifies positive modes below one. Decomposing its spectral transformation energy (squared singular-value change $\sum_{\lambda_i>0}(\sqrt{\lambda_i}-1)^2$) reveals that the top-32 modes ($\lambda_i > 1$) consistently account for approximately 62%–67% across checkpoints. By concentrating strictly on these modes, localized preconditioning covers a substantial portion of the global transformation to yield the optimization gains observed in Figure 1. Further details and training dynamics are provided in Appendix A.3.

4.3 Comparisons of NormPre-G and NormPre-L

Drawing on the preceding findings, we discuss the complementary strengths of two variants.

(a) NormPre-G yields superior performance via full-spectrum transformations. By operating on the entire eigenspectrum, NormPre-G consistently achieves the lowest validation loss across all evaluations. This indicates that transforming the long-tail modes alongside the dominant spectral energy provides crucial optimization benefits that push the absolute performance limit.

(b) NormPre-L offers a performance-efficiency trade-off via local-spectrum transformations. The normalized base update motivates a regularized steepest-descent problem limiting deviations from this reference, giving NormPre-L a principled formulation (Propositions 1 and 2). Although empirical results favor the broader transformation of NormPre-G in validation loss, Figure 4c illustrates a clear trade-off on GPT-2 Small that larger r lowers validation loss at higher optimizer cost. We use $r = 32$ as the default spectral budget to balance performance and efficiency. At this budget, NormPre-L yields clear optimization gains with lower optimizer latency and end-to-end step time than NormPre-G (Table 2), serving as a practical and scalable alternative.



(a) (b) (c)

Figure 4: (a) Eigenspectra before and after marginal normalization; (b) Spectral transformation percentages under global and localized preconditioning; (c) Performance-efficiency trade-off.

5 Related Work

LLM Optimizers. Recent optimizers span coordinate-wise methods such as AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019), Lion (Chen et al., 2023), Sophia (Liu et al., 2024), and Cautious Optimizers (Liang et al., 2026), and block-wise approaches like Adam-mini (Zhang et al., 2025) and Blockwise LR (Wang et al., 2025). Alongside these, matrix-level methods like K-FAC (Martens & Grosse, 2015), Shampoo (Gupta et al., 2018; Shi et al., 2023), and SOAP (Vyas et al., 2025) construct preconditioners from curvature or accumulated statistics. We focus on matrix optimizers operating directly on update geometry via orthogonalization or row/column normalization.

Orthogonalization and Normalization. Matrix structure is increasingly exploited via orthogonalization or lightweight row/column normalization. On the orthogonalization front, Muon applies a matrix-sign transformation to momentum via Newton-Schulz iterations (Jordan et al., 2024; Liu et al., 2025), while broader norm-constrained frameworks connect these updates to steepest descent and spectral spheres (Bernstein & Newhouse, 2024; Pethick et al., 2025b; Xie et al., 2026). Alternatively, row/column normalization offers a lightweight way. For example, MANO combines momentum projection with alternating row/column normalization (Gu & Xie, 2026), while RMNP uses only row normalization (Deng et al., 2026). MOGA derives such normalized updates from mean-normalized operator norms under the steepest-descent perspective (Xu et al., 2026).

Closely Relevant Work. Several methods retain orthogonalization as a core operation and incorporate normalization or adaptive scaling. One line modifies the post-orthogonalization update: Muon+ applies row/column normalization to the orthogonalized output (Zhang et al., 2026), while NorMuon and AdaMuon introduce neuron-wise and element-wise adaptive scaling (Li et al., 2025; Si et al., 2025). Another line applies pre-orthogonalization normalization for numerical stability. For instance, SWAN normalizes instantaneous gradients before whitening (Ma et al., 2024) and MuonEq (Chang et al., 2026) equilibrates momentum to improve Newton-Schulz conditioning. While these approaches discuss the utility of incorporating normalization, it remains underexplored to formulate a systematic hierarchical framework and investigate if targeted spectral schemes can be effectively developed. We establish the **Normalize-Then-Precondition** framework, introducing two realizations of spectral preconditioning from spectral steepest descent and its regularized form with leading mode selection.

6 Conclusion

This paper formalizes the Normalize-Then-Precondition framework to hierarchically organize marginal scales and interactions in matrix optimizers. Building on this, we propose NormPre, which achieves superior LLM pretraining performance over established baselines with a clear performance-efficiency trade-off. We highlight three promising directions for future research addressing current limitations: (a) Developing hardware-aware implementations to improve the efficiency of spectral transformations; (b) Scaling our empirical validation beyond 1.7B parameters; (c) Exploring adaptive spectral schemes to narrow the performance gap between localized and global preconditioning while preserving efficiency. Overall, our work provides a systematic framework for integrating normalization and spectral preconditioning, inspiring more effective and efficient LLM training.

AI USE STATEMENT

We used OpenAI Codex and Anthropic Claude Code to assist with core code implementation, data processing, experimental evaluation, translation, and the verification and refinement of mathematical proofs. Additionally, we used these tools for literature search, manuscript drafting and revision. The research questions and ideas, conceptual framework, methodology, proof strategies and initial proof drafts, and experimental design were developed by the authors without AI assistance. We have reviewed all AI-assisted work, including manuscript text, source code, data processing, and training and evaluation pipelines. We retain full responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies effective and efficient optimization methods for large language model training. It does not involve human-subject experiments, collection of personal information or deployment in real-world decision-making systems. Our experiments use established language-modeling datasets, including OpenWebText, C4 and Pile, and we do not introduce or release new training data. We follow standard research practices for experimental evaluation, reporting and reproducibility.

REPRODUCIBILITY STATEMENT

Section 3 provides a detailed description of the proposed optimizer NormPre, including two variants NormPre-G and NormPre-L. Section 4 presents the experimental setup, datasets, evaluation protocol and main results, with additional configurations, results, dynamics analyses and ablations provided in Appendix A. Implementation details for spectral preconditioning and computational complexity analyses are given in Appendices B and C. The assumptions, derivations and complete proofs supporting our theoretical results are provided in Appendix D. An anonymous repository containing the source code and materials for reproducing the experiments is linked in the abstract.

REFERENCES

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

Da Chang, Qiankun Shi, Lvgang Zhang, Yu Li, Ruijie Zhang, Yao Lu, Yongxiang Liu, and Ganzhao Yuan. Muoneq: Balancing before orthogonalization with lightweight equilibration. *arXiv preprint arXiv:2603.28254*, 2026.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. *Advances in neural information processing systems*, 36:49205–49233, 2023.

Shenyang Deng, Zhuoli Ouyang, Tianyu Pang, Zihang Liu, Ruochen Jin, Shuhua Yu, and Yaoqing Yang. Rmnp: Row-momentum normalized preconditioning for scalable matrix-based optimization. *arXiv preprint arXiv:2603.20527*, 2026.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Athanasios Glentis, Jiaxiang Li, Andi Han, and Mingyi Hong. A minimalist optimizer design for llm pretraining. In *ES-FoMo III: 3rd Workshop on Efficient Systems for Foundation Models*, 2025.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Yufei Gu and Zeke Xie. Mano: Restriking manifold optimization for llm training. *arXiv preprint arXiv:2601.23000*, 2026.

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

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2): 217–288, 2011.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan. github. io/posts/muon*, 6(3):4, 2024.

Andrej Karpathy. nanogpt. https://github.com/karpathy/nanoGPT, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. Normuon: Making muon more efficient and scalable. *arXiv preprint arXiv:2510.05491*, 2025.

Kaizhao Liang, Lizhang Chen, Bo Liu, and Qiang Liu. Cautious optimizers: Improving training with one line of code. In *International Conference on Learning Representations*, volume 2026, pp. 106538–106563, 2026.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *International conference on learning representations*, volume 2024, pp. 1621–1650, 2024.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

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

Chao Ma, Wenbo Gong, Meyer Scetbon, and Edward Meeds. Swan: Sgd with normalization and whitening enables stateless llm training. *arXiv preprint arXiv:2412.13148*, 2024.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015.

Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1):50–59, 1960.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos. *arXiv preprint arXiv:2502.07529*, 2025a.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training Deep Learning Models with Norm-Constrained LMOs. In *ICML*, 2025b.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Hao-Jun Michael Shi, Tsung-Hsien Lee, Shintaro Iwasaki, Jose Gallego-Posada, Zhijing Li, Kaushik Rangadurai, Dheevatsa Mudigere, and Michael Rabbat. A distributed data-parallel pytorch implementation of the distributed shampoo optimizer for training neural networks at-scale. *arXiv preprint arXiv:2309.06497*, 2023.

Chongjie Si, Debing Zhang, and Wei Shen. Adamuon: Adaptive muon optimizer. *arXiv preprint arXiv:2507.11005*, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam for language modeling. In *International Conference on Learning Representations*, volume 2025, pp. 93423–93444, 2025.

Jinbo Wang, Mingze Wang, Zhanpeng Zhou, Junchi Yan, Lei Wu, et al. The sharpness disparity principle in transformers for accelerating language model pre-training. *arXiv preprint arXiv:2502.19002*, 2025.

Tian Xie, Haoming Luo, Haoyu Tang, Hu Yiwen, Jason Klein Liu, Qingnan Ren, Yang Wang, Xin Zhao, Rui Yan, Bing Su, Chong Luo, and Baining Guo. Controlled LLM training on spectral sphere. In *Forty-third International Conference on Machine Learning*, 2026.

Ruihan Xu, Jiajin Li, and Yiping Lu. On the width scaling of neural optimizers under matrix operator norms i: Row/column normalization and hyperparameter transfer. *arXiv preprint arXiv:2603.09952*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Ruijie Zhang, Yequan Zhao, Ziyue Liu, Zhengyang Wang, and Zheng Zhang. Muon+: Towards better muon via one additional normalization step. *arXiv e-prints*, pp. arXiv–2602, 2026.

Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik Durk Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. In *International Conference on Learning Representations*, volume 2025, pp. 28033–28063, 2025.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection. *International Conference on Machine Learning*, 2024.

Appendix

A Experimental Details . . . 14
A.1 Experimental Configurations . . . 14
A.2 Additional Experimental Results . . . 15
A.3 Additional Details on Spectral and Training Dynamics . . . 16
A.4 Additional Ablations . . . 18
B Detailed Implementations on Spectral Preconditioning . . . 20
C Detailed Computational Complexity Analysis . . . 21
C.1 Shared Components in NormPre-G and NormPre-L . . . 21
C.2 NormPre-G and Muon . . . 21
C.3 NormPre-L . . . 22
D Proofs . . . 23
D.1 Optimization Formulation of Localized Spectral Refinement . . . 23
D.2 Convergence Analysis of NormPre . . . 25
D.2.1 Lemmas . . . 25
D.2.2 Convergence of NormPre-G and NormPre-L (Exact) . . . 29
D.2.3 Convergence of NormPre-L (Sketch) . . . 30

A Experimental Details

A.1 Experimental Configurations

Models and Data. Our experiments use GPT-2 Small on OpenWebText (Radford et al., 2019; Gokaslan & Cohen, 2019), LLaMA-{130M, 350M, 1.3B} on C4 (Touvron et al., 2023; Raffel et al., 2020) and Qwen3-{0.6B, 1.7B} on Pile (Yang et al., 2025; Gao et al., 2020). GPT-2 Small is reproduced using the nanoGPT codebase (Karpathy, 2022). We follow the general pretraining setup in Zhao et al. (2024); Raffel et al. (2020). Table 3 summarizes the model and training configurations.

OpenWebText is processed with the GPT-2 BPE tokenizer, and the preprocessed training and validation data are loaded into CPU memory before training. For C4, we tokenize the English corpus with the T5-base tokenizer. For Pile, we recover documents from the Pythia/Pile token stream and re-tokenize them with the Qwen3 tokenizer. For C4 and Pile, we construct fixed preprocessed datasets of $5{,}120{,}000$ training sequences and $2{,}048$ validation sequences, each with length 1024. Documents are processed independently by truncating long documents and padding shorter ones without cross-document packing. The resulting data order is fixed and shared across all optimizer runs. With sequence length 1024, effective batch size 512 and $10{,}000$ optimization steps, each training trajectory processes approximately 5.24B nominal token positions.

