PRO-KLSHAMPOO: PROJECTED KL-SHAMPOO WITH WHITENING RECOVERED BY ORTHOGONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimizers that exploit the matrix structure of gradients are central to modern LLM pre-training, with two distinct frontiers: explicit Kronecker-factored preconditioning—most recently KL-Shampoo, which estimates the preconditioner via KL divergence minimization—and orthogonalization of the gradient momentum, exemplified by Muon and analyzed as steepest descent under the spectral norm. The two routes are typically developed in isolation. We make a structural observation about KL-Shampoo's Kronecker preconditioners: their eigenvalue spectra exhibit a *spike-and-flat* shape—a few dominant eigenvalues followed by an approximately uniform tail—across layers and training stages, holding exactly under a rank-ρ signal-plus-noise gradient model. We exploit this structure by restricting one of KL-Shampoo's Kronecker factors to a parametric family aligned with the spike-and-flat shape: full spectral structure on a tracked r-dimensional subspace, single shared eigenvalue across the remaining $n - r$ directions. On these directions, we apply orthogonalization. An identity shows that this orthogonalization recovers the algebraic form of full KL-Shampoo's preconditioner, with the running estimate replaced by an instantaneous one. On four pre-training scales (GPT-2 124M / 350M, LLaMA 134M / 450M), Pro-KLShampoo consistently outperforms KL-Shampoo at every subspace rank we test in validation loss, peak per-GPU memory, and wallclock time to reach KL-Shampoo's final validation loss.

1 INTRODUCTION

Optimizers that exploit the matrix structure of gradients now rival AdamW (Kingma & Ba, 2014; Loshchilov & Hutter, 2017) for LLM pre-training. The Kronecker-factored preconditioning approach, originating with Shampoo (Gupta et al., 2018), has seen rapid recent development. SOAP (Vyas et al., 2024) reduces per-iteration runtime by running Adam in Shampoo's eigenbasis. KL-Shampoo (Lin et al., 2025) instead replaces the underlying estimation principle—recasting it as KL divergence minimization—and outperforms SOAP and Shampoo on LLM pretraining tasks. A separate approach, exemplified by Muon (Jordan et al., 2024), orthogonalizes the gradient momentum, realizing steepest descent under the spectral norm (Bernstein & Newhouse, 2024b).

In this work, we make a structural observation about KL-Shampoo's Kronecker preconditioners: their eigenvalue spectra exhibit a *spike-and-flat* shape—a few dominant eigenvalues followed by an approximately uniform tail across layers and training stages (Figure 1), echoing low-rank gradient and spiked Hessian phenomena reported in neural network training (Gur-Ari et al., 2018; Ghorbani et al., 2019; Jaiswal et al., 2024). Under a rank-ρ signal-plus-noise gradient model, the derived preconditioners in KL-Shampoo exhibit the spike-and-flat structure exactly: at least $n - \rho$ bottom eigenvalues coincide. This motivates restricting one of the Kronecker preconditioners to a parametric family aligned with the spike-and-flat shape.

Concretely, the parametric family takes the following form: full spectral structure on a tracked r-dimensional subspace, and a single shared eigenvalue across the remaining $n - r$ directions. This scalar tail update is the KL-optimal choice within the restricted family when spike-and-flat holds exactly. In practice the tail is only approximately uniform, and the scalar tail update loses information. Pro-KLShampoo retains the spike-and-flat parametric family but *replaces the scalar tail update with*

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.

orthogonalization. An identity (Eq. (12)) motivates this replacement: on the remaining directions, orthogonalization recovers the algebraic form of full KL-Shampoo's complement update for the current minibatch. Our contributions:

1. We identify a spike-and-flat eigenvalue structure in KL-Shampoo's preconditioners—empirically robust across layers and training stages, theoretically exact under a low-rank gradient model—and exploit it by restricting the KL objective to a parametric family aligned with this structure.
2. On the directions outside the tracked subspace, we show that orthogonalization recovers the algebraic form of full KL-Shampoo's complement update for the current minibatch, connecting Muon-style orthogonalization to the KL estimation framework. A stationarity bound up to a noise-dependent neighborhood follows for the idealized update under operator-norm smoothness.
3. We evaluate Pro-KLShampoo on GPT-2 (124M / 350M) and LLaMA (134M / 450M) at multiple subspace ranks. Pro-KLShampoo improves on KL-Shampoo in validation loss, peak memory, and wallclock time to reach matched final loss in every configuration; an ablation study examines the contribution of each design component.

2 Background

We use $\mathbb{S}^k_{++}$ for the cone of $k \times k$ symmetric positive-definite (SPD) matrices, and $\mathrm{St}(n, r) \coloneqq \{U \in \mathbb{R}^{n\times r} : U^\top U = I_r\}$ for the Stiefel manifold. We write $\|\cdot\|_F$, $\|\cdot\|_{\mathrm{op}}$, and $\|\cdot\|_*$ for the Frobenius, operator (spectral), and nuclear norms, respectively.

Consider a weight matrix $W \in \mathbb{R}^{m\times n}$ and stochastic gradient $G \in \mathbb{R}^{m\times n}$. Following Gupta et al. (2018), we use row-major vectorization: $\mathrm{vec}(G) \coloneqq (g_1^\top, \ldots, g_m^\top)^\top \in \mathbb{R}^{mn}$ where $g_i^\top$ is the i-th row of G. We write $\Sigma \coloneqq \mathbb{E}[\mathrm{vec}(G)\mathrm{vec}(G)^\top] \in \mathbb{S}^{mn}_{++}$ for the gradient second moment.

2.1 Kronecker-factored preconditioning: K-FAC, Shampoo, and KL-Shampoo

Kronecker product approximation of the preconditioner has seen substantial development in neural-network training. **K-FAC** (Martens & Grosse, 2015) approximates the Fisher information matrix as a Kronecker product of two factors, enabling efficient natural-gradient updates. **Shampoo** (Gupta et al., 2018) adopts the same Kronecker structure for the gradient second moment, applying the update $\Delta W \propto L^{-p} G R^{-p}$ where $L \in \mathbb{S}^m_{++}, R \in \mathbb{S}^n_{++}$ are estimated as $L \propto \mathbb{E}[GG^\top]$ and $R \propto \mathbb{E}[G^\top G]$. The original Shampoo (Gupta et al., 2018) uses $p = 1/4$ for the regret guarantee; subsequent theoretical and empirical works (Morwani et al., 2024; Shi et al., 2023) support $p = 1/2$ and we follow this convention throughout.

KL-Shampoo (Lin et al., 2025) reformulates Shampoo's estimation as a KL divergence minimization. Treating Σ as the covariance of a zero-mean Gaussian, the goal is to find the best Kronecker-structured covariance $L \otimes R$ that approximates Σ:

$$\min_{L\in\mathbb{S}^m_{++},\, R\in\mathbb{S}^n_{++}} D_{\mathrm{KL}}\big(\mathcal{N}(0,\Sigma) \,\|\, \mathcal{N}(0, L\otimes R)\big). \tag{1}$$

Unlike Shampoo's Frobenius-based estimation, the KL divergence naturally respects the SPD constraint on L and R. The stationarity conditions of (1) (Lin et al. 2025, Claim 2) couple L, R:

$$L^* = \tfrac{1}{n}\,\mathbb{E}\big[G\,(R^*)^{-1}G^\top\big], \qquad R^* = \tfrac{1}{m}\,\mathbb{E}\big[G^\top\,(L^*)^{-1}\,G\big]. \tag{2}$$

Each side is the gradient second moment whitened by the other side—a coupling absent in Shampoo. This KL-based estimation removes the need for step-size grafting, which vanilla Shampoo typically requires. In practice, L and R are maintained via exponential moving averages (EMA).

2.2 Orthogonalization

For a matrix $M = U\Sigma V^\top$ (SVD), define $\mathrm{polar}(M) \coloneqq UV^\top = M(M^\top M)^{\dagger/2}$; this operation is called *orthogonalization*. The gradient step $W \leftarrow W - \eta\,\mathrm{polar}(\nabla f(W))$ realizes steepest descent under the spectral norm (Bernstein & Newhouse, 2024b;a; Large et al., 2024). Muon (Jordan et al., 2024) applies this update to the momentum, with polar approximated by Newton–Schulz iteration.

3 Pro-KLShampoo: Exploiting Spike-and-Flat Structure

In this section, we present **Pro-KLShampoo: Projected KL-Shampoo with Whitening Recovered by Orthogonalization**, by restricting KL-Shampoo's preconditioner to a parametric family with a

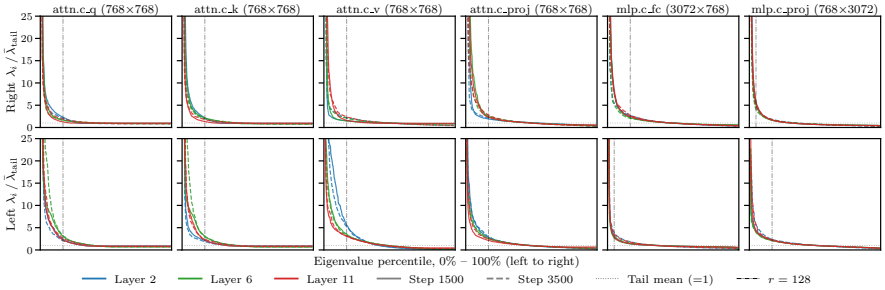

Figure 1: Eigenvalue spectra of practical KL-Shampoo's Kronecker preconditioners on GPT-2 (124M), normalized by the tail mean; the dash-dot line marks $r = 128$. The right-side R (top row) and left-side L (bottom row) show a consistent spike-and-flat pattern across all layer types (panels), three depths (color), and two training stages (line style); the first four panels are attention (square) layers, the last two MLP (rectangular) layers. For square layers the structure is most pronounced on the right side, most clearly in c_v; for rectangular layers $r = 128$ still captures the spike on the larger side, though the tail occupies a larger fraction of the spectrum and varies more.

spike-and-flat eigenvalue structure. The motivation comes from two parts: an empirical observation and an argument from a low-rank gradient model.

Empirical observation. Figure 1 visualizes the eigenvalue spectra of KL-Shampoo's Kronecker preconditioners. Both sides exhibit a *spike-and-flat* pattern: a small number of dominant eigenvalues followed by a nearly uniform tail. This is a robust empirical pattern across the layers, depths, and training stages we measure. The same structure also holds on LLaMA (Appendix D). Appendix E quantifies the pattern by the tail ratio ρ_r, the ratio of the arithmetic to the geometric mean of the eigenvalues beyond the leading r. At $r = 128$ the attention tails are within 2.5% of flat at both scales and training stages, and the least flat case, GPT-2 124M MLP, has $\rho_{128} \approx 1.19$. Related spectral observations—low-rank gradient subspaces (Jaiswal et al., 2024; Gur-Ari et al., 2018) and spiked Hessian spectra (Ghorbani et al., 2019)—suggest that low-dimensional spectral structure is pervasive in deep network optimization.

Low rank gradient supports the spike-and-flat. The rank-ρ signal-plus-noise gradient model is a fundamental building block of spiked covariance theory (Johnstone, 2001; Paul, 2007) and is consistent with low-rank gradient phenomena reported in transformer training (Zhao et al., 2024; Jaiswal et al., 2024). Under this model, the KL stationary points L^* and R^* have at least $m - \rho$ and $n - \rho$ equal bottom eigenvalues, respectively (Lemma 1; Appendix I).

Structural restriction. Together, these motivate restricting KL-Shampoo's Kronecker preconditioner to a parametric family that encodes the spike-and-flat structure explicitly: the preconditioner need not be maintained in full. We apply this restriction to the right-side preconditioner R only, orienting rectangular layers so that R is the larger side ($m \leq n$).[1] We choose an r-dimensional projected subspace (spanned by a particular $U \in \mathrm{St}(n, r)$) in which R retains its full spectral structure, and collapse the remaining $n - r$ dimensions to a single scalar:

$$\hat{R} \;=\; USU^\top + \mu_\perp P_\perp. \tag{3}$$

Here $S \in \mathbb{S}^r_{++}$ is the preconditioner maintained in the tracked subspace U, and $\mu_\perp > 0$ is the scalar approximating the uniform tail of the spectrum on the complement, with $P_\perp := I_n - UU^\top$. Ideally, r captures the dominant eigenvalues of R (Figure 1). When $r \ll n$, the right-side storage drops from $O(n^2)$ to $O(nr)$ and its QR cost from $O(n^3)$ to $O(r^3)$; the left-side is unchanged.

$$\mathcal{J}^{\text{restr}} \;:=\; \min_{\substack{L \in \mathbb{S}^m_{++},\, U \in \mathrm{St}(n,r) \\ S \in \mathbb{S}^r_{++},\, \mu_\perp > 0}} D_{\mathrm{KL}}\big(\mathcal{N}(0,\Sigma) \,\|\, \mathcal{N}\left(0, L \otimes (USU^\top + \mu_\perp P_\perp)\right)\big) \tag{4}$$

[1] For square layers, Figure 1 shows the spike on R to be more pronounced than on L; for rectangular layers, the larger side has a much more expensive QR decomposition (e.g., GPT-2's MLP has a larger side $4\times$ the smaller, making its QR $64\times$ as expensive; LLaMA's SwiGLU MLP has ratio about $8/3$).

Restricting $\hat{R}$ to spike-and-flat structure on the top-r eigenspace of the full optimum R^* discards information whenever its bottom $n-r$ eigenvalues are not exactly uniform. Intuitively, the flatter the tail, the more accurate the approximation. The following claim quantifies the intuition by the AM/GM ratio (the ratio of arithmetic to geometric mean) of the tail eigenvalues.

Claim 1 (Approximation gap)**.** *Denote the optimal full KL objective in* (1) *as* $\mathcal{J}^{\mathrm{full}}$*. Let* (L^*, R^*) *be the full KL optimum with* R^* *having eigenvalues* $\mu_1^* \geq \cdots \geq \mu_n^* > 0$*. Then:*

$$0 \;\leq\; \mathcal{J}^{\mathrm{restr}} - \mathcal{J}^{\mathrm{full}} \;\leq\; \frac{m(n-r)}{2}\,\log \frac{\mathrm{AM}(\mu_{r+1}^*, \ldots, \mu_n^*)}{\mathrm{GM}(\mu_{r+1}^*, \ldots, \mu_n^*)}. \tag{5}$$

When $\mu_{r+1}^* = \cdots = \mu_n^*$*, the gap is zero.*

The AM/GM ratio is commonly used in majorization theory to represent the spread of a positive sequence (Marshall et al., 1979).

Decomposition of the preconditioned gradient. Since $USU^\top$ and $\mu_\perp P_\perp$ have orthogonal ranges, $\hat{R}^{-1/2}$ (using $p = 1/2$ for the preconditioner exponent) decomposes as $US^{-1/2}U^\top + \mu_\perp^{-1/2} P_\perp$, and the preconditioned gradient splits into

$$L^{-1/2}\, G\, \hat{R}^{-1/2} = \underbrace{L^{-1/2}\, GU\, S^{-1/2}\, U^\top}_{\text{full KL-Shampoo in the subspace}} \quad + \quad \underbrace{\mu_\perp^{-1/2}\, L^{-1/2}\, GP_\perp}_{\substack{\text{one-sided KL-Shampoo in the complement} \\ \text{with the other side approximated by a scalar tail}}} \tag{6}$$

We refer to this preconditioned gradient (and the resulting algorithm) as **Smok-Hop**, the letters of "Pro-KLShampoo" minus those of "polar": it is our method without the orthogonalization introduced in §3.3. On the projected subspace, preconditioning is two-sided (as in KL-Shampoo), while the complement is left-preconditioned and uniformly scaled by a scalar on the right side. The scalar tail discards information when the bottom eigenvalues are not exactly uniform; relatedly, two-sided preconditioning is known to outperform one-sided variants more generally (Eschenhagen et al., 2026). We improve the complement of the preconditioned gradient by introducing orthogonalization (in §3.3). The split uses the running estimate only where the spectrum has structure to capture (the spike); on the flat tail it replaces the running per-direction estimate by whitening evaluated on the current minibatch, which by (12) is what the polar step computes. The overall Pro-KLShampoo update reads:

$$\Delta W = -\alpha_{\mathrm{kl}} L^{-1/2}\, GU\, S^{-1/2}\, U^\top - c_a \mathrm{polar}\big(\mu_\perp^{-1/2}\, L^{-1/2}\, GP_\perp\big), \tag{7}$$

where $c_a = \sqrt{\max(1, m/n)}$ is the aspect-ratio scaling of the official Muon implementation (Jordan et al., 2024) (set by the original shape) for the orthogonalized complement, and α_{kl} is added since $\mathrm{polar}(\cdot)$ changes the scale of the complement. We characterize the magnitude of α_{kl} in section 3.3.

3.1 Stationarity of the restricted problem

We fix U for now (the choice of U is addressed in §3.2) and derive the optimal $L, S, \mu_\perp$. Note that once U is given, the gradient decomposes into the projected subspace and the complement component, i.e., $G = \widetilde{G}U^\top + G_\perp$ where $\widetilde{G} \coloneqq GU \in \mathbb{R}^{m\times r}$ and $G_\perp \coloneqq GP_\perp \in \mathbb{R}^{m\times n}$.