Table 3: Model and Training Configurations.

Model	Model Size	Dataset	Layers	Hidden	FFN	Heads	Seq. Len.	Eff. Batch	Base Peak LR	Warmup	Steps
GPT-2 Small	124M	OpenWebText	12	768	3,072	12	1,024	512	6×10^{-4}	2,000	10,000
LLaMA-130M	130M	C4	12	768	2,048	12	1,024	512	6×10^{-4}	1,000	10,000
LLaMA-350M	350M	C4	24	1,024	2,736	16	1,024	512	3×10^{-4}	1,000	10,000
LLaMA-1.3B	1.3B	C4	24	2,048	5,461	32	1,024	512	3×10^{-4}	1,000	10,000
Qwen3-0.6B	0.6B	Pile	28	1,024	3,072	16	1,024	512	3×10^{-4}	1,000	10,000
Qwen3-1.7B	1.7B	Pile	28	2,048	6,144	16	1,024	512	3×10^{-4}	1,000	10,000

Optimizers. We compare NormPre-G and NormPre-L with AdamW (Loshchilov & Hutter, 2019), Muon (Jordan et al., 2024) and MANO (Gu & Xie, 2026). Table 4 summarizes the default optimizer configurations. We use $(\beta_1, \beta_2) = (0.9, 0.95)$ for AdamW and momentum coefficient $\mu = 0.95$ for Muon, MANO and both NormPre variants. Muon and NormPre-G use five Newton-Schulz iterations. NormPre-L uses the Sketch-based implementation with interaction rank $r = 32$, sketch oversampling $o = 8$ and one power iteration $p = 1$. Both NormPre variants match the post-preconditioning update RMS to a target value of 0.2 before the parameter update. For GPT-2 Small, Muon follows the Keller–Jordan convention (Jordan et al., 2024) for hidden matrix parameters, using an effective peak matrix learning rate of 0.02 together with width-dependent update scaling. For LLaMA and Qwen3, Muon follows the scalable RMS-matching convention (Liu et al., 2025) and uses the base learning rate of the corresponding model setting. Muon, MANO and both NormPre variants are applied to hidden two-dimensional attention and MLP weight matrices, with token embeddings, output heads, normalization parameters, biases and other one-dimensional parameters optimized by AdamW.

Table 4: Default Optimizer Configurations. A dash indicates that the corresponding hyperparameter is not applicable.

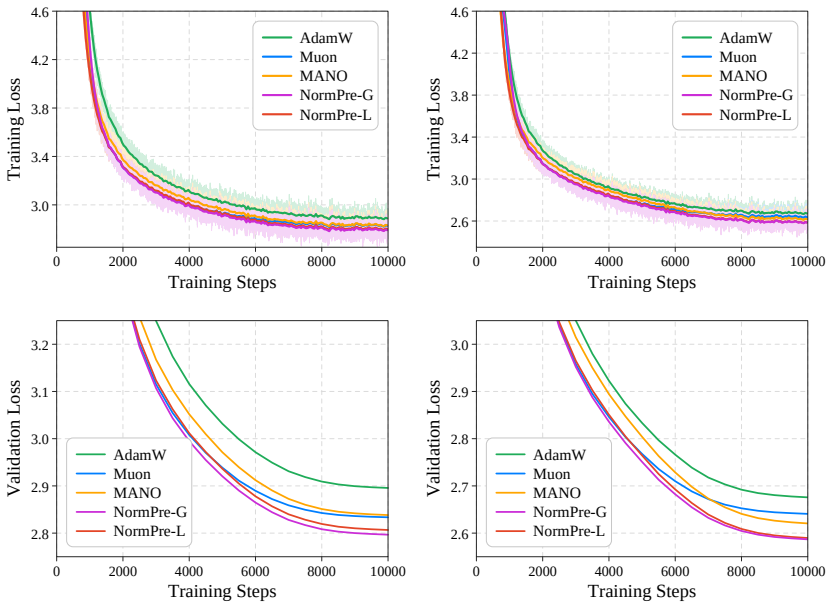
Hyperparameter	AdamW	Muon	MANO	NormPre-G	NormPre-L
β_1	0.9	–	–	–	–
β_2	0.95	–	–	–	–
Momentum μ	–	0.95	0.95	0.95	0.95
Newton-Schulz steps	–	5	–	5	–
Weight decay	0.1	0.1	0.1	0.1	0.1
Interaction rank r	–	–	–	–	32
Sketch oversampling o	–	–	–	–	8
Power iterations p	–	–	–	–	1

Training. All models are trained for $10{,}000$ optimization steps with sequence length 1024 and effective batch size 512. We use linear warmup followed by cosine decay to 10% of the peak learning rate, with the model-specific learning rates and warmup steps given in Table 3. Weight decay is

0.1 and gradients are clipped at 1.0. All models are evaluated every 500 optimization steps. GPT-2 Small uses 200 randomly sampled validation mini-batches per evaluation, while LLaMA and Qwen3 evaluate on the complete fixed validation split of 2,048 sequences. Within each setting, all compared optimizers use matched model initialization, data order, training budget, base learning-rate schedule and validation protocol. All experiments use the same PyTorch training codebase and NVIDIA A100 80GB GPUs. GPT-2 Small uses one GPU per training trajectory and LLaMA and Qwen3 use four-rank data parallelism.

A.2 Additional Experimental Results

Training and Validation Loss for Qwen3 Models. We provide the training and validation loss curves for Qwen3-0.6B and Qwen3-1.7B on Pile in Figure 5.



(a) Qwen3-0.6B on Pile (b) Qwen3-1.7B on Pile

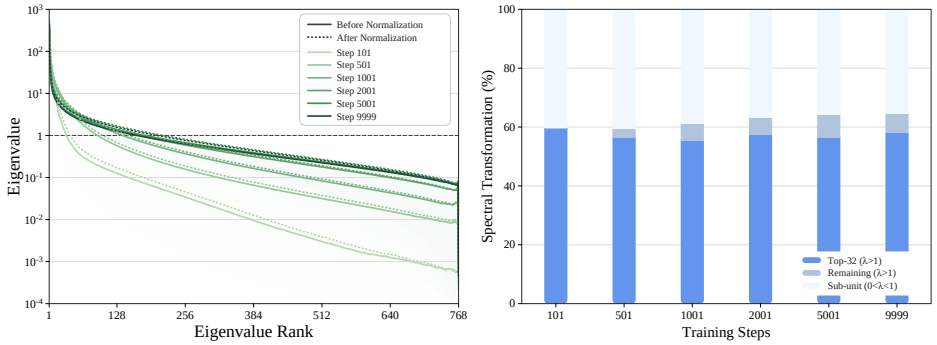
Figure 5: **Qwen3 Models on Pile.** Training and validation loss for Qwen3-{0.6B, 1.7B} with AdamW, Muon, MANO, NormPre-G and NormPre-L. For training loss, dark curves denote 50-step moving averages and light curves the raw trajectories.

Training Efficiency Configurations. Efficiency measurements mirror the main pretraining configurations (including model architecture, sequence length of 1024, effective batch size of 512, and gradient clipping) with the following specific profiling protocols:

- *Precision & Data Loading:* GPT-2 Small and LLaMA use FP32 parameters with BF16 autocast, while Qwen3 uses pure BF16. OpenWebText is fully materialized in host memory, whereas C4 and Pile are read dynamically from persistent NumPy memmaps (transferring only the active batch to the GPU to avoid full-dataset materialization).
- *Measurement Window:* We profile 100 optimizer steps following 20 warmup steps. To isolate core algorithmic overhead, we disable optimizer diagnostics (e.g., descent alignment), validation, checkpointing, W&B logging, and `torch.compile`.
- *Metric Definitions:* Optimizer latency isolates `optimizer.step()` using strict CUDA synchronization immediately before and after the call. E2E step time encompasses the complete training loop (data transfer, forward/backward passes, gradient accumulation, clipping, and optimizer update). Throughput is computed as global tokens per E2E step time, and peak memory tracks the maximum CUDA allocated memory during a measured step.

Table 5: **Training Efficiency on Qwen3-1.7B.** Measurements use the same training configuration as the main Qwen3-1.7B experiment and are averaged over 100 optimizer steps after 20 warmup steps. We report global throughput across 4 GPUs and per-GPU peak allocated memory.

Model	Dataset	Optimizer	Optimizer Latency ↓ (ms/step)	E2E Step Time ↓ (ms/step)	Throughput ↑ (k tokens/s)	Peak Memory ↓ (GiB)
Qwen3-1.7B	Pile	**AdamW**	18.3	12102.7	43.32	53.17
		Muon	940.0	13050.4	40.17	53.17
		MANO	118.7	12217.4	42.91	53.17
		NormPre-G	1003.6 ↑6.77%	13119.7 ↑0.53%	39.96 ↓0.52%	53.17
		NormPre-L	402.7 ↓57.16%	12501.4 ↓4.21%	41.94 ↑4.41%	53.17



(a) (b)

Figure 6: **Spectral Dynamics.** (a) Eigenspectra before and after marginal normalization in the column-active orientation. (b) Spectral transformation percentages across modes under global and localized preconditioning (column-active orientation).

- *TF32 Settings:* Unless otherwise stated, TF32 refers to the CUDA matrix-multiplication TF32 setting. To isolate intrinsic algorithmic scaling from hardware-specific Tensor Core accelerations at small dimensions, the main GPT-2 Small efficiency results disable TF32.

Training Efficiency on Qwen3-1.7B. We further evaluate training efficiency on Qwen3-1.7B with Pile under the main experimental configuration. Similar to the observations in Section 4.2, relative to Muon, NormPre-G increases optimizer latency by 6.77%, with an end-to-end step-time overhead of 0.53%. NormPre-L reduces optimizer latency by 57.16%, yielding a 4.21% reduction in step time and a 4.41% increase in throughput. Both variants remain slower per step than AdamW and MANO. All evaluated optimizers have the same whole-training peak allocated memory of 53.17 GiB per GPU. This shared peak is determined by forward and backward computation, whose memory peak exceeds that of the optimizer step.

A.3 Additional Details on Spectral and Training Dynamics

Spectral Dynamics. We collect spectral diagnostics from a NormPre-L training run on GPT-2 Small with OpenWebText and random seed 1337. We use the Sketch implementation with rank $r = 32$, oversampling $o = 8$, and one power iteration ($p = 1$). Diagnostics cover the 48 attention and MLP weight matrices at checkpoints corresponding to steps 100 (101), 500 (501), 1,000 (1,001), 2,000 (2,001), 5,000 (5,001), and 10,000 (9,999). Figure 4a presents row-active spectra, and Figure 6a provides the column-active counterpart.

(1) Eigenspectra Before and After Normalization. Following Section 3, let X denote the update after relaxed tangent processing (i.e., before normalization) and Ψ denote its marginally normalized counterpart. Both are expressed in the active orientation and measured before spectral preconditioning.

To isolate changes in spectral shape from differences in global update magnitude, we compare

$$\widehat{G}_X = \frac{d}{\mathrm{tr}(XX^\top)} XX^\top, \quad \Gamma = \Psi\Psi^\top,$$

where d is the active-axis dimension. Both matrices have trace d and then the same total spectral mass. The scalar rescaling of $XX^\top$ preserves its eigenvalue ratios and relative row-norm differences, which is used only for analysis in spectral dynamics. We sort the nonzero eigenvalues of $\widehat{G}_X$ and Γ in descending order and average them across matrices at each rank. The shaded regions show the corresponding 25th–75th percentile ranges. Different colors correspond to different training steps. Solid curves show the spectra before marginal normalization ($\widehat{G}_X$), and dotted curves show those after normalization (Γ). The horizontal line at $\lambda = 1$ indicates the target under full-spectrum preconditioning and localized preconditioning.