Claim 2 (Restricted stationarity)**.** *The optimal solution of the restricted KL objective* (4) *with respect to* $(L, S, \mu_\perp)$ *for fixed* U *should satisfy:*

$$S^* = \tfrac{1}{m}\,\mathbb{E}\big[\widetilde{G}^\top (L_{\mathrm{restr}}^*)^{-1} \widetilde{G}\big], \quad \textit{(Covariance estimation of the subspace preconditioner)} \tag{8}$$

$$\mu_\perp^* = \tfrac{1}{m(n-r)}\,\mathrm{Tr}\big(\mathbb{E}\big[G_\perp^\top (L_{\mathrm{restr}}^*)^{-1} G_\perp\big]\big), \quad \textit{(Scalar estimation of the complement's flat tail)} \tag{9}$$

$$L_{\mathrm{restr}}^* = \tfrac{1}{n}\,\mathbb{E}\big[G\,(\hat{R}^*)^{-1} G^\top\big], \quad \textit{(Covariance estimation of the unrestricted-side preconditioner)} \tag{10}$$

where $\hat{R}^* = US^*U^\top + \mu_\perp^* P_\perp$*. The proof is in Appendix G.*

The optimal solutions are coupled (as in KL-Shampoo): L_{restr}^* and $\hat{R}^*$ are gradient second moments mutually whitened by each other's inverse, with $\hat{R}^*$ decomposing into S^* on the projected subspace and the scalar $\mu_\perp^*$ averaging the whitened second moment on the complement. Explicit characterization is infeasible—L_{restr}^* is needed to solve for S^* but itself unknown—so we use the EMA scheme of Lin et al. (2025), interpreted as a stochastic proximal gradient step for the KL objective.

3.2 Optimal subspace

Claim 2 gives the optimal $L, S, \mu_\perp$ for any fixed U. Plugging this back, we characterize the minimizer of the resulting objective in U, which depends on the matrix $\Phi_L \coloneqq \mathbb{E}[G^\top L^{-1} G] \in \mathbb{S}^n_{++}$; recall that under full KL-Shampoo stationarity (2), $\Phi_{L^*} = mR^*$.

Claim 3 (Optimal subspace)**.** *Fix $L \in \mathbb{S}^m_{++}$ and let Φ_L have eigenvalues $\phi_1 \geq \cdots \geq \phi_n > 0$. Every global minimizer U^* of (4) over $U \in \mathrm{St}(n, r)$ is an eigenspace of Φ_L. The optimal index set I^* is the size-r subset whose complement has the smallest AM/GM: $I^* \in \arg\min_{|I|=r} \frac{\mathrm{AM}(\phi_j : j \notin I)}{\mathrm{GM}(\phi_j : j \notin I)}$. When the bottom-$(n-r)$ eigenvalues form the flattest subset (spike-and-flat shape), $I^* = \{1, \ldots, r\}$.*

See the proof in Appendix H. In general, the optimal subspace need not be the top-r eigenspace. For example, $\Phi_L = \mathrm{Diag}(10, 9, 1)$ with $r = 1$: the top-1 choice gives complement $\mathrm{AM/GM} \approx 1.67$, while capturing the *bottom* eigenvalue gives $\mathrm{AM/GM} \approx 1.00$. Note that the stationary condition of full KL-Shampoo (2) gives $R^* = \frac{1}{m}\Phi_{L^*}$, so R^* and Φ_{L^*} share the same spike-and-flat structure (Figure 1); since we choose r to capture the dominant eigenvalues of R^*, it also captures those of Φ_{L^*}. We thus conjecture that $\Phi_{L^*_{\mathrm{restr}}}$ has similar structure with its dominant eigenvalues captured within the top r; we check this empirically (Figure 6, Appendix C). Under this structure, Claim 3 gives the top-r eigenspace as optimal. Since L^*_{restr} is unknown, in practice we run subspace iteration on Φ_L, where L is the algorithm's running EMA estimate of L^*_{restr}.

Remark. Tracking the top-r eigenspace is also used in prior subspace-based optimizers (Zhao et al., 2024; Liu et al., 2025); under the restricted KL objective and the spike-and-flat conjecture on $\Phi_{L^*_{\mathrm{restr}}}$, this construction emerges as the KL-optimal choice (Claim 3).

3.3 Orthogonalization recovers the algebraic form of whitening

The update decomposition (6) applies the same scalar $\mu_\perp^{-1/2}$ to all $n-r$ directions in the complement subspace, rather than a preconditioner matrix scaling each direction. We call the latter *per-direction whitening*. We argue below that orthogonalization recovers the algebraic form of full KL-Shampoo's complement update for the current minibatch.

Decomposition of full KL-Shampoo's update. Full KL-Shampoo applies the update $\Delta W^{\mathrm{full}} = -L^{*-1/2}\, G\, R^{*-1/2}$. Let U_f be the top-r eigenspace of R^* and $U_{f,\perp}$ be an orthonormal basis for U_f's orthogonal complement. Since R^* is block-diagonal in the basis $[U_f, U_{f,\perp}]$, the full-KL update decomposes as

$$\Delta W^{\mathrm{full}} = -L^{*-1/2}\, G\, U_f\, (U_f^\top R^* U_f)^{-1/2}\, U_f^\top \;-\; \underbrace{L^{*-1/2}\, G\, U_{f,\perp}\, (U_{f,\perp}^\top R^* U_{f,\perp})^{-1/2}\, U_{f,\perp}^\top}_{\Delta_\perp^{\mathrm{full}}}.$$

Substituting the full-KL stationarity condition (2) into $\Delta_\perp^{\mathrm{full}}$:

$$\Delta_\perp^{\mathrm{full}} = \sqrt{m}\, L^{*-1/2}\, G\, U_{f,\perp} \left(U_{f,\perp}^\top \mathbb{E}[\boldsymbol{G^\top L^{*-1} G}]\, U_{f,\perp}\right)^{-1/2} U_{f,\perp}^\top. \tag{11}$$

Pro-KLShampoo's orthogonalization on the complement. Denote $U_\perp$ as an orthonormal basis for U's orthogonal complement, equivalently $P_\perp = U_\perp U_\perp^\top$. The complement update of (7) reads (the aspect-ratio scaling c_a is absorbed into the learning rate):

$$\begin{aligned}\mathrm{polar}(\mu_\perp^{-1/2} L^{-1/2} G\, P_\perp) &= \mathrm{polar}(L^{-1/2} G\, U_\perp)\, U_\perp^\top \\ &= L^{-1/2} G\, U_\perp \left(U_\perp^\top \boldsymbol{G^\top L^{-1} G}\, U_\perp\right)^{\dagger/2} U_\perp^\top. \end{aligned} \tag{12}$$

Identity (12) recovers the algebraic form of full KL-Shampoo's complement update (11) for the current minibatch. The right preconditioner depends explicitly on L^{-1}, with the stationary statistic $\mathbb{E}[G^\top L^{*-1} G]$ replaced by the instantaneous $G^\top L^{-1} G$, which we do not claim approximates it in a statistical sense. A second difference is scale. Since $\mathrm{polar}(\alpha M) = \mathrm{polar}(M)$ for $\alpha > 0$, (12) recovers the per-direction structure but not the magnitude of the complement update, which is set by c_a and the calibrated α_{kl} (Appendix J.3). At the restricted-KL stationary point with $L^*_{\mathrm{restr}} = L^*$, $\Phi_{L^*_{\mathrm{restr}}} = mR^*$ by (2), so $\Phi_{L^*_{\mathrm{restr}}}$ and R^* share the top-r eigenspace spanned by U_f (Claim 3).

Claim 1 bounds the KL approximation gap of the scalar-tail family; the effect of the polar complement is assessed empirically in §4.2, and the convergence analysis (§3.4) treats the polar update directly. Beyond Pro-KLShampoo, identity (12) supports the broader idea of composing Muon-style orthogonalization with Shampoo-style whitening.

Calibrating α_{kl}. The mixing weight α_{kl} is set by matching the operator-norm of the orthogonalized complement to that of the original complement update at the restricted KL stationary state. This yields a per-layer principled range, which we intersect across layers to obtain $\alpha_{\mathrm{kl}} \in [1.4\times10^{-3},\ 1.6\times10^{-2}]$ (Appendix J.3). In experiments we sweep $\alpha_{\mathrm{kl}} \in \{0.005, 0.01, 0.015\}$ within this range.

Algorithm 1 Pro-KLShampoo (idealized): one step $t \to t+1$

State at step t: $(W_t, L_t, S_t, U_t, \mu_{\perp,t})$ (predictable); after sampling G_t, set $P_{\perp,t} = I_n - U_t U_t^\top$.

1: Parameter update

$$D_t = \alpha_{\mathrm{kl}} L_t^{-1/2} G_t U_t S_t^{-1/2} U_t^\top + c_a \,\mathrm{polar}\Big(\mu_{\perp,t}^{-1/2} L_t^{-1/2} G_t P_{\perp,t}\Big), \qquad W_{t+1} = W_t - \eta D_t.$$

2: Statistics update in the U_t coordinates

$$\Delta L_t = \tfrac{1}{n}\Big(G_t U_t S_t^{-1} U_t^\top G_t^\top + \mu_{\perp,t}^{-1} G_t P_{\perp,t} G_t^\top\Big), \qquad \Delta S_t = \tfrac{1}{m} U_t^\top G_t^\top L_t^{-1} G_t U_t,$$

$$\delta_{\perp,t} = \tfrac{1}{m(n-r)} \Big\| L_t^{-1/2} G_t P_{\perp,t} \Big\|_F^2.$$

Set $L_{t+1} = \beta_2 L_t + (1-\beta_2)\Delta L_t, \quad \bar{S}_{t+1} = \beta_2 S_t + (1-\beta_2)\Delta S_t, \quad \mu_{\perp,t+1} = \beta_2 \mu_{\perp,t} + (1-\beta_2)\delta_{\perp,t}.$

3: Subspace update:

$$U_{t+1} = \mathrm{qr}\Big(\beta_2 U_t S_t + \tfrac{1-\beta_2}{m} G_t^\top L_t^{-1} G_t U_t\Big),\ T_t = U_t^\top U_{t+1},\ S_{t+1} = T_t^\top \bar{S}_{t+1} T_t.$$

4: Exact eigendecomposition:

$$(\Lambda_{L,t+1}, Q_{L,t+1}) = \mathrm{eig}(L_{t+1}), \quad (\Lambda_{S,t+1}, Q_{S,t+1}) = \mathrm{eig}(S_{t+1}).$$

The next update uses $L_{t+1}^{-1/2} = Q_{L,t+1} \Lambda_{L,t+1}^{-1/2} Q_{L,t+1}^\top$ and $S_{t+1}^{-1/2} = Q_{S,t+1} \Lambda_{S,t+1}^{-1/2} Q_{S,t+1}^\top$.

In Algorithm 1, the update direction D_t uses G_t and the step-t state, which is refreshed only afterwards. Step 2 forms the statistics in the U_t coordinates, and Step 3 moves $\bar{S}_{t+1}$ to the U_{t+1} coordinates through $T_t := U_t^\top U_{t+1}$. Step 3 tracks the top-r eigenspace of Φ_L via one step of power iteration; under the spike-and-flat conjecture on $\Phi_{L^*_{\mathrm{restr}}}$, this eigenspace is the KL-optimal subspace (Claim 3). We estimate $\frac{1}{m}\Phi_L U_t$ by EMA over a fresh minibatch sample $\frac{1}{m} G_t^\top L_t^{-1} G_t U_t$ and a running estimate $U_t S_t$, exact at the restricted KL optimum because $L = L^*_{\mathrm{restr}}$ (Claim 2) and U is an eigenspace of $\Phi_{L^*_{\mathrm{restr}}}$ (Claim 3) together give $\Phi_{L^*_{\mathrm{restr}}} U = m\,US$. The EMA covariance estimation (Step 2) follows Lin et al. (2025). All EMAs in Algorithm 1 share a single coefficient β_2. Algorithm 2 (Appendix K) replaces the exact eigendecompositions by maintained spectral estimates refreshed by periodic QR.

The practical implementation, which adds Nesterov momentum, Newton–Schulz iteration for orthogonalization, decoupled weight decay, and eigenvalue clipping, is given in Appendix K.

3.4 Convergence analysis

The idealized update of Algorithm 1 admits a stationarity bound up to a noise-dependent neighborhood for the loss f as a function of $W \in \mathbb{R}^{m\times n}$. We assume, as listed in Appendix J, lower boundedness, operator-norm smoothness with constant L_{op} (Bernstein & Newhouse, 2024b; Large et al., 2024), unbiased stochastic gradients with bounded operator norm and conditional variance $\sigma_{F,t}^2$ given the past $\mathcal{F}_t := \sigma(G_0, \ldots, G_{t-1})$, and explicit lower bounds on the eigenvalues of L_t, S_t and on $\mu_{\perp,t}$ (in practice by clipping decoupled eigenvalue estimates as in KL-Shampoo (Lin et al., 2025)).

Theorem 1 (Stationarity bound for the idealized update)**.** *Let Assumptions (i)–(v) hold along the trajectory of Algorithm 1. Let σ_{kl}^2 be a deterministic constant, independent of η, such that $\mathbb{E}\Big[\|L_t^{-1/2} G_t U_t S_t^{-1/2}\|_{\mathrm{op}}^2 \,\Big|\, \mathcal{F}_t\Big] \le \sigma_{kl}^2$ a.s. for all $t \ge 0$ (one exists under Assumptions (iv)–(v); see §J.1.6), and $K := c_a^2 + \alpha_{\mathrm{kl}}^2 \sigma_{kl}^2$. For $t \ge 0$ let*

$$M_t \;:=\; c_a \sqrt{\lambda_{\min}(L_t)}\, \big\| L_t^{-1/2} \nabla f(W_t)\, P_{\perp,t} \big\|_* \;+\; \alpha_{\mathrm{kl}} \big\| L_t^{-1/4} \nabla f(W_t)\, U_t\, S_t^{-1/4} \big\|_F^2. \tag{13}$$

Then Algorithm 1 with step size $\eta = \sqrt{\Delta_0/(T\, L_{\mathrm{op}}\, K)}$ satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[M_t] \;\le\; 2\sqrt{\frac{\Delta_0\, L_{\mathrm{op}}\, K}{T}} + 2\, c_a \sqrt{k}\, \frac{1}{T}\sum_{t=0}^{T-1} \sigma_{F,t}, \tag{14}$$

where $\Delta_0 := f(W_0) - f^*$, c_a as in (7), and $k := \min(m, n-r)$.

M_t is zero if and only if $\nabla f(W_t) = 0$ (Lemma 3, Appendix J.1). The left-hand side mirrors the two update geometries: the orthogonalized complement contributes a *nuclear-norm* term (dual of spectral-norm), and the subspace update contributes a *squared Frobenius* term. The asymmetry is because orthogonalization normalizes the gradient's magnitude away in the complement update, so the descent is linear in ∇f; the subspace update preserves the gradient's magnitude, so ∇f contributes quadratically. The noise term comes from the nonlinearity of orthogonalization and scales with the average noise level. It vanishes when the noise decays, for example with a growing batch, as long as the eigenvalue lower bounds are enforced by clipping, and it stays at $O(B^{-1/2})$ for a fixed batch B, as in signSGD (Bernstein et al., 2018). Analyses of Muon remove this fixed-batch floor with scheduled momentum or variance reduction (Shen et al., 2025; Chang et al., 2025; Sfyraki & Wang, 2025); our analysis is momentum-free and keeps it. The proof and a Frobenius-norm corollary are in Appendix J.1.

4 Experiments

We evaluate Pro-KLShampoo at four model scales and two architectures: GPT-2 (Radford et al., 2019) (124M, 350M) trained on FineWeb-10B (Penedo et al., 2024) and LLaMA (Touvron et al., 2023) (134M, 450M) trained on C4 (Raffel et al., 2020), all on A100 GPUs. Our primary baseline is KL-Shampoo (Lin et al., 2025); we also compare against AdamW and Muon (Jordan et al., 2024), and provide a head-to-head comparison with COSMOS (Liu et al., 2025) in Appendix N. For Muon, KL-Shampoo, and Pro-KLShampoo, the embedding and output weights are trained by AdamW, following Eschenhagen et al. (2026); Lin et al. (2025).

We tune AdamW first, then fix its learning rate and weight decay for the embedding and output layers across all optimizers; the remaining hyperparameters of each method are tuned independently. For Pro-KLShampoo, we additionally sweep $\alpha_{\text{kl}} \in \{0.005, 0.01, 0.015\}$ within the principled range derived in §3.3, and report results at $r \in \{32, 64, 128\}$ (sensitivity to α_{kl}: Table 12). Other hyperparameters (momentum, β_2, preconditioner refresh frequency, Newton–Schulz step count) follow the source-release defaults of Lin et al. (2025); Jordan et al. (2024). KL-Shampoo and Pro-KLShampoo maintain the optimizer state in half precision following Lin et al. (2025); AdamW and Muon use the full-precision optimizer state of their official releases. Reported memory therefore reflects each method's reference configuration; we did not re-validate other methods at half precision.