Effect of Marginal Normalization on the Spectrum. Before normalization, correlations between row directions are weighted by the products of their row norms in the Gram matrix. Normalization gives each row unit norm, removing these weights and increasing the relative influence of smaller-norm rows. To illustrate this effect, consider mutually orthogonal rows with norms $a_1, \ldots, a_d > 0$. The trace-matched Gram matrix is diagonal, with eigenvalues $da_i^2 / \sum_j a_j^2 = a_i^2/\overline{a^2}$, where $\overline{a^2} = \frac{1}{d}\sum_j a_j^2$. After marginal normalization, the rows remain orthogonal and each has unit norm, so all eigenvalues equal one. Thus, the eigenvalue corresponding to a row increases if its squared norm is below the average and decreases if it is above the average. For nonorthogonal rows, eigenvalues do not directly correspond to individual row norms. Removing row-norm weighting still changes the spectral structure formed jointly by the rows and might increase eigenvalues across multiple ranks. This provides a possible explanation for the dotted curves lying above the solid curves over many ranks in Figures 4a and 6a.

(2) Global and Localized Spectral Preconditioning. We use row-active and column-active snapshots from the same NormPre-L trajectory described above. For each snapshot, we compare the spectral transformations induced by global and localized preconditioning on the same normalized interaction spectrum.

Let $\{\lambda_i\}$ denote the nonzero eigenvalues of $\Gamma = \Psi\Psi^\top$. Global preconditioning maps each corresponding singular value $\sqrt{\lambda_i}$ to one. We define the spectral transformation energy of each mode as its squared singular-value change, $e_i = (\sqrt{\lambda_i} - 1)^2$, and partition the total into three components:

$$E_{\mathrm{top}} = \sum_{i \in C} e_i, \quad E_{\mathrm{remain}} = \sum_{\lambda_i > 1, i \notin C} e_i, \quad E_{\mathrm{sub}} = \sum_{0 < \lambda_i < 1} e_i,$$

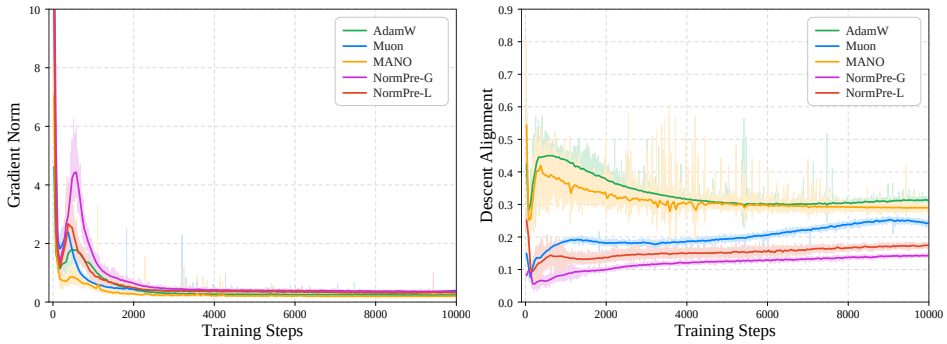
where C contains the indices of the largest at most $r = 32$ eigenvalues exceeding one. Modes with $\lambda_i = 1$ contribute zero. The total spectral transformation energy is $E_{\mathrm{global}} = E_{\mathrm{top}} + E_{\mathrm{remain}} + E_{\mathrm{sub}}$. For each matrix, we express each component as a percentage of E_{global}. The stacked bars show these percentages averaged equally across the 48 matrices at each checkpoint. This decomposition distinguishes the spectral scope of the two variants. Localized preconditioning contracts only the modes in C and leaves the others unchanged. Global preconditioning additionally contracts the remaining modes above one and amplifies the positive modes below one.

Figure 6b presents the column-active results, which follow a similar pattern to the row-active results in Figure 4b. The top-32 modes account for approximately 55%–60% of the total spectral transformation energy across checkpoints.

Training Dynamics. We examine training dynamics on GPT-2 Small with OpenWebText. Gradient norm is measured over all trainable parameters before clipping. Descent alignment is computed over the attention and MLP weight matrices. For each matrix W, let $G_t^{(W)}$ denote its gradient and $\Phi_t^{(W)}$ its optimizer update direction before learning-rate scaling, excluding decoupled weight decay. We report

$$\mathrm{Align}_t = \frac{1}{|\mathcal{M}|} \sum_{W \in \mathcal{M}} \frac{\langle G_t^{(W)}, \Phi_t^{(W)} \rangle_F}{\|G_t^{(W)}\|_F \|\Phi_t^{(W)}\|_F},$$

where $\mathcal{M}$ is the set of evaluated weight matrices. This averages the per-matrix cosine similarities. Each cosine equivalently measures the alignment between the optimizer-induced parameter change and the current negative gradient.



(a) Gradient Norm (b) Descent Alignment

Figure 7: **Training Dynamics.** (a) Pre-clipping global gradient norm over all trainable parameters. (b) Descent alignment between the gradient and optimizer update direction over matrix-optimized parameters. Dark curves show 50-step moving averages and light curves show the corresponding raw trajectories.

Table 6: Ablations on Marginally Normalized Update. Results are reported on GPT-2 Small with OpenWebText using seed 1337 under the same 10k-step training budget.

(a) Momentum Processing

Variant	Validation Loss	
	NormPre-G	NormPre-L
No Tangent	3.0945	3.1139
Strict Tangent	3.1869	3.1196
Relaxed Tangent	3.0711	3.0864

(b) Normalization Orientation

Orientation	Validation Loss	
	NormPre-G	NormPre-L
Row	3.0604	3.0817
Alternating Row/Column	3.0711	3.0864
Row+Column/Step	3.0735	3.0878

Analysis. Both NormPre variants exhibit transient peaks in gradient norm during early training, with a more pronounced peak for NormPre-G. Their descent alignment is also lower than Muon's, indicating greater deviation from the current negative gradient. These distinct early dynamics do not develop into persistent gradient instability. As training progresses, their gradient norms decrease and stabilize, with fewer isolated spikes overall than Muon. Directional differences persist after the gradient norms stabilize: both variants maintain lower alignment than Muon, with NormPre-G remaining the lowest. Thus, NormPre continues to reshape the update direction throughout training. Together with the lower validation losses, these observations demonstrate an effective update geometry: NormPre adjusts update directions while maintaining positive average descent alignment, sustains relatively steady optimization dynamics in later training, and achieves better final solutions.

A.4 Additional Ablations

Ablations on Marginally Normalized Update. Table 6 examines two design choices in constructing the marginally normalized update for both NormPre-G and NormPre-L. In part (a), relaxed tangent projection achieves the lowest validation loss for both variants, outperforming both leaving the momentum unchanged and strict tangent projection. The degradation under strict tangent projection is particularly pronounced for NormPre-G, suggesting that completely removing the radial component can discard useful information before spectral refinement. These results support the relaxed design, which prioritizes direction-changing information while retaining a limited scale-dependent component. From part (b), the choice of normalization orientation has a smaller effect. For NormPre-L, row-only and alternating row/column normalization achieve comparable validation losses, with a slight advantage for the row-only variant in this controlled comparison. For NormPre-G, the row-only variant shows a clearer improvement over alternating normalization, while normalizing both axes at every step performs similarly to the alternating scheme. We adopt alternating normalization as the default to treat the two matrix axes symmetrically and avoid introducing a fixed preferred orientation.

Table 7: Ablations on Interaction Geometry Source. Results on GPT-2 Small with OpenWebText using seed 1337 and the same 10k-step training budget.

Interaction Source	NormPre-G		NormPre-L	
	Training Loss	Validation Loss	Training Loss	Validation Loss
Update before normalization (X)	3.0844	3.0900	3.1375	3.1305
Update after normalization (Ψ)	**3.0655**	**3.0711**	**3.0810**	**3.0864**

Table 8: Spectral Budget in NormPre-L. Results are reported on GPT-2 Small with OpenWebText using seed 1337 under the same 10k-step training budget.

Rank r	Training Loss	Validation Loss
8	3.0911	3.0962
16	3.0857	3.0908
32 (default)	3.0810	3.0864
64	3.0768	3.0827

Ablations on Interaction Geometry Source. We compare interaction geometry extracted from the update before normalization (X) with that extracted from the update after normalization (Ψ). Both settings apply preconditioning to the same normalized base update Ψ and differ only in the source used to construct the preconditioner.

Table 7 shows that extracting interaction geometry after normalization yields lower training and validation loss for both variants. This supports constructing the preconditioner from the geometry of the update on which it acts. Using the update before normalization causes a larger degradation for NormPre-L, suggesting greater sensitivity to the source when refinement is restricted to selected modes. The marginal scales in X can affect the spectral ordering, so its leading modes may differ from those of Ψ. Normalization removes this scale weighting first, allowing the top-r eigenspace to capture interaction structure without the influence of unequal row norms. This offers a possible explanation for the greater sensitivity of localized refinement. Together, these results support using marginal normalization both to construct the base update and to establish the interaction geometry for subsequent spectral preconditioning.

Spectral Budget in NormPre-L. Table 8 reports training and validation losses for different spectral budgets, complementing Figure 4c. Both losses decrease as r increases over the evaluated range. We use $r = 32$ as the default to balance these performance gains against the optimizer cost.

B Detailed Implementations on Spectral Preconditioning

From Section 3.2, NormPre-G realizes global spectral preconditioning by applying Newton-Schulz iteration to the normalized update Ψ, approximating $T^{\mathrm{G}} = \mathrm{msign}(\Psi)$. This follows the standard Newton-Schulz implementation used by Muon (Jordan et al., 2024; Liu et al., 2025). NormPre-L operates on a selected interaction eigenspace, which can be obtained through exact eigendecomposition or a randomized sketch. In the following, we focus on the implementation details of NormPre-L.

Exact Eigenspace Extraction. The exact implementation explicitly forms the interaction matrix $\Gamma = \Psi\Psi^\top$ and computes its eigendecomposition $\Gamma = \widetilde{U}\Lambda\widetilde{U}^\top$. Following the localized construction in Section 2.3, we define the active set $\mathcal{A} := \{i : \lambda_i > 1\}$ and let $\mathcal{C} \subseteq \mathcal{A}$ index the $\min\{r, |\mathcal{A}|\}$ largest active eigenvalues. The selected eigenpairs $(\widetilde{U}_\mathcal{C}, \Lambda_\mathcal{C})$ are directly used in the localized preconditioner

$$P^{\mathrm{L}} = I_m + \widetilde{U}_\mathcal{C}(\Lambda_\mathcal{C}^{-1/2} - I)\widetilde{U}_\mathcal{C}^\top.$$

Sketch-Based Eigenspace Extraction. To avoid forming and decomposing the full interaction matrix, we approximate its leading eigenspace through a randomized sketch (Halko et al., 2011). Among the top-r estimated interaction modes, those satisfying $\widehat{\lambda}_i > 1$ are retained for the localized preconditioner

$$\widehat{P}^{\mathrm{L}} = I_m + \widehat{U}_\mathcal{C}(\widehat{\Lambda}_\mathcal{C}^{-1/2} - I)\widehat{U}_\mathcal{C}^\top.$$

Algorithm 2 gives the complete procedure.

Algorithm 2 Randomized Interaction Sketch

Require: Marginally normalized update $\Psi \in \mathbb{R}^{m\times n}$, rank $r \le m$, oversampling o, power iterations p.