Table 1: Validation loss (*loss*) and peak per-GPU memory in GiB (*mem*) across pretraining configurations. Best loss across hyperparameter sweeps; memory is from that run.

Optimizer	r	GPT-2 124M		GPT-2 350M		LLaMA 134M		LLaMA 450M	
		loss	mem	loss	mem	loss	mem	loss	mem
AdamW	–	3.4944	30.62	3.2063	74.20	3.0871	32.19	2.8248	73.91
Muon	–	3.2841	30.14	3.0407	72.23	3.0536	31.71	2.7578	71.75
KL-Shampoo	–	3.2796	30.63	3.0397	73.07	3.0504	32.09	2.7650	73.42
Pro-KLShampoo	32	3.2758	30.16	3.0364	72.23	3.0358	31.74	2.7461	71.85
	64	3.2754	30.16	3.0348	72.24	3.0345	31.75	**2.7453**	71.86
	128	**3.2745**	30.17	**3.0344**	72.25	**3.0335**	31.76	2.7469	71.89

4.1 Pretraining experiments

We pretrain two GPT-2 models using the modded-nanogpt framework (Jordan et al., 2024): a 124M model for $5{,}100$ iterations and a 350M model for $13{,}000$ iterations, corresponding to 2.7B and 6.8B tokens respectively. We additionally pretrain two LLaMA models on C4 following the GaLore (Zhao et al., 2024) convention with a linear warmup–linear decay schedule, 10% warmup, and zero weight decay: a 134M model on a 5B subset for $5{,}000$ iterations and a 450M model on a 10B subset for $6{,}250$ iterations. Architecture details, sweep ranges, and selected hyperparameters are in Appendix O.

Main results. Across all configurations, Pro-KLShampoo attains the lowest validation loss at every tested rank, with peak per-GPU memory below KL-Shampoo's and on par with Muon's (Table 1). Pro-KLShampoo also reaches KL-Shampoo's final validation loss in less total wallclock time at every rank (Figure 2); we call this reduction *wallclock saving*. The best wallclock savings are 4.12% (GPT-2 350M), 2.28% (GPT-2 124M), 13.43% (LLaMA 450M), and 10.96% (LLaMA 134M). At GPT-2 350M, the smallest-gain setting, four seeds per method give losses of 3.0352 ± 0.0009 vs. 3.0401 ± 0.0014 with disjoint ranges (exact one-sided rank-sum $p = 0.014$, Table 6).

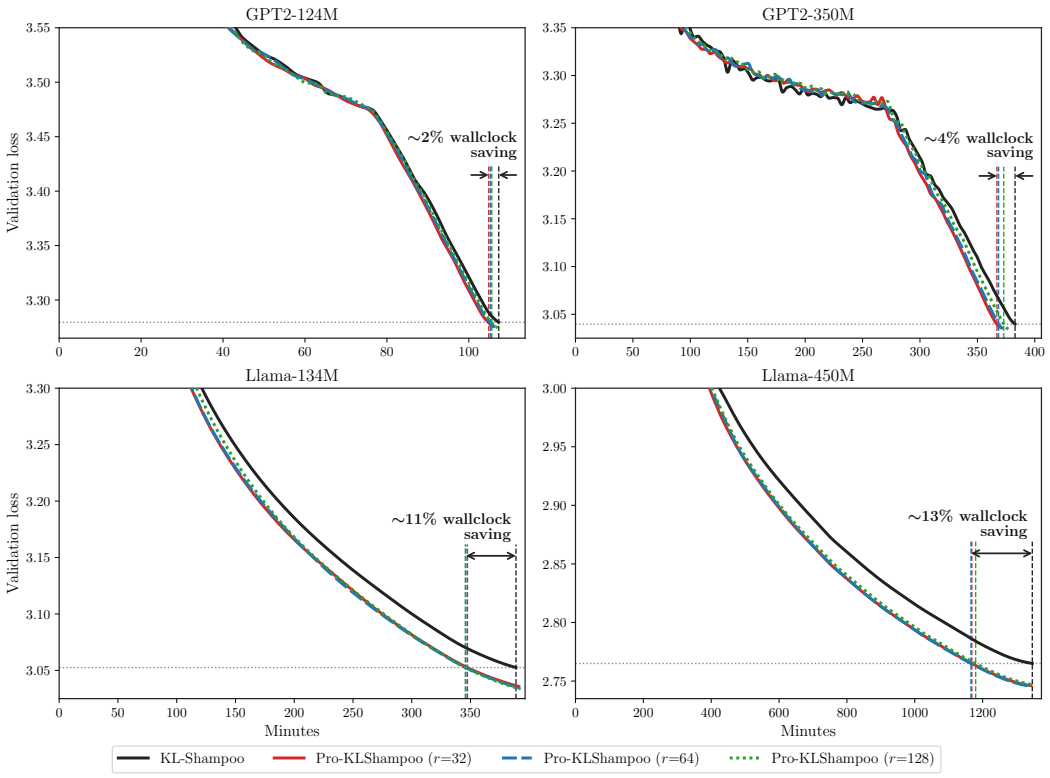

Figure 2: Validation loss versus time for KL-Shampoo and Pro-KLShampoo with $r \in \{32, 64, 128\}$ across all four pretraining configurations. Pro-KLShampoo reaches KL-Shampoo's final validation loss (dotted line) faster at every rank in every panel. On GPT-2, its curves also terminate earlier at every rank; under the fixed iteration count of each panel, this implies shorter per-step time.

Mechanism of the wallclock saving. Wallclock saving has three potential sources: reduced per-step time (each iteration is faster), faster convergence (fewer iterations to reach a given training loss), and a generalization effect (lower validation loss at matched training loss). On GPT-2, per-step time is the primary source: Pro-KLShampoo replaces the QR decomposition on the larger Kronecker factor with one of size r, and GPT-2's largest weight dimension (3072 at 124M, 4096 at 350M) is where this reduction outweighs the overhead of subspace tracking (Appendix L compares per-step cost and memory). Convergence does not contribute: Pro-KLShampoo's training loss is comparable to or slightly above KL-Shampoo's (Figure 4, Appendix B). A small generalization effect may also contribute: on GPT-2, Pro-KLShampoo's validation loss is lower than KL-Shampoo's despite comparable training loss; we leave its mechanism to future work. On LLaMA, the largest weight dimension is smaller (2048 at 134M, 2816 at 450M), so per-step time is comparable between the two methods. The wallclock saving is instead driven by faster convergence: Pro-KLShampoo's training loss lies visibly below KL-Shampoo's throughout training (Figure 5, Appendix B).

At smaller ranks such as r=32, the subspace may not capture all dominant eigenvectors of the preconditioner (Figure 1); these fall into the complement, where a single scalar cannot rescale them individually but orthogonalization compensates by recovering per-direction whitening on the current minibatch, and the consistent improvement at r=32 across all four scales (Table 1) supports this.

Choosing r. The validation loss varies by at most 0.003 over $r \in \{32, 64, 128\}$ at all four scales, and r did not need to grow with model size in our range. We use r=128 by default; Appendix O gives a pilot-based guideline for other model widths.

4.2 Ablation Studies

We ablate two ingredients of Pro-KLShampoo: orthogonalization on the complement subspace (§3.3) and the spike-and-flat decomposition (7). All variants run at r=128. Validation losses are reported in Table 2; loss curves on GPT-2 124M and LLaMA 134M are shown in Figure 3.

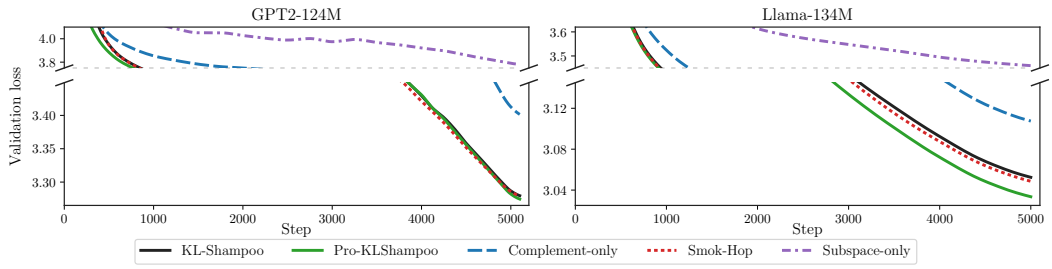

Figure 3: Validation loss versus training step on GPT-2 124M (left) and LLaMA 134M (right) for Pro-KLShampoo, KL-Shampoo, and three ablation variants. Subspace-only and complement-only correspond to the spike-and-flat decomposition ablation; Smok-Hop corresponds to the orthogonalization ablation. All Pro-KLShampoo variants run at $r{=}128$.

Orthogonalization on the complement subspace. We compare Pro-KLShampoo against Smok-Hop, the variant corresponding to the preconditioned gradient (6) without orthogonalization on the complement; following KL-Shampoo, it uses EMA momentum (Appendix K). Table 2 shows Smok-Hop is competitive with KL-Shampoo at the smaller scales but degrades at 450M, where the scalar approximation on the complement appears insufficient. Pro-KLShampoo leads at every scale; the gap is larger at LLaMA-450M (0.024) than at LLaMA-134M (0.014), suggesting orthogonalization is more beneficial at the larger LLaMA scale. Zeroing the complement altogether (a truncated pseudo-inverse variant of KL-Shampoo, Table 15) worsens validation loss on GPT-2 124M by 0.43.

Spike-and-flat decomposition. We further isolate the two update components: *subspace-only* zeros out the complement update, and *complement-only* zeros out the subspace update. Neither alone matches Pro-KLShampoo or KL-Shampoo (Table 2, Figure 3). Complement-only is consistently stronger than subspace-only at both scales. Combining the two gives the full Pro-KLShampoo update (7), which neither alone can express.

Ranks beyond 128. We also evaluate $r{=}256$ and $r{=}512$ on GPT-2 124M and LLaMA-450M (Tables 13 and 14). The gain peaks at $r{=}256$ and falls back to the $r{=}128$ level at $r{=}512$, likely as the tracked subspace takes over more of the update and the method moves toward full KL-Shampoo.

Table 2: Final validation loss at $r{=}128$. *Subspace-only* and *Complement-only* ablate the spike-and-flat decomposition; *Smok-Hop* ablates orthogonalization. Spike-and-flat ablations on 450M were not run.

	Subspace-only	Complement-only	Smok-Hop	Pro-KLShampoo	KL-Shampoo
GPT-124M	3.7803	3.4015	3.2791	**3.2745**	3.2796
LLaMA-134M	3.4598	3.1075	3.0476	**3.0335**	3.0504
LLaMA-450M	—	—	2.7704	**2.7469**	2.7650

5 Conclusion

We proposed Pro-KLShampoo, which restricts KL-Shampoo's preconditioner to a spike-and-flat structure and orthogonalizes the complement. On different pre-training architectures and scales, Pro-KLShampoo improves over KL-Shampoo at every tested rank in validation loss, peak per-GPU memory, and wallclock to matched final loss.

Limitations. The rank r is set empirically (see the guideline in §4.1). Adapting r over time would require rules for transferring the maintained states between ranks; given the observed insensitivity to fixed ranks, we leave this and per-layer ranks to future work. Per-layer or adaptive calibration of α_{kl}, a convergence analysis of the EMA-based subspace tracking and preconditioner estimation, and the extension of the stationarity bound (Theorem 1) to the practical variant of Appendix K, which uses maintained spectral estimates in place of exact eigendecompositions and adds Nesterov momentum, Newton–Schulz orthogonalization, damping, and weight decay, are also left to future work. Our experiments cover batch sizes from 0.5M to 1.7M tokens per step; smaller batches increase the noise term of Theorem 1, and we have not tested them.

AI USE STATEMENT

In this work, we used generative AI tools to draft sections of the paper from the authors' notes, results, and derivations, and to aid and polish the writing (editing for grammar and wording, restructuring and condensing passages). In research execution, we used generative AI tools only to write plotting scripts (including color schemes and multi-panel layouts) and to format result tables; the optimizer and training code was adapted by the authors from existing public repositories (Lin et al., 2025; Jordan et al., 2024), and all AI-assisted scripts were checked by the authors. We have not used generative AI tools for research ideation, for the design of experiments, for the derivation of proofs, for generating datasets, or for producing any numerical result. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The practical algorithm and its implementation details are given in Appendix K. The complete hyperparameter sweeps and selected values for every method and scale, the batch sizes, sequence lengths, and the reporting protocol are listed in Appendix O. The tail-flatness measurement protocol is described in Appendix E. The assumptions and full proofs of every stated result are given in Appendices F–J. The code of the optimizer and a training script for LLaMA-450M are provided as supplementary material.

REFERENCES

Kang An, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, and Tong Zhang. Asgo: Adaptive structured gradient optimization. *arXiv preprint arXiv:2503.20762*, 2025.

Rohan Anil, Vineet Gupta, Tomer Koren, Kevin Regan, and Yoram Singer. Scalable second order optimization for deep learning. *arXiv preprint arXiv:2002.09018*, 2020.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. *arXiv preprint arXiv:2410.21265*, 2024a.

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

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International conference on machine learning*, pp. 560–569. PMLR, 2018.

David E Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned spectral descent for deep learning. *Advances in neural information processing systems*, 28, 2015.

Da Chang, Yongxiang Liu, and Ganzhao Yuan. On the convergence of muon and beyond. *arXiv preprint arXiv:2509.15816*, 2025.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Pierre Dutilleul. The mle algorithm for the matrix normal distribution. *Journal of statistical computation and simulation*, 64(2):105–123, 1999.

Runa Eschenhagen, Aaron Defazio, Tsung-Hsien Lee, Richard E Turner, and Hao-Jun Michael Shi. Purifying shampoo: Investigating shampoo's heuristics by decomposing its preconditioner. *arXiv preprint arXiv:2506.03595*, 2025.

Runa Eschenhagen, Anna Cai, Tsung-Hsien Lee, and Hao-Jun Michael Shi. Clarifying shampoo: Adapting spectral descent to stochasticity and the parameter trajectory. *arXiv preprint arXiv:2602.09314*, 2026.

Thomas George, César Laurent, Xavier Bouthillier, Nicolas Ballas, and Pascal Vincent. Fast approximate natural gradient descent in a kronecker factored eigenbasis. *Advances in neural information processing systems*, 31, 2018.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pp. 2232–2241. PMLR, 2019.

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

Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Yutong He, Pengrui Li, Yipeng Hu, Chuyan Chen, and Kun Yuan. Subspace optimization for large language models with convergence guarantees. *arXiv preprint arXiv:2410.11289*, 2024.

De Huang, Jonathan Niles-Weed, and Rachel Ward. Streaming k-pca: Efficient guarantees for oja's algorithm, beyond rank-one updates. In *Conference on Learning Theory*, pp. 2463–2498. PMLR, 2021.

Zhendong Huang, Hengjie Cao, Fang Dong, Ruijun Huang, Mengyi Chen, Yifeng Yang, Xin Zhang, Anrui Chen, Mingzhi Dong, Yujiang Wang, et al. Spectra: Rethinking optimizers for llms under spectral anisotropy. *arXiv preprint arXiv:2602.11185*, 2026.

Prateek Jain, Chi Jin, Sham M Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming pca: Matching matrix bernstein and near-optimal finite sample guarantees for oja's algorithm. In *Conference on learning theory*, pp. 1147–1164. PMLR, 2016.

Ajay Jaiswal, Yifan Wang, Lu Yin, Shiwei Liu, Runjin Chen, Jiawei Zhao, Ananth Grama, Yuandong Tian, and Zhangyang Wang. From low rank gradient subspace stabilization to low-rank weights: Observations, theories, and applications. *arXiv preprint arXiv:2407.11239*, 2024.

Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327, 2001.

Keller Jordan, Jeremy Bernstein, Brendan Rappazzo, @fernbear.bsky.social, Boza Vlado, You Jiacheng, Franz Cesista, Braden Koszarsky, and @Grad62304977. modded-nanogpt: Speedrunning the nanogpt baseline, 2024. URL `https://github.com/KellerJordan/modded-nanogpt`.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Tim Large, Yang Liu, Minyoung Huh, Hyojin Bahng, Phillip Isola, and Jeremy Bernstein. Scalable optimization in the modular norm. *Advances in Neural Information Processing Systems*, 37:73501–73548, 2024.

Kaizhao Liang, Bo Liu, Lizhang Chen, and Qiang Liu. Memory-efficient llm training with online subspace descent. *Advances in Neural Information Processing Systems*, 37:64412–64432, 2024.

Wu Lin, Scott C Lowe, Felix Dangel, Runa Eschenhagen, Zikun Xu, and Roger B Grosse. Understanding and improving shampoo and soap via kullback-leibler minimization. *arXiv preprint arXiv:2509.03378*, 2025.

Bing Liu, Wenjie Zhou, and Chengcheng Zhao. Rethinking bregman divergences in kronecker-factored optimizers. *arXiv preprint arXiv:2606.00542*, 2026.