1: $\ell \leftarrow \min\{m, r+o\}$
2: Sample $\Omega \in \mathbb{R}^{n\times\ell}$ with $\Omega_{ij} \sim \mathcal{N}(0,1)$ and set $Y \leftarrow \Psi\Omega$ ▷ Gaussian Sketch
3: **for** $j = 1, \ldots, p$ **do**
4: $\quad Y \leftarrow \Psi(\Psi^\top Y)$
5: **end for** ▷ Power Iteration
6: $Q \leftarrow \mathrm{qr}(Y)$ ▷ Candidate Subspace
7: $\Gamma_Q \leftarrow (Q^\top\Psi)(Q^\top\Psi)^\top$ ▷ Rayleigh–Ritz Matrix
8: $(Z, \widehat{\Lambda}) \leftarrow$ top-r eigenpairs of Γ_Q ▷ Ritz Pairs
9: $\widehat{U} \leftarrow QZ$ ▷ Lifted Ritz Vectors
10: $\mathcal{C} \leftarrow \{i : \widehat{\lambda}_i > 1\}$
11: **return** $(\widehat{U}_\mathcal{C}, \widehat{\Lambda}_\mathcal{C})$

The main steps of Algorithm 2 are as follows:

- *Randomized Subspace Construction (Lines 2–6).* The Gaussian sketch $Y = \Psi\Omega$ probes the range of Ψ and the power iterations $Y \leftarrow \Psi(\Psi^\top Y)$ amplify the separation of its leading singular directions. Then, QR decomposition yields an orthonormal basis $Q \in \mathbb{R}^{m\times\ell}$ for the candidate interaction subspace.
- *Rayleigh–Ritz Eigenspace Extraction (Lines 7–10).* The interaction matrix is projected onto the candidate subspace through

  $$\Gamma_Q = Q^\top\Gamma Q = (Q^\top\Psi)(Q^\top\Psi)^\top \in \mathbb{R}^{\ell\times\ell}.$$

  Computing the top-r eigenpairs $(Z, \widehat{\Lambda})$ of Γ_Q gives the corresponding Ritz pairs and the Ritz vectors are lifted to the original interaction space through $\widehat{U} = QZ$. We retain the estimated modes satisfying $\widehat{\lambda}_i > 1$, yielding $(\widehat{U}_\mathcal{C}, \widehat{\Lambda}_\mathcal{C})$ for localized spectral preconditioning. Since the Ritz values are ordered in descending order, applying $\widehat{\lambda}_i > 1$ after top-r selection is equivalent to first identifying the active modes and then retaining the largest r, consistent with Section 2.3.

C Detailed Computational Complexity Analysis

This section analyzes the optimizer-side computational complexity. For NormPre, the analysis covers momentum update, relaxed tangent projection, diagonal-Gram normalization, spectral preconditioning, update RMS scaling and parameter update. We consider both NormPre-G and NormPre-L, and include Muon as a reference for the Newton-Schulz computation. Table 9 summarizes the resulting complexities.

Let $m \times n$ denote the matrix dimensions entering the spectral preconditioning stage and define $s := \min\{m, n\}$. For the Newton-Schulz implementations in Muon and NormPre-G, the input is transposed when necessary so that the Gram matrix is $s \times s$. For NormPre-L, $m \times n$ follows the current active orientation. Let q denote the number of Newton-Schulz iterations, r the interaction rank, o the oversampling parameter, p the number of power iterations and $\ell := \min\{m, r + o\}$. We omit constant factors and lower-order terms.

Table 9: Computational Complexity of Muon and NormPre Variants.

Method	**Computational Complexity**
Muon	$\mathcal{O}(mn + qmns)$
NormPre-G	$\mathcal{O}(mn + qmns)$
NormPre-L (Exact)	$\mathcal{O}(m^2n + m^3)$
NormPre-L (Sketch)	$\mathcal{O}\big((p+1)mn\ell + (m+n)\ell^2 + \ell^3\big)$

C.1 Shared Components in NormPre-G and NormPre-L

Following Algorithm 1, NormPre-G and NormPre-L share the same optimization pipeline except for the spectral preconditioning step. We analyze these shared operations under the active orientation.

- *Momentum Update and Relaxed Tangent Projection.* Updating the momentum requires $\mathcal{O}(mn)$ operations. The tangent projection is applied independently to each row with $\mathcal{O}(n)$ cost per row (inner products and vector rescaling), giving $\mathcal{O}(mn)$ operations in total.
- *Diagonal-Gram Normalization.* Computing the row norms of X and forming the normalized update $\Psi = D(X)^{-1}X$ require $\mathcal{O}(mn)$ operations.
- *Update RMS Scaling and Parameter Update.* Computing $\mathrm{RMS}(T)$ and rescaling the update both require $\mathcal{O}(mn)$ operations. Applying weight decay and updating the parameter matrix also require $\mathcal{O}(mn)$ operations.

Overall, these shared operations contribute $\mathcal{O}(mn)$ computation to both variants and are included in the total costs reported in Table 9.

C.2 NormPre-G and Muon

Muon. Muon applies a Newton-Schulz approximation of the matrix-sign transformation to its momentum update. The input is transposed when necessary so that the Gram matrix is $s \times s$. Each Newton-Schulz iteration involves matrix products between the $s \times s$ Gram matrix and the matrix update, with dominant cost $\mathcal{O}(mns)$. For q iterations, the matrix transformation requires $\mathcal{O}(qmns)$. The remaining optimizer operations (*e.g.*, momentum update, update RMS scaling and parameter update) contribute $\mathcal{O}(mn)$, giving the total complexity $\mathcal{O}(mn + qmns)$.

NormPre-G. NormPre-G applies the same q-step Newton-Schulz transformation to the normalized update Ψ. Thus, its spectral preconditioning stage requires $\mathcal{O}(qmns)$ computation. Together with the shared $\mathcal{O}(mn)$ operations analyzed in Section C.1, the total complexity is $\mathcal{O}(mn + qmns)$. Therefore, Muon and NormPre-G have the same leading asymptotic computational complexity for a fixed number of Newton-Schulz iterations. The relaxed tangent projection and diagonal-Gram normalization in NormPre-G add only $\mathcal{O}(mn)$ computation and do not change the asymptotic order.

C.3 NormPre-L

For the Exact implementation, the selected eigenvectors $\widetilde{U}_{\mathcal{C}} \in \mathbb{R}^{m\times|\mathcal{C}|}$ and eigenvalues $\Lambda_{\mathcal{C}}$ give $P^{\mathrm{L}}\Psi = \Psi + \widetilde{U}_{\mathcal{C}}\left(\Lambda_{\mathcal{C}}^{-1/2} - I\right)\widetilde{U}_{\mathcal{C}}^{\top}\Psi$. For the Sketch-based implementation, the estimated eigenpairs $(\widehat{U}_{\mathcal{C}}, \widehat{\Lambda}_{\mathcal{C}})$ yield $\widehat{P}^{\mathrm{L}}\Psi = \Psi + \widehat{U}_{\mathcal{C}}\left(\widehat{\Lambda}_{\mathcal{C}}^{-1/2} - I\right)\widehat{U}_{\mathcal{C}}^{\top}\Psi$. In both cases, the selected eigenvector matrix has dimension $m \times |\mathcal{C}|$ with $|\mathcal{C}| \le r$. Computing $\widetilde{U}_{\mathcal{C}}^{\top}\Psi$ or $\widehat{U}_{\mathcal{C}}^{\top}\Psi$ requires $\mathcal{O}(mn|\mathcal{C}|)$ operations. Applying the diagonal spectral scaling costs $\mathcal{O}(n|\mathcal{C}|)$ and multiplying by $\widetilde{U}_{\mathcal{C}}$ or $\widehat{U}_{\mathcal{C}}$ requires another $\mathcal{O}(mn|\mathcal{C}|)$ operations. Therefore, the localized spectral application costs $\mathcal{O}(mnr)$. The exact and sketch-based implementations differ in the cost of extracting the selected interaction eigenspace.

NormPre-L (Exact). The exact implementation forms the interaction matrix $\Gamma = \Psi\Psi^{\top}$, which costs $\mathcal{O}(m^2n)$. The resulting $m \times m$ matrix is then processed by a dense symmetric eigendecomposition with cost $\mathcal{O}(m^3)$. Together with the shared $\mathcal{O}(mn)$ operations analyzed in Section C.1 and the $\mathcal{O}(mnr)$ localized spectral application, we obtain $\mathcal{O}\big(mn + mnr + m^2n + m^3\big) = \mathcal{O}(m^2n + m^3)$, where the simplification follows from $r \le m$.

NormPre-L (Sketch). The sketch-based implementation avoids forming and decomposing the full $m \times m$ interaction matrix by using the randomized sketch in Algorithm 2. Its eigenspace-extraction cost can be decomposed as follows, the main intermediate dimensions are summarized in Table 10.

- *Randomized Sketch and Power Iteration.* Forming $Y = \Psi\Omega$ costs $\mathcal{O}(mn\ell)$. Each power iteration computes $\Psi^{\top}Y$ followed by $\Psi(\Psi^{\top}Y)$, with both matrix multiplications requiring $\mathcal{O}(mn\ell)$ operations. Over p iterations, the power-iteration cost is $\mathcal{O}(pmn\ell)$.
- *Candidate Subspace.* Computing the QR decomposition of $Y \in \mathbb{R}^{m\times\ell}$ costs $\mathcal{O}(m\ell^2)$.
- *Rayleigh–Ritz Matrix.* Forming $Q^{\top}\Psi \in \mathbb{R}^{\ell\times n}$ costs $\mathcal{O}(mn\ell)$ and constructing $\Gamma_Q = (Q^{\top}\Psi)(Q^{\top}\Psi)^{\top} \in \mathbb{R}^{\ell\times\ell}$ costs $\mathcal{O}(n\ell^2)$.
- *Rayleigh–Ritz Extraction.* A dense eigendecomposition of $\Gamma_Q \in \mathbb{R}^{\ell\times\ell}$ costs $\mathcal{O}(\ell^3)$. Computing $\widehat{U} = QZ$ costs $\mathcal{O}(m\ell r)$, which is bounded by $\mathcal{O}(m\ell^2)$ because $r \le \ell$.

We have $\mathcal{O}\big(mn + mnr + (p+1)mn\ell + (m+n)\ell^2 + \ell^3\big)$ when combining these terms with the shared $\mathcal{O}(mn)$ operations and $\mathcal{O}(mnr)$ localized spectral application. Since $r \le \ell$, the $\mathcal{O}(mn)$ and $\mathcal{O}(mnr)$ terms are absorbed into the sketching cost. Therefore, the total computational complexity of NormPre-L (Sketch) is $\mathcal{O}\big((p+1)mn\ell + (m+n)\ell^2 + \ell^3\big)$.

Table 10: Dimensions of the Main Quantities in Algorithm 2

Symbol	Dimension	Description
Ψ	$m \times n$	Marginally normalized update
Ω	$n \times \ell$	Gaussian random matrix
Y	$m \times \ell$	Gaussian sketch
Q	$m \times \ell$	Candidate subspace
Γ_Q	$\ell \times \ell$	Rayleigh–Ritz matrix
Z	$\ell \times r$	Top-r eigenvectors of Γ_Q
$\widehat{\Lambda}$	$r \times r$	Estimated eigenvalues
$\widehat{U}$	$m \times r$	Estimated eigenvectors

D PROOFS

D.1 OPTIMIZATION FORMULATION OF LOCALIZED SPECTRAL REFINEMENT

Proposition 1 (Regularized Spectral Steepest Descent). *Suppose that $\Gamma(X) \succ 0$. The closest update to Ψ within the unit spectral-norm ball is*

$$T_+ := \arg\min_{\|T\|_{\mathrm{op}} \le 1} \frac{1}{2}\|T - \Psi\|_F^2 = \arg\max_{\|T\|_{\mathrm{op}} \le 1} \left\{ \langle \Psi, T \rangle_F - \frac{1}{2}\|T\|_F^2 \right\}. \tag{6}$$

Its unique solution is $T_+ = P_+\Psi$, where the preconditioner is

$$P_+ = \widetilde{U}\,\mathrm{Diag}\left(\min\{1, \lambda_1^{-1/2}\}, \ldots, \min\{1, \lambda_m^{-1/2}\}\right)\widetilde{U}^\top. \tag{7}$$

In particular, $0 \prec P_+ \preceq I_m$ and $\sigma_i(T_+) = \min\{\widetilde{\sigma}_i, 1\}$.

Remark 3. Proposition 1 gives the exact projection of Ψ onto the unit spectral-norm ball, clipping singular values above one to one while leaving the remaining singular values unchanged. Define the active set $\mathcal{A} := \{i : \lambda_i > 1\}$, together with $\widetilde{U}_{\mathcal{A}} := [\widetilde{u}_i]_{i\in\mathcal{A}}$ and $\Lambda_{\mathcal{A}} := \mathrm{Diag}(\lambda_i)_{i\in\mathcal{A}}$. Equivalently, the preconditioner can be written as

$$P_+ = I_m + \widetilde{U}_{\mathcal{A}}(\Lambda_{\mathcal{A}}^{-1/2} - I)\widetilde{U}_{\mathcal{A}}^\top.$$

Its non-identity component is $\Delta_{\mathcal{A}} := P_+ - I_m$. Hence, $P_+\Psi = \Psi + \Delta_{\mathcal{A}}\Psi$, where $\Delta_{\mathcal{A}}\Psi$ is the induced update refinement relative to the normalized update Ψ.

Proof. Let $\Psi = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top$ be a thin SVD, where $\widetilde{\Sigma} = \mathrm{Diag}(\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_m)$ with $\widetilde{\sigma}_1 \ge \cdots \ge \widetilde{\sigma}_m > 0$. Suppose that $\Gamma(X) = \Psi\Psi^\top = \widetilde{U}\Lambda\widetilde{U}^\top \succ 0$, i.e., Ψ has full row rank and $\lambda_i = \widetilde{\sigma}_i^2 > 0$ for every i.

For the regularized steepest-descent problem,

$$\frac{1}{2}\|T - \Psi\|_F^2 = \frac{1}{2}\|\Psi\|_F^2 - \left(\langle \Psi, T \rangle_F - \frac{1}{2}\|T\|_F^2\right).$$

Since $\|\Psi\|_F^2/2$ is independent of T, the two optimization problems in Equation 6 are equivalent.

Following the standard spectral steepest-descent derivation of the matrix-sign update, we solve the regularized problem under the same unit spectral-norm constraint by optimizing its singular values modewise.

Let $\tau_1, \ldots, \tau_m$ denote the singular values of a feasible T. The constraint $\|T\|_{\mathrm{op}} \le 1$ implies $0 \le \tau_i \le 1$. By von Neumann's trace inequality, $\langle \Psi, T \rangle_F \le \sum_{i=1}^m \widetilde{\sigma}_i\tau_i$, with equality when T shares the singular directions of Ψ. Since $\|T\|_F^2 = \sum_{i=1}^m \tau_i^2$, the regularized objective satisfies

$$\langle \Psi, T \rangle_F - \frac{1}{2}\|T\|_F^2 \le \sum_{i=1}^m \left(\widetilde{\sigma}_i\tau_i - \frac{1}{2}\tau_i^2\right).$$

Thus, the matrix problem reduces to independently maximizing $\widetilde{\sigma}_i\tau_i - \frac{1}{2}\tau_i^2$ over $0 \le \tau_i \le 1$ for each singular mode. For each i, consider the scalar objective $f_i(\tau_i) := \widetilde{\sigma}_i\tau_i - \frac{1}{2}\tau_i^2$. Since $f_i'(\tau_i) = \widetilde{\sigma}_i - \tau_i$ and $f_i''(\tau_i) = -1 < 0$, its unique unconstrained maximizer is $\tau_i = \widetilde{\sigma}_i$. With the unit spectral-norm constraint $0 \le \tau_i \le 1$, we have

$$\tau_i^\star = \min\{\widetilde{\sigma}_i, 1\}.$$

Therefore, the unique solution is

$$T_+ = \widetilde{U}\,\mathrm{Diag}\left(\min\{\widetilde{\sigma}_1, 1\}, \ldots, \min\{\widetilde{\sigma}_m, 1\}\right)\widetilde{V}^\top.$$

The uniqueness follows from the strict concavity of $\langle \Psi, T \rangle_F - \frac{1}{2}\|T\|_F^2$ over the convex feasible set.

We next express the optimal solution as a left spectral preconditioning of Ψ. Since $\lambda_i = \widetilde{\sigma}_i^2$, define

$$P_+ := \widetilde{U}\,\mathrm{Diag}\left(\min\{1, \lambda_1^{-1/2}\}, \ldots, \min\{1, \lambda_m^{-1/2}\}\right)\widetilde{U}^\top.$$

Then $P_+\Psi = \widetilde{U}\,\mathrm{Diag}\left(\min\{1,\widetilde{\sigma}_1^{-1}\}\widetilde{\sigma}_1,\ldots,\min\{1,\widetilde{\sigma}_m^{-1}\}\widetilde{\sigma}_m\right)\widetilde{V}^\top = T_+$, i.e., $T_+ = P_+\Psi$. Since

$$\min\{1,\lambda_i^{-1/2}\} = \begin{cases} 1, & \lambda_i \le 1, \\ \lambda_i^{-1/2}, & \lambda_i > 1, \end{cases}$$

letting $\mathcal{A} := \{i : \lambda_i > 1\}$ gives the equivalent representation of the preconditioner

$$P_+ = I_m + \widetilde{U}_{\mathcal{A}}\left(\Lambda_{\mathcal{A}}^{-1/2} - I\right)\widetilde{U}_{\mathcal{A}}^\top.$$

Thus, P_+ leaves modes with $\lambda_i \le 1$ unchanged and contracts modes with $\lambda_i > 1$ by $\lambda_i^{-1/2}$. Moreover, every eigenvalue of P_+ lies in $(0,1]$, so $0 \prec P_+ \preceq I_m$. □

Proposition 2 (Rank-Constrained Spectral Refinement)**.** *Suppose that $\Gamma(X) \succ 0$. A closest rank-constrained update refinement to $\Delta_{\mathcal{A}}\Psi$ is obtained by solving*

$$\Delta^\star \in \arg\min_{\mathrm{rank}(\Delta)\le r} \frac{1}{2}\left\|(\Delta - \Delta_{\mathcal{A}})\Psi\right\|_F^2. \tag{8}$$

Let $\mathcal{C} \subseteq \mathcal{A}$ index the $\min\{r,|\mathcal{A}|\}$ largest eigenvalues satisfying $\lambda_i > 1$. Then one solution is

$$\Delta^\star = \widetilde{U}_{\mathcal{C}}\left(\Lambda_{\mathcal{C}}^{-1/2} - I\right)\widetilde{U}_{\mathcal{C}}^\top. \tag{9}$$

The resulting preconditioner and update are $P := I_m + \Delta^\star$ and $T := P\Psi$.

Remark 4. Proposition 2 characterizes the optimal rank-constrained spectral refinement. Since $\Gamma(X) \succ 0$, Ψ has full row rank and $\mathrm{rank}(\Delta\Psi) = \mathrm{rank}(\Delta)$, so the rank constraint on the non-identity component is equivalent to that on its induced update refinement. The result then follows from the Eckart–Young–Mirsky theorem (Eckart & Young, 1936; Mirsky, 1960) applied to $\Delta_{\mathcal{A}}\Psi$. Its nonzero singular values are $|1-\sqrt{\lambda_i}|$ for $i \in \mathcal{A}$ and increase monotonically with λ_i, so the optimal truncation retains the active modes associated with the largest λ_i indexed by $\mathcal{C}$.

Proof. Recall that $\Psi = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top$ is a thin SVD, where $\widetilde{\Sigma} = \mathrm{Diag}(\widetilde{\sigma}_1,\ldots,\widetilde{\sigma}_m)$ with $\widetilde{\sigma}_1 \ge \cdots \ge \widetilde{\sigma}_m > 0$. Suppose that $\Gamma(X) = \Psi\Psi^\top = \widetilde{U}\Lambda\widetilde{U}^\top \succ 0$, i.e., Ψ has full row rank and $\lambda_i = \widetilde{\sigma}_i^2 > 0$ for every i. From Proposition 1, with $\mathcal{A} := \{i : \lambda_i > 1\}$ and $P_+ = I_m + \widetilde{U}_{\mathcal{A}}\left(\Lambda_{\mathcal{A}}^{-1/2} - I\right)\widetilde{U}_{\mathcal{A}}^\top$,

$$\Delta_{\mathcal{A}} = P_+ - I_m = \widetilde{U}_{\mathcal{A}}\left(\Lambda_{\mathcal{A}}^{-1/2} - I\right)\widetilde{U}_{\mathcal{A}}^\top.$$

Applying this refinement to Ψ, and using that $\widetilde{U}_{\mathcal{A}}$ and $\widetilde{V}_{\mathcal{A}}$ collect the singular directions indexed by $\mathcal{A}$, we have

$$\begin{aligned}\Delta_{\mathcal{A}}\Psi &= \widetilde{U}_{\mathcal{A}}\left(\Lambda_{\mathcal{A}}^{-1/2} - I\right)\widetilde{U}_{\mathcal{A}}^\top\widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top \\ &= \widetilde{U}_{\mathcal{A}}\left(\Lambda_{\mathcal{A}}^{-1/2} - I\right)\widetilde{\Sigma}_{\mathcal{A}}\widetilde{V}_{\mathcal{A}}^\top \\ &= \widetilde{U}_{\mathcal{A}}\,\mathrm{Diag}\left(1-\widetilde{\sigma}_i\right)_{i\in\mathcal{A}}\widetilde{V}_{\mathcal{A}}^\top,\end{aligned}$$

where the last equality follows from $\lambda_i = \widetilde{\sigma}_i^2$. Since $\lambda_i > 1$ for $i \in \mathcal{A}$, we have $\widetilde{\sigma}_i = \sqrt{\lambda_i} > 1$. Therefore, the nonzero singular values of $\Delta_{\mathcal{A}}\Psi$ are $|1-\widetilde{\sigma}_i| = \sqrt{\lambda_i} - 1$ for $i \in \mathcal{A}$.

For any feasible Δ with $\mathrm{rank}(\Delta) \le r$, we have $\mathrm{rank}(\Delta\Psi) \le \mathrm{rank}(\Delta) \le r$. Replacing $\Delta\Psi$ by an arbitrary matrix Z with rank at most r enlarges the feasible set. Therefore,

$$\min_{\mathrm{rank}(Z)\le r}\|Z - \Delta_{\mathcal{A}}\Psi\|_F^2 \le \min_{\mathrm{rank}(\Delta)\le r}\|\Delta\Psi - \Delta_{\mathcal{A}}\Psi\|_F^2.$$

By the Eckart–Young–Mirsky theorem, the left-hand side is minimized by retaining the $\min\{r,|\mathcal{A}|\}$ largest singular values of $\Delta_{\mathcal{A}}\Psi$, whose rank is $|\mathcal{A}|$. Since $\sqrt{\lambda_i} - 1$ is increasing in λ_i, the optimal rank-r approximation retains the modes with the largest active eigenvalues. Let $\mathcal{C} \subseteq \mathcal{A}$ index these modes. The resulting truncated approximation is

$$Z^\star = \widetilde{U}_{\mathcal{C}}\,\mathrm{Diag}\left(1-\widetilde{\sigma}_i\right)_{i\in\mathcal{C}}\widetilde{V}_{\mathcal{C}}^\top.$$

Define $\Delta^\star := \widetilde{U}_{\mathcal{C}} \left(\Lambda_{\mathcal{C}}^{-1/2} - I \right) \widetilde{U}_{\mathcal{C}}^\top$. Since $|\mathcal{C}| \le r$, we have $\operatorname{rank}(\Delta^\star) \le r$, so $\Delta^\star$ is feasible for the original problem. Moreover,

$$\begin{aligned} \Delta^\star \Psi &= \widetilde{U}_{\mathcal{C}} \left(\Lambda_{\mathcal{C}}^{-1/2} - I \right) \widetilde{\Sigma}_{\mathcal{C}} \widetilde{V}_{\mathcal{C}}^\top \\ &= \widetilde{U}_{\mathcal{C}} \operatorname{Diag}\left(1 - \widetilde{\sigma}_i\right)_{i \in \mathcal{C}} \widetilde{V}_{\mathcal{C}}^\top \\ &= Z^\star, \end{aligned}$$

Thus, $\Delta^\star$ attains the optimal rank-r approximation obtained from the relaxed problem. Together with the preceding lower bound, this gives

$$\min_{\operatorname{rank}(\Delta) \le r} \|\Delta\Psi - \Delta_{\mathcal{A}}\Psi\|_F^2 = \min_{\operatorname{rank}(Z) \le r} \|Z - \Delta_{\mathcal{A}}\Psi\|_F^2.$$

Therefore, $\Delta^\star$ is a solution to Equation 8. The resulting preconditioner and refined update are

$$P = I_m + \Delta^\star, \quad T = P\Psi,$$

where $\Delta^\star := \widetilde{U}_{\mathcal{C}} \left(\Lambda_{\mathcal{C}}^{-1/2} - I \right) \widetilde{U}_{\mathcal{C}}^\top$. □

D.2 Convergence Analysis of NormPre

This section establishes convergence guarantees for NormPre-G and NormPre-L as stated in Theorem 1, with the Sketch-based implementation for NormPre-L given in Corollary 1. Throughout this convergence analysis, we use a fixed orientation $k_t \equiv k$ for both variants. For NormPre-G, we use the exact matrix-sign transformation and assume $\Gamma_t \succ 0$ for all t.