Liming Liu, Zhenghao Xu, Zixuan Zhang, Hao Kang, Zichong Li, Chen Liang, Weizhu Chen, and Tuo Zhao. Cosmos: A hybrid adaptive optimizer for memory-efficient training of llms. *arXiv preprint arXiv:2502.17410*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Albert W Marshall, Ingram Olkin, and Barry C Arnold. Inequalities: theory of majorization and its applications. 1979.

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

Alan Milligan, Zikun Xu, Simon Lacoste-Julien, Felix Dangel, and Wu Lin. Reparametrizing shampoo and soap for subspace basis updates and bfloat16 storage. *arXiv preprint arXiv:2605.26327*, 2026.

Ionut-Vlad Modoranu, Philip Zmushko, Erik Schultheis, Mher Safaryan, and Dan Alistarh. Dash: Faster shampoo via batched block preconditioning and efficient inverse-root solvers. *arXiv preprint arXiv:2602.02016*, 2026.

Depen Morwani, Itai Shapira, Nikhil Vyas, Eran Malach, Sham Kakade, and Lucas Janson. A new perspective on shampoo's preconditioner. *arXiv preprint arXiv:2406.17748*, 2024.

Shuntaro Nagashima and Hideaki Iiduka. Improved convergence rates of muon optimizer for nonconvex optimization. *arXiv preprint arXiv:2601.19400*, 2026.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 2006.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3): 267–273, 1982.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pp. 1617–1642, 2007.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

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.

Sahar Rajabi, Nayeema Nonta, and Sirisha Rambhatla. Subtrack++: Gradient subspace tracking for scalable llm training. *arXiv preprint arXiv:2502.01586*, 2025.

Beata Roś, Fetsje Bijma, Jan C de Munck, and Mathisca CM de Gunst. Existence and uniqueness of the maximum likelihood estimator for models with a kronecker product covariance structure. *Journal of Multivariate Analysis*, 143:345–361, 2016.

Maria-Eleni Sfyraki and Jun-Kun Wang. Lions and muons: Optimization via stochastic frank-wolfe. *arXiv preprint arXiv:2506.04192*, 2025.

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

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.

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, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321*, 2024.

Nikhil Vyas, Rosie Zhao, Depen Morwani, Mujin Kwun, and Sham Kakade. Improving soap using iterative whitening and muon, 2025.

Shuo Xie, Tianhao Wang, Sashank Reddi, Sanjiv Kumar, and Zhiyuan Li. Structured preconditioners in adaptive optimization: A unified analysis. *arXiv preprint arXiv:2503.10537*, 2025.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Yechen Zhang, Shuhao Xing, Junhao Huang, Kai Lv, Yunhua Zhou, Xipeng Qiu, Qipeng Guo, and Kai Chen. Mousse: Rectifying the geometry of muon with curvature-aware preconditioning. *arXiv preprint arXiv:2603.09697*, 2026.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*, 2024.

A Related work

Kronecker-factored preconditioning. Shampoo (Gupta et al., 2018) introduced Kronecker-factored preconditioning as a tractable approximation to full-matrix Adagrad (Duchi et al., 2011); distributed implementations (Anil et al., 2020; Shi et al., 2023) scaled it to large models, with the Shampoo submission winning the AlgoPerf training-algorithm benchmark. SOAP (Vyas et al., 2024) improved its practical efficiency by running Adam in Shampoo's eigenbasis. Recent work analyzes Shampoo from multiple angles: convergence theory of one-sided variants (Xie et al., 2025), investigation of its heuristics by preconditioner decomposition (Eschenhagen et al., 2025), connection to optimal Kronecker approximation (Morwani et al., 2024), and broader theory for structured preconditioners and adaptive structured optimization (An et al., 2025). KL-Shampoo (Lin et al., 2025) recasts Shampoo- and SOAP-style estimation through KL minimization, while Clarifying Shampoo (Eschenhagen et al., 2026) sharpens the relationship between Shampoo and Muon-style spectral descent. An earlier line of work, K-FAC (Martens & Grosse, 2015) and its eigenbasis variant E-KFAC (George et al., 2018), predates Shampoo and develops Kronecker-factored curvature approximations for natural-gradient training, though it has seen comparatively slower adoption at LLM scale. Classical Kronecker covariance MLE (Dutilleul, 1999; Roś et al., 2016) provides the estimation-theoretic background. Recent work reduces the cost of Shampoo-type methods by batching preconditioner blocks and using faster inverse-root solvers (Modoranu et al., 2026), or by updating only part of the eigenbasis at a time (Milligan et al., 2026). Liu et al. (2026) instead modify the preconditioner, preconditioning the top part of the Kronecker spectrum by its eigenvalues and scaling the rest by a single adaptive constant, in the spirit of our scalar-tail variant Smok-Hop. Pro-KLShampoo keeps the full spectral estimate of one factor only on a tracked r-dimensional subspace, which also lowers its memory and QR cost, and orthogonalizes the update on its complement.

Orthogonalization and norm geometry. A complementary line of work interprets matrix-valued updates through norm geometry. Muon (Jordan et al., 2024) applies a Newton–Schulz approximation to an orthogonalized momentum update, and subsequent work interprets orthogonalization-style updates through spectral- or modular-norm steepest-descent viewpoints (Bernstein & Newhouse, 2024b; Large et al., 2024); the orthogonalization argument also inherits structure from matrix sign-SGD (Bernstein et al., 2018), with early roots in spectral descent for deep networks (Carlson et al., 2015). We use these geometric perspectives for interpreting the complement step, not for the restricted-KL derivation itself. Vyas et al. (2025) analyze the interaction between SOAP and Muon-style updates through an iterative-whitening lens; Mousse (Zhang et al., 2026) reformulates spectral-norm steepest descent in a Kronecker-whitened geometry.

Subspace structure in optimization. Several recent methods exploit low-rank or subspace structure in gradient updates: GaLore (Zhao et al., 2024) projects gradients into a low-rank subspace; other approaches include online subspace descent (Liang et al., 2024), SubTrack++ (Rajabi et al., 2025), and GoLore (He et al., 2024). COSMOS (Liu et al., 2025) decomposes updates into a SOAP-preconditioned subspace and a Muon-treated complement. Unlike low-rank gradient or update methods, Pro-KLShampoo restricts only the fine-grained spectral estimate to rank r; the update itself remains full-rank, with the complement orthogonalized rather than discarded. In our work, the subspace/complement decomposition emerges from restricting the KL objective to a spike-and-flat parametric family (Claims 2 and 3), rather than being imposed as an architectural choice.

Spectral structure and subspace tracking. Our spike-and-flat motivation is related to prior observations that deep-network optimization often concentrates in low-dimensional spectral structure. Gradient Descent Happens in a Tiny Subspace (Gur-Ari et al., 2018) and related observations on low-rank gradient structure (Jaiswal et al., 2024) support the relevance of low-dimensional gradient subspaces, while Ghorbani et al. (2019) study spectral concentration in Hessian eigenvalues during training, and the same spike–tail anisotropy is reported independently for the gradients themselves by Huang et al. (2026). Our use of a spike-and-flat model is also related to classical spiked-covariance theory (Johnstone, 2001; Paul, 2007). For the subspace-tracking component, we build on the classical online PCA update of Oja (1982), and finite-sample analyses of streaming PCA (Jain et al., 2016; Huang et al., 2021), and Davis–Kahan-type eigenspace perturbation tools (Yu et al., 2015). Nonconvex stochastic $O(1/\sqrt{T})$ convergence is classical (Ghadimi & Lan, 2013).

Convergence analyses of Muon. Several recent works analyze the convergence of Muon. Shen et al. (2025) provide nuclear-norm stationarity bounds for momentum Muon, matching the geometry of the complement term in our bound; Chang et al. (2025) remove the fixed-batch noise term with coupled step-size and momentum schedules and further improve the rate with two-batch variance reduction; Nagashima & Iiduka (2026) analyze Muon with and without Nesterov momentum and remove the fixed-batch term through batch growth; and Sfyraki & Wang (2025) interpret Muon with weight decay as stochastic Frank–Wolfe on a spectral-norm ball and measure convergence by the Frank–Wolfe gap, an interpretation that applies to the pure polar update and therefore does not directly cover our hybrid update. The batch-size dependence of our noise term matches that of signSGD (Bernstein et al., 2018). Theorem 1 analyzes the idealized Pro-KLShampoo update without momentum or variance reduction, and extending these mechanisms to our setting remains future work.