In the deterministic setting without momentum, let $G_t := \nabla L(W_t)$ and define the actively oriented parameter and gradient as $\overline{W}_t := \mathcal{O}_{k_t}(W_t)$ and $\overline{G}_t := \mathcal{O}_{k_t}(G_t)$. For their i-th rows $\bar{w}_{t,i}$ and $\bar{g}_{t,i}$, respectively, define the relaxed tangent gradient

$$x_{t,i} := \bar{g}_{t,i} - \langle \bar{g}_{t,i}, \bar{w}_{t,i} \rangle \bar{w}_{t,i}.$$

Decompose $\bar{g}_{t,i}$ into the tangent and radial components with respect to $\bar{w}_{t,i}$:

$$\bar{g}_{t,i} = g_{t,i}^{\tan} + g_{t,i}^{\mathrm{rad}}, \quad g_{t,i}^{\mathrm{rad}} := \frac{\langle \bar{g}_{t,i}, \bar{w}_{t,i} \rangle}{\|\bar{w}_{t,i}\|_2^2} \bar{w}_{t,i}.$$

Then $x_{t,i} = g_{t,i}^{\tan} + \left(1 - \|\bar{w}_{t,i}\|_2^2\right) g_{t,i}^{\mathrm{rad}}$. Thus, $x_{t,i}$ preserves the tangent component of $\bar{g}_{t,i}$ while retaining a norm-dependent radial component. In particular, it reduces to the exact tangent projection when $\|\bar{w}_{t,i}\|_2 = 1$. Collecting the resulting rows gives $X_t := \mathcal{T}_{W_t,k_t}(G_t)$.

As in Section 2, we present the derivation for matrices in $\mathbb{R}^{m \times n}$ under the row orientation; the column-oriented case follows by applying the same argument to the transpose. We assume throughout that the row normalization is well defined. Through row normalization of X_t, we get the normalized update $\Psi_t := D(X_t)^{-1} X_t$ and its i-th row $\psi_{t,i}$. Applying spectral preconditioning, we get the refined update T_t. As Algorithm 1, the resulting chain is

$$\overline{G}_t \xrightarrow{\text{Relaxed Tangent Projection}} X_t \xrightarrow{\text{Diagonal-Gram Normalization}} \Psi_t \xrightarrow{\text{Spectral Preconditioning}} T_t.$$

The following lemmas quantify the gradient contribution retained by the normalized update, the spectral properties of the global and localized refinements and the radial interaction induced by each preconditioner. These quantities are used to lower bound the gradient-update inner product and control the smoothness term in the subsequent descent analyses.

D.2.1 Lemmas

Lemma 1. *Let $\phi_{t,i}$ denote the angle between $\bar{g}_{t,i}$ and $\bar{w}_{t,i}$. Suppose $\sin(\phi_{t,i}) \ge \gamma > 0$ for every row i, where γ represents the tangential fraction. With the diagonal-gram normalization $\Psi_t := D(X_t)^{-1} X_t$ in Algorithm 1,*

$$\langle X_t, \Psi_t \rangle_F \ge \gamma \|\overline{G}_t\|_F.$$

Proof. Since $x_{t,i}$ preserves the tangent component of $\bar{g}_{t,i}$ while retaining a norm-dependent radial component along $\bar{w}_{t,i}$, letting $\phi_{t,i}$ denote the angle between $\bar{g}_{t,i}$ and $\bar{w}_{t,i}$ gives

$$\|x_{t,i}\|_2^2 = \|\bar{g}_{t,i}\|_2^2 \left[\sin^2(\phi_{t,i}) + \left(1 - \|\bar{w}_{t,i}\|_2^2\right)^2 \cos^2(\phi_{t,i})\right] \geq \|\bar{g}_{t,i}\|_2^2 \sin^2(\phi_{t,i}).$$

Thus, we get $\|x_{t,i}\|_2 \geq \|\bar{g}_{t,i}\|_2 \sin(\phi_{t,i})$. Moreover, since $\Psi_t = D(X_t)^{-1} X_t$, its i-th row satisfies $\psi_{t,i} = x_{t,i}/\|x_{t,i}\|_2$. Then

$$\begin{aligned}\langle X_t, \Psi_t \rangle_F &= \sum_i \langle x_{t,i}, \psi_{t,i} \rangle = \sum_i \left\langle x_{t,i}, \frac{x_{t,i}}{\|x_{t,i}\|_2} \right\rangle = \sum_i \|x_{t,i}\|_2 \\ &\geq \sum_i \|\bar{g}_{t,i}\|_2 \sin(\phi_{t,i}).\end{aligned}$$

The lower bound becomes zero when the full gradient vanishes or when the gradient becomes parallel to the corresponding weight direction, $\sin(\phi_{t,i}) = 0$. If we assume that the gradient is not perfectly aligned with the weight, *i.e.*, the gradient maintains a tangential fraction $\sin(\phi_{t,i}) \geq \gamma > 0$, we obtain

$$\langle X_t, \Psi_t \rangle_F \geq \gamma \sum_i \|\bar{g}_{t,i}\|_2 \geq \gamma \left(\sum_i \|\bar{g}_{t,i}\|_2^2 \right)^{1/2} = \gamma \|\overline{G}_t\|_F.$$

Thus, this lemma quantifies the first-order gradient signal retained through the relaxed tangent projection and row normalization. □

Lemma 2. *Consider the* NormPre-G *update with matrix-sign transformation and the* NormPre-L *update with Exact implementation in Algorithm 1 and Section 3.2. For* NormPre-G*, suppose* $\Gamma_t = \Psi_t \Psi_t^\top \succ 0$*. Let* P_t *denote the corresponding spectral preconditioner and* $T_t = P_t \Psi_t$*. Then*

(a) Inner product with relaxed tangent gradient: $\langle X_t, T_t \rangle_F \geq \lambda_{\min}(P_t) \langle X_t, \Psi_t \rangle_F$.

(b) Norm of the refined update: $\|T_t\|_F \leq \max\{m, n\}^{1/2}$.

(c) Minimum eigenvalue of preconditioner: $\lambda_{\min}(P_t) \geq \lambda_{t,1}^{-1/2} \geq \max\{m, n\}^{-1/2}$*, where the first inequality holds with equality for* NormPre-G.

Proof. Recall that X_t is the relaxed tangent gradient under the active orientation, $\Psi_t = D(X_t)^{-1} X_t$ is the normalized update and $T_t = P_t \Psi_t$ is the update after spectral preconditioning. We follow the chain

$$\overline{G}_t \xrightarrow{\text{Relaxed Tangent Projection}} X_t \xrightarrow{\text{Diagonal-Gram Normalization}} \Psi_t \xrightarrow{\text{Spectral Preconditioning}} T_t.$$

We first characterize how spectral preconditioning changes the inner product with the relaxed tangent gradient. Let $\Gamma_t = \Psi_t \Psi_t^\top = \widetilde{U}_t \Lambda_t \widetilde{U}_t^\top$, and let $\tilde{u}_{t,i}$ denote the i-th column of $\widetilde{U}_t$, with corresponding eigenvalue $\lambda_{t,i}$ of Γ_t. For NormPre-G, $\Gamma_t \succ 0$ and

$$P_t = \Gamma_t^{-1/2} = \widetilde{U}_t \Lambda_t^{-1/2} \widetilde{U}_t^\top.$$

We have $P_t \tilde{u}_{t,i} = \lambda_{t,i}^{-1/2} \tilde{u}_{t,i}$ for every i. For NormPre-L,

$$P_t = I + \widetilde{U}_{\mathcal{C}_t} (\Lambda_{\mathcal{C}_t}^{-1/2} - I) \widetilde{U}_{\mathcal{C}_t}^\top.$$

For $i \in \mathcal{C}_t$, $P_t \tilde{u}_{t,i} = \lambda_{t,i}^{-1/2} \tilde{u}_{t,i}$, while for $i \notin \mathcal{C}_t$, $P_t \tilde{u}_{t,i} = \tilde{u}_{t,i}$. Since every selected mode satisfies $\lambda_{t,i} > 1$, the eigenvalues of P_t are $\lambda_{t,i}^{-1/2} \in (0, 1)$ on the selected modes and 1 on the unselected modes.

Therefore, for both realizations, P_t and Γ_t share the same eigenbasis and $P_t \succ 0$. Since $P_t \succeq \lambda_{\min}(P_t) I$,

$$P_t \Gamma_t \succeq \lambda_{\min}(P_t) \Gamma_t. \tag{10}$$

With $\Psi_t = D(X_t)^{-1}X_t$, let $D_t := D(X_t) = \mathrm{Diag}\left(\|x_{t,1}\|_2, \ldots, \|x_{t,m}\|_2\right)$. Then $X_t = D_t\Psi_t$ and $D_t \succeq 0$. Using $T_t = P_t\Psi_t$ and $\Gamma_t = \Psi_t\Psi_t^\top$,

$$\begin{aligned}\langle X_t, T_t\rangle_F &= \mathrm{tr}\big(X_t^\top T_t\big) = \mathrm{tr}\big((D_t\Psi_t)^\top P_t\Psi_t\big)\\ &= \mathrm{tr}\big(D_tP_t\Psi_t\Psi_t^\top\big) = \mathrm{tr}(D_tP_t\Gamma_t).\end{aligned} \tag{11}$$

Together with Equation 10, we have $\mathrm{tr}(D_t\left(P_t\Gamma_t - \lambda_{\min}(P_t)\Gamma_t\right)) \geq 0$, and thus

$$\mathrm{tr}(D_tP_t\Gamma_t) \geq \lambda_{\min}(P_t)\,\mathrm{tr}(D_t\Gamma_t). \tag{12}$$

Since each row of Ψ_t has unit norm, $\mathrm{diag}(\Gamma_t) = \mathbf{1}$. Then

$$\mathrm{tr}(D_t\Gamma_t) = \sum_i (D_t)_{ii}(\Gamma_t)_{ii} = \sum_i \|x_{t,i}\|_2. \tag{13}$$

Combining Equations 11–13 gives $\langle X_t, T_t\rangle_F \geq \lambda_{\min}(P_t)\sum_i \|x_{t,i}\|_2$. Since $\psi_{t,i} = x_{t,i}/\|x_{t,i}\|_2$,

$$\langle X_t, \Psi_t\rangle_F = \sum_i \langle x_{t,i}, \psi_{t,i}\rangle = \sum_i \|x_{t,i}\|_2.$$

Therefore,

$$\langle X_t, T_t\rangle_F \geq \lambda_{\min}(P_t)\langle X_t, \Psi_t\rangle_F. \tag{14}$$

This inequality quantifies the change in the inner product with the relaxed tangent gradient induced by exact spectral preconditioning: $\langle X_t, \Psi_t\rangle_F$ before preconditioning is transformed into $\langle X_t, T_t\rangle_F$ after preconditioning, while retaining at least a $\lambda_{\min}(P_t)$ fraction. Together with Lemma 1, this provides the positive first-order descent signal used later in the descent analysis.

We next bound the norm of the refined update T_t. Since P_t is symmetric and $T_t = P_t\Psi_t$,

$$T_tT_t^\top = P_t\Psi_t\Psi_t^\top P_t = P_t\Gamma_tP_t.$$

Recall that P_t and Γ_t share the same eigenbasis, and let $p_{t,i}$ denote the eigenvalue of P_t associated with $\tilde{u}_{t,i}$. Then $T_tT_t^\top = \widetilde{U}_t\,\mathrm{Diag}(p_{t,i}^2\lambda_{t,i})\widetilde{U}_t^\top$.

For NormPre-G, $p_{t,i} = \lambda_{t,i}^{-1/2}$ for every i. We have $p_{t,i}^2\lambda_{t,i} = 1$ and

$$\|T_t\|_F^2 = \mathrm{tr}(T_tT_t^\top) = m.$$

For NormPre-L, $p_{t,i} = \lambda_{t,i}^{-1/2}$ for $i \in \mathcal{C}_t$ and $p_{t,i} = 1$ for $i \notin \mathcal{C}_t$. We have

$$\|T_t\|_F^2 = \mathrm{tr}(T_tT_t^\top) = \sum_{i\in\mathcal{C}_t} 1 + \sum_{i\notin\mathcal{C}_t}\lambda_{t,i} \leq \sum_i \lambda_{t,i} = \mathrm{tr}(\Gamma_t) = m.$$

The inequality follows because every selected mode satisfies $\lambda_{t,i} > 1$. Therefore, for both realizations,

$$\|T_t\|_F \leq m^{1/2} \leq \max\{m,n\}^{1/2},$$

which later controls the quadratic term in the smoothness inequality.

Finally, we bound the minimum eigenvalue of the preconditioner. For NormPre-G, the eigenvalues of $P_t = \Gamma_t^{-1/2}$ are $\lambda_{t,i}^{-1/2}$. We have $\lambda_{\min}(P_t) = \lambda_{t,1}^{-1/2}$. For NormPre-L, the eigenvalues of P_t are $\lambda_{t,i}^{-1/2}$ on the selected modes and 1 on the unselected modes. Since $\lambda_{t,i} \leq \lambda_{t,1}$ for every selected mode and $\lambda_{t,1} \geq 1$, every eigenvalue of P_t is at least $\lambda_{t,1}^{-1/2}$. Therefore, for both realizations,

$$\lambda_{\min}(P_t) \geq \lambda_{t,1}^{-1/2}.$$

Since $\mathrm{tr}(\Gamma_t) = m$, we have $\lambda_{t,1} \leq m$. Thus,

$$\lambda_{\min}(P_t) \geq \lambda_{t,1}^{-1/2} \geq m^{-1/2} \geq \max\{m,n\}^{-1/2}.$$

□

Lemma 3. *Under the setting of Lemma 2, let* $\nu_t := \max_i \frac{\|\bar{g}_{t,i} - x_{t,i}\|_2}{\|x_{t,i}\|_2}$ *denote the maximum radial ratio. Let* $\phi'_{t,j,i}$ *denote the angle between* $\psi_{t,j}$ *and* $\bar{w}_{t,i}$*, and suppose* $\max_i(\sum_j \cos^2(\phi'_{t,j,i}))^{1/2} \le \gamma'$*. For* NormPre-G*, let* $\kappa_t := \sqrt{\lambda_{t,1}/\lambda_{t,m}}$ *and define*

$$\chi_t := \begin{cases} \kappa_t, & \text{NormPre-G,} \\ \sqrt{\lambda_{t,1}}, & \text{NormPre-L.} \end{cases}$$

Then, we have

$$\left|\langle \overline{G}_t - X_t, T_t \rangle_F\right| \le \nu_t \gamma' \chi_t \langle X_t, T_t \rangle_F.$$

Proof. Recall that X_t is the rowwise relaxed tangent gradient of $\overline{G}_t$ and $\overline{G}_t - X_t$ is the corresponding radial residual. Using $T_t = P_t \Psi_t$, we have

$$\langle \overline{G}_t - X_t, T_t \rangle_F = \langle \overline{G}_t - X_t, P_t \Psi_t \rangle_F.$$

By the definition of ν_t,

$$\frac{\|\bar{g}_{t,i} - x_{t,i}\|_2}{\|x_{t,i}\|_2} \le \nu_t,$$

or equivalently, $\|\bar{g}_{t,i} - x_{t,i}\|_2 \le \nu_t \|x_{t,i}\|_2$.

We next bound the interaction between the radial residual and the spectrally refined update. From the definition of the relaxed tangent projection, $\bar{g}_{t,i} - x_{t,i} = \langle \bar{g}_{t,i}, \bar{w}_{t,i} \rangle \bar{w}_{t,i}$. Let e_i denote the i-th standard basis vector. Expanding the Frobenius inner product rowwise gives

$$\begin{aligned} \left|\langle \overline{G}_t - X_t, T_t \rangle_F\right| &= \left|\langle \overline{G}_t - X_t, P_t \Psi_t \rangle_F\right| = \left|\sum_i \left\langle \bar{g}_{t,i} - x_{t,i}, e_i^\top P_t \Psi_t \right\rangle\right| \\ &= \left|\sum_i \langle \bar{g}_{t,i}, \bar{w}_{t,i} \rangle e_i^\top P_t \Psi_t \bar{w}_{t,i}^\top\right| \le \sum_i \|\bar{g}_{t,i} - x_{t,i}\|_2 \|e_i^\top P_t\|_2 \frac{\|\Psi_t \bar{w}_{t,i}^\top\|_2}{\|\bar{w}_{t,i}\|_2} \\ &\le \nu_t \|P_t\|_{\mathrm{op}} \sum_i \|x_{t,i}\|_2 \frac{\|\Psi_t \bar{w}_{t,i}^\top\|_2}{\|\bar{w}_{t,i}\|_2}. \end{aligned}$$

For all j and i, let $\phi'_{t,j,i}$ denote the angle between $\psi_{t,j}$ and $\bar{w}_{t,i}$. Since each row of Ψ_t has unit norm,

$$\frac{\|\Psi_t \bar{w}_{t,i}^\top\|_2}{\|\bar{w}_{t,i}\|_2} = \left(\sum_j \cos^2(\phi'_{t,j,i})\right)^{1/2} \le \gamma'.$$

This quantity measures the aggregate non-orthogonality between the i-th weight direction and the normalized relaxed tangent gradients. A smaller γ' corresponds to stronger orthogonality between the normalized relaxed tangent gradients and the weight directions. Therefore,

$$\left|\langle \overline{G}_t - X_t, T_t \rangle_F\right| \le \nu_t \gamma' \|P_t\|_{\mathrm{op}} \sum_i \|x_{t,i}\|_2. \tag{15}$$

From Lemma 2(a), $\langle X_t, T_t \rangle_F \ge \lambda_{\min}(P_t) \sum_i \|x_{t,i}\|_2$. Combining this inequality with Equation 15 gives

$$\left|\langle \overline{G}_t - X_t, T_t \rangle_F\right| \le \nu_t \gamma' \frac{\|P_t\|_{\mathrm{op}}}{\lambda_{\min}(P_t)} \langle X_t, T_t \rangle_F.$$

For NormPre-G, $P_t = \Gamma_t^{-1/2}$ and the eigenvalues of Γ_t satisfy $\lambda_{t,1} \ge \cdots \ge \lambda_{t,m} > 0$. We have

$$\|P_t\|_{\mathrm{op}} = \lambda_{t,m}^{-1/2}, \quad \lambda_{\min}(P_t) = \lambda_{t,1}^{-1/2}.$$

Therefore,

$$\frac{\|P_t\|_{\mathrm{op}}}{\lambda_{\min}(P_t)} = \sqrt{\frac{\lambda_{t,1}}{\lambda_{t,m}}} = \kappa_t.$$

For NormPre-L, all eigenvalues of P_t lie in $(0, 1]$, so $\|P_t\|_{\mathrm{op}} \le 1$. From Lemma 2(c), $\lambda_{\min}(P_t) \ge \lambda_{t,1}^{-1/2}$. Therefore,

$$\frac{\|P_t\|_{\mathrm{op}}}{\lambda_{\min}(P_t)} \le \sqrt{\lambda_{t,1}}.$$

Combining the two cases gives

$$\left|\langle \overline{G}_t - X_t, T_t\rangle_F\right| \le \nu_t \gamma' \chi_t \langle X_t, T_t\rangle_F.$$

In summary, the radial interaction in the spectrally refined update is controlled relative to the inner product between the refined update and the relaxed tangent gradient. Its relative magnitude is determined by the radial ratio ν_t, the non-orthogonality γ' between the normalized row $\psi_{t,j}$ and the weight $\bar{w}_{t,i}$ and the spectral factor χ_t. □

D.2.2 Convergence of NormPre-G and NormPre-L (Exact)

We now combine Lemmas 1, 2 and 3 to prove the NormPre-G and NormPre-L (Exact) cases of Theorem 1.

Proof. We consider NormPre-G with matrix-sign transformation and NormPre-L with Exact implementation in the deterministic setting without momentum, RMS scaling or weight decay. Let $\chi_t = \kappa_t$ for NormPre-G and $\chi_t = \sqrt{\lambda_{t,1}}$ for NormPre-L, as in Lemma 3. Taking $G_t = \nabla \mathcal{L}(W_t)$, the final update is

$$W_{t+1} = W_t - \eta\, \mathcal{O}_{k_t}^{-1}(T_t).$$

Since $\mathcal{O}_{k_t}$ is either the identity or transpose, it preserves Frobenius inner products and norms. Hence, $\langle G_t, \mathcal{O}_{k_t}^{-1}(T_t)\rangle_F = \langle \overline{G}_t, T_t\rangle_F$ and $\|\mathcal{O}_{k_t}^{-1}(T_t)\|_F = \|T_t\|_F$. Moreover, $\|\overline{G}_t\|_F = \|G_t\|_F = \|\nabla \mathcal{L}(W_t)\|_F$.

By the L-smoothness of $\mathcal{L}$,

$$\begin{aligned}
\mathcal{L}(W_{t+1}) &\le \mathcal{L}(W_t) + \langle \nabla \mathcal{L}(W_t), W_{t+1} - W_t\rangle_F + \frac{L}{2}\|W_{t+1} - W_t\|_F^2 \\
&= \mathcal{L}(W_t) - \eta \left\langle G_t, \mathcal{O}_{k_t}^{-1}(T_t)\right\rangle_F + \frac{L\eta^2}{2}\left\|\mathcal{O}_{k_t}^{-1}(T_t)\right\|_F^2 \\
&= \mathcal{L}(W_t) - \eta \langle \overline{G}_t, T_t\rangle_F + \frac{L\eta^2}{2}\|T_t\|_F^2. \qquad (16)
\end{aligned}$$

Thus, the descent analysis reduces to lower bounding the first-order gradient signal $\langle \overline{G}_t, T_t\rangle_F$ and upper bounding the refined update norm $\|T_t\|_F$.

We first lower bound $\langle \overline{G}_t, T_t\rangle_F$. Using the decomposition $\overline{G}_t = X_t + (\overline{G}_t - X_t)$,

$$\begin{aligned}
\langle \overline{G}_t, T_t\rangle_F &= \langle X_t, T_t\rangle_F + \langle \overline{G}_t - X_t, T_t\rangle_F \ge (1 - \nu_t \gamma' \chi_t)\, \langle X_t, T_t\rangle_F \\
&\ge \epsilon \langle X_t, T_t\rangle_F \ge \epsilon \lambda_{\min}(P_t) \langle X_t, \Psi_t\rangle_F \\
&\ge \epsilon \gamma \lambda_{\min}(P_t) \|\overline{G}_t\|_F \ge \epsilon \gamma \max\{m, n\}^{-1/2} \|\overline{G}_t\|_F.
\end{aligned}$$

We now explain each step in this bound. The first inequality follows from Lemma 3, which gives

$$\left|\langle \overline{G}_t - X_t, T_t\rangle_F\right| \le \nu_t \gamma' \chi_t \langle X_t, T_t\rangle_F.$$

The second inequality follows from the assumption that there exists $\epsilon > 0$ such that

$$\max_{0 \le t \le T} \nu_t \gamma' \chi_t \le 1 - \epsilon.$$

This condition ensures that, in the worst case, the radial interaction can cancel at most a $(1-\epsilon)$ fraction of the positive first-order term $\langle X_t, T_t\rangle_F$, leaving at least an ϵ fraction in the full gradient-update inner product. Since Lemma 3 bounds the sign-indefinite term $\langle \overline{G}_t - X_t, T_t\rangle_F$ only in magnitude, this provides a sufficient but not necessary condition for convergence.

The third inequality follows from Lemma 2(a), which gives $\langle X_t, T_t\rangle_F \ge \lambda_{\min}(P_t) \langle X_t, \Psi_t\rangle_F$. The fourth inequality follows from Lemma 1, which gives $\langle X_t, \Psi_t\rangle_F \ge \gamma \|\overline{G}_t\|_F$. Finally, Lemma 2(c) gives $\lambda_{\min}(P_t) \ge \max\{m, n\}^{-1/2}$.

We next control the quadratic term in Equation 16. Lemma 2(b) gives $\|T_t\|_F^2 \le \max\{m,n\}$. Substituting the preceding lower and upper bounds into Equation 16 and using $\|\overline{G}_t\|_F = \|\nabla\mathcal{L}(W_t)\|_F$, gives

$$\mathcal{L}(W_{t+1}) \le \mathcal{L}(W_t) - \eta\epsilon\gamma \max\{m,n\}^{-1/2}\|\nabla\mathcal{L}(W_t)\|_F + \frac{L\eta^2}{2}\max\{m,n\}. \tag{17}$$

Summing Equation 17 from $t=0$ to T gives

$$\eta\epsilon\gamma \max\{m,n\}^{-1/2}\sum_{t=0}^{T}\|\nabla\mathcal{L}(W_t)\|_F \le \mathcal{L}(W_0) - \mathcal{L}(W_{T+1}) + \frac{(T+1)L\eta^2}{2}\max\{m,n\}.$$

Since $\mathcal{L}(W_{T+1}) \ge \mathcal{L}_{\inf}$ and $\Delta_0 := \mathcal{L}(W_0) - \mathcal{L}_{\inf}$,

$$\eta\epsilon\gamma \max\{m,n\}^{-1/2}\sum_{t=0}^{T}\|\nabla\mathcal{L}(W_t)\|_F \le \Delta_0 + \frac{(T+1)L\eta^2}{2}\max\{m,n\}.$$

Dividing by $\eta\epsilon\gamma \max\{m,n\}^{-1/2}(T+1)$ and using that the minimum does not exceed the average yields

$$\min_{0\le t\le T}\|\nabla\mathcal{L}(W_t)\|_F \le \frac{\Delta_0 \max\{m,n\}^{1/2}}{\eta\epsilon\gamma(T+1)} + \frac{L\eta\max\{m,n\}^{3/2}}{2\epsilon\gamma}.$$

For any constant $C>0$, choosing $\eta = C/\sqrt{T+1}$ gives

$$\min_{0\le t\le T}\|\nabla\mathcal{L}(W_t)\|_F \le \frac{1}{\sqrt{T+1}}\left(\frac{\Delta_0 \max\{m,n\}^{1/2}}{\epsilon\gamma C} + \frac{LC\max\{m,n\}^{3/2}}{2\epsilon\gamma}\right).$$

For $\Delta_0 > 0$, the bound is minimized by

$$C^\star = \sqrt{\frac{2\Delta_0}{L\max\{m,n\}}}, \qquad \eta^\star = \frac{C^\star}{\sqrt{T+1}} = \sqrt{\frac{2\Delta_0}{L\max\{m,n\}(T+1)}}.$$

Substituting $C^\star$ gives

$$\min_{0\le t\le T}\|\nabla\mathcal{L}(W_t)\|_F \le \frac{\max\{m,n\}\sqrt{2L\Delta_0}}{\epsilon\gamma\sqrt{T+1}}.$$

Thus, both NormPre-G with matrix-sign transformation and NormPre-L with Exact implementation achieve an $\mathcal{O}(T^{-1/2})$ convergence rate under their corresponding conditions. □

D.2.3 Convergence of NormPre-L (Sketch)

We next extend the preceding analysis to NormPre-L with the Sketch-based implementation by viewing the sketched localized update as a perturbation of the exact update constructed from the same normalized update Ψ_t.

Corollary 1 (Convergence of NormPre-L (Sketch) w/o Momentum)**.** *Suppose that the assumptions of Theorem 1 hold for the sequence generated by* NormPre-L *with the Sketch-based implementation. Let P_t and $\widehat{P}_t$ denote the exact and sketch-based localized spectral preconditioners constructed from the same Ψ_t, respectively. Assume that there exists $\delta \ge 0$ such that $\|\widehat{P}_t - P_t\|_{\mathrm{op}} \le \delta$ and $\delta \max_{0\le t\le T}\lambda_{t,1} < \epsilon\gamma$. For any constant $C>0$, choosing $\eta = C/\sqrt{T+1}$ gives*

$$\min_{0\le t\le T}\|\nabla\mathcal{L}(W_t)\|_F \le \frac{1}{\sqrt{T+1}}\left(\frac{\Delta_0 \max\{m,n\}^{1/2}}{(\epsilon\gamma - \delta\max_{0\le t\le T}\lambda_{t,1})\,C} + \frac{LC\max\{m,n\}^{3/2}}{2\,(\epsilon\gamma - \delta\max_{0\le t\le T}\lambda_{t,1})}\right).$$

Proof. Let P_t and $\widehat{P}_t$ denote the localized spectral preconditioners constructed from the exact and sketch-based interaction eigenspaces, respectively (Section 3.2):

$$P_t := I + \widetilde{U}_{\mathcal{C}_t}\left(\Lambda_{\mathcal{C}_t}^{-1/2} - I\right)\widetilde{U}_{\mathcal{C}_t}^\top, \quad \widehat{P}_t := I + \widehat{U}_{\mathcal{C}_t}\left(\widehat{\Lambda}_{\mathcal{C}_t}^{-1/2} - I\right)\widehat{U}_{\mathcal{C}_t}^\top.$$

For notational simplicity, $\mathcal{C}_t$ denotes the selected index set in each implementation; the exact and sketch-based sets may differ in both membership and cardinality. The corresponding updates are

$T_t := P_t\Psi_t$ and $\widehat{T}_t := \widehat{P}_t\Psi_t$. Both are evaluated from the same Ψ_t and differ only in the interaction eigenspace extraction.

We first bound the resulting update perturbation. Using the assumption $\|\widehat{P}_t - P_t\|_{\mathrm{op}} \le \delta$, we have

$$\begin{aligned}\|\widehat{T}_t - T_t\|_{\mathrm{op}} &= \|(\widehat{P}_t - P_t)\Psi_t\|_{\mathrm{op}} \\ &\le \|\widehat{P}_t - P_t\|_{\mathrm{op}}\|\Psi_t\|_{\mathrm{op}} \\ &\le \delta\sqrt{\lambda_{t,1}},\end{aligned}$$

where $\|\Psi_t\|_{\mathrm{op}} = \sqrt{\lambda_{t,1}}$ follows from $\Gamma_t = \Psi_t\Psi_t^\top$.

We next lower bound the first-order gradient signal. By nuclear/operator norm duality,

$$\langle \overline{G}_t, \widehat{T}_t\rangle_F \ge \langle \overline{G}_t, T_t\rangle_F - \|\overline{G}_t\|_*\|\widehat{T}_t - T_t\|_{\mathrm{op}}. \tag{18}$$

As in Lemma 1, $\|x_{t,i}\|_2 \ge \|\bar{g}_{t,i}\|_2 \sin(\phi_{t,i})$ with $\sin(\phi_{t,i}) \ge \gamma$. Then, $\|\overline{G}_t\|_* \le \sum_i \|\bar{g}_{t,i}\|_2 \le \gamma^{-1}\sum_i \|x_{t,i}\|_2 = \gamma^{-1}\langle X_t, \Psi_t\rangle_F$. Therefore,

$$\|\overline{G}_t\|_*\|\widehat{T}_t - T_t\|_{\mathrm{op}} \le \frac{\delta\sqrt{\lambda_{t,1}}}{\gamma}\langle X_t, \Psi_t\rangle_F. \tag{19}$$

As derived in Appendix D.2.2 for NormPre-L (Exact),

$$\langle \overline{G}_t, T_t\rangle_F \ge \epsilon\lambda_{\min}(P_t)\langle X_t, \Psi_t\rangle_F. \tag{20}$$

Combining Equations 18–20 with $\lambda_{\min}(P_t) \ge \lambda_{t,1}^{-1/2}$ from Lemma 2, we obtain

$$\begin{aligned}\langle \overline{G}_t, \widehat{T}_t\rangle_F &\ge \left(\epsilon\lambda_{t,1}^{-1/2} - \frac{\delta\sqrt{\lambda_{t,1}}}{\gamma}\right)\langle X_t, \Psi_t\rangle_F \\ &= \frac{\epsilon\gamma - \delta\lambda_{t,1}}{\gamma\sqrt{\lambda_{t,1}}}\langle X_t, \Psi_t\rangle_F \\ &\ge \frac{\epsilon\gamma - \delta\lambda_{t,1}}{\sqrt{\lambda_{t,1}}}\|\overline{G}_t\|_F \\ &\ge \left(\epsilon\gamma - \delta\max_{0\le t\le T}\lambda_{t,1}\right)\max\{m,n\}^{-1/2}\|\overline{G}_t\|_F.\end{aligned}$$

Thus, the assumption $\delta\max_{0\le t\le T}\lambda_{t,1} < \epsilon\gamma$ preserves a positive first-order gradient signal.

It remains to bound the sketched update norm. From $\widehat{P}_t = I + \widehat{U}_{\mathcal{C}_t}\left(\widehat{\Lambda}_{\mathcal{C}_t}^{-1/2} - I\right)\widehat{U}_{\mathcal{C}_t}^\top$, since the columns of $\widehat{U}_{\mathcal{C}_t}$ are orthonormal, $\widehat{P}_t$ has eigenvalues $\widehat{\lambda}_{t,i}^{-1/2}$ along the selected directions and 1 on the non-selected ones. Since the selected modes satisfy $\widehat{\lambda}_{t,i} > 1$, all eigenvalues of $\widehat{P}_t$ lie in $(0, 1]$. Therefore, $\|\widehat{T}_t\|_F \le \|\widehat{P}_t\|_{\mathrm{op}}\|\Psi_t\|_F \le \|\Psi_t\|_F$, which gives $\|\widehat{T}_t\|_F^2 \le \max\{m,n\}$.

Substituting the preceding first-order and norm bounds into the same L-smoothness derivation as in the proof of Appendix D.2.2 gives

$$\mathcal{L}(W_{t+1}) \le \mathcal{L}(W_t) - \eta\left(\epsilon\gamma - \delta\max_{0\le t\le T}\lambda_{t,1}\right)\max\{m,n\}^{-1/2}\|\nabla\mathcal{L}(W_t)\|_F + \frac{L\eta^2}{2}\max\{m,n\}.$$

Summing from $t = 0$ to T, using $\mathcal{L}(W_{T+1}) \ge \mathcal{L}_{\inf}$ and $\Delta_0 = \mathcal{L}(W_0) - \mathcal{L}_{\inf}$ and choosing $\eta = C/\sqrt{T+1}$ as in Theorem 1 yields

$$\min_{0\le t\le T}\|\nabla\mathcal{L}(W_t)\|_F \le \frac{1}{\sqrt{T+1}}\left(\frac{\Delta_0\max\{m,n\}^{1/2}}{(\epsilon\gamma - \delta\max_{0\le t\le T}\lambda_{t,1})\,C} + \frac{LC\max\{m,n\}^{3/2}}{2\,(\epsilon\gamma - \delta\max_{0\le t\le T}\lambda_{t,1})}\right).$$

Therefore, under the condition $\delta\max_{0\le t\le T}\lambda_{t,1} < \epsilon\gamma$ with a positive margin independent of T, NormPre-L with the Sketch-based implementation preserves the $\mathcal{O}(T^{-1/2})$ convergence rate established for the Exact implementation. □