B TRAINING LOSS CURVES

These training loss curves support the wallclock saving mechanism analysis in §4.1. On GPT-2 (Figure 4), Pro-KLShampoo and KL-Shampoo training losses are close throughout training at every rank, indicating that convergence speed is not the source of wallclock saving. On LLaMA (Figure 5), Pro-KLShampoo's training loss lies visibly below KL-Shampoo's throughout training at every rank, indicating that faster convergence is the dominant source.

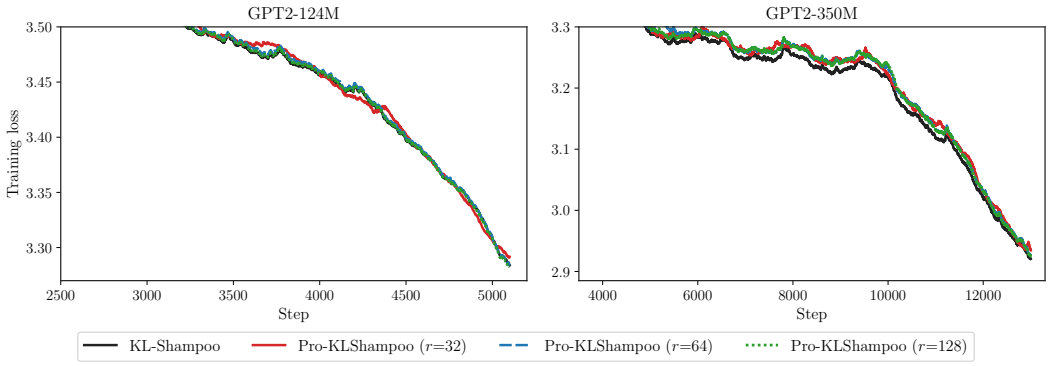

Figure 4: Training loss versus training step on GPT-2 (124M, left; 350M, right) for KL-Shampoo and Pro-KLShampoo at $r \in \{32, 64, 128\}$.

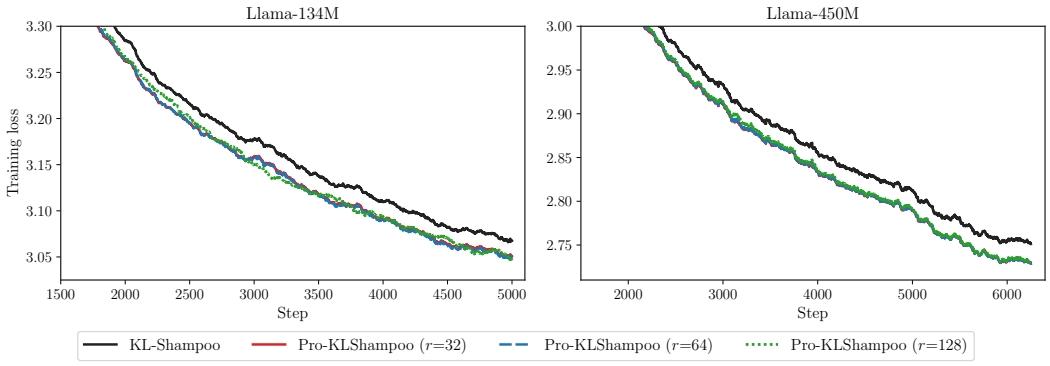

Figure 5: Training loss versus training step on LLaMA (134M, left; 450M, right) for KL-Shampoo and Pro-KLShampoo at $r \in \{32, 64, 128\}$.

C SPIKE-AND-FLAT STRUCTURE OF $\Phi_{L^*_{\text{restr}}}$

The conjecture in §3.2 concerns $\Phi_{L^*_{\text{restr}}}$, the matrix at the restricted-KL stationary point. Since L^*_{restr} is unobtainable in practice, we substitute the algorithm's running EMA estimate L_t and approximate $\Phi_{L_t} := \mathbb{E}[G^\top L_t^{-1} G]$ by the sample mean over a window of $W = 10$ consecutive minibatches: $\Phi_{L_t} \approx \frac{1}{W}\sum_{i=0}^{W-1} G_{t+i}^\top L_{t+i}^{-1} G_{t+i}$. We repeat this measurement every 250 training steps. Figure 6 shows the resulting eigenvalue spectra on GPT-2 (124M).

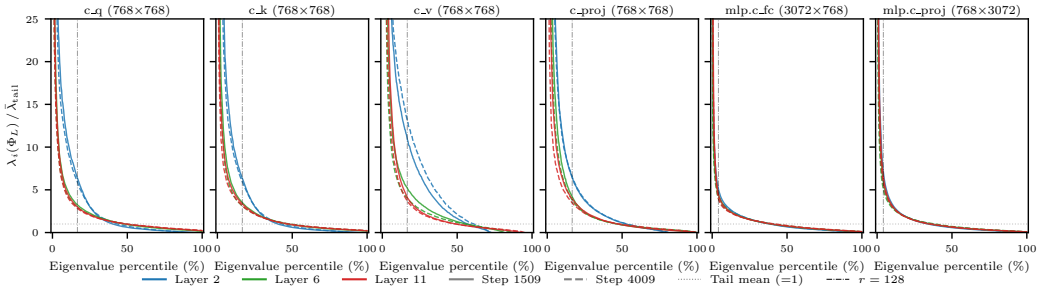

Figure 6: Eigenvalue spectra of Φ_{L_t} on GPT-2 (124M), normalized by the tail mean (vertical dash-dot line at $r = 128$). The spike-and-flat shape is present across all layer types, depths, and training stages, supporting the conjecture in §3.2, though less pronounced than the corresponding spectra of full KL-Shampoo's preconditioner (Figure 1).

D SPIKE-AND-FLAT STRUCTURE ON LLAMA

We confirm the spike-and-flat observation on LLaMA. Figure 7 shows the eigenvalue spectra of KL-Shampoo's Kronecker preconditioners during LLaMA training, in the same format as Figure 1.

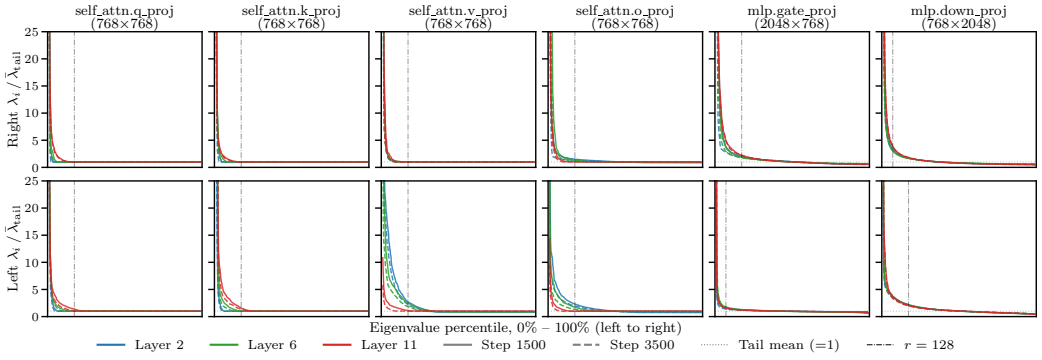

Figure 7: Eigenvalue spectra of KL-Shampoo's Kronecker preconditioners on LLaMA, normalized by the tail mean (vertical dash-dot line at $r = 128$). Both the right-side preconditioner R (top row) and the left-side preconditioner L (bottom row) show a spike-and-flat pattern across all layer types, depths, and training stages, mirroring the structure observed on GPT-2 (Figure 1).

E TAIL-FLATNESS STATISTICS

To quantify the spike-and-flat structure, we measure for each Kronecker factor the tail ratio

$$\rho_r = \frac{\text{AM}(\lambda_{r+1}, \ldots, \lambda_n)}{\text{GM}(\lambda_{r+1}, \ldots, \lambda_n)},$$

where $\lambda_1 \geq \cdots \geq \lambda_n$ are the eigenvalues of the factor and ρ_0 uses the full spectrum; it satisfies $\rho_r \geq 1$, and smaller values mean a flatter tail. We compute ρ_r for every layer and measured step of the KL-Shampoo runs, and Table 3 reports the median over the corresponding layer $\times$ measured-step records in the early and late stages of training, rounded to three decimals (entries are early / late).

Table 3: Tail ratios ρ_r of KL-Shampoo's Kronecker factors: medians over layer $\times$ measured-step records, early / late stage of training.

Scale	Layer type	ρ_0	ρ_{32}	ρ_{128}
GPT-2 124M	Attention	8.570 / 8.415	1.145 / 1.055	1.025 / 1.006
GPT-2 124M	MLP	2.529 / 1.894	1.388 / 1.348	1.195 / 1.186
LLaMA-134M	Attention	10.682 / 12.138	1.017 / 1.004	1.000 / 1.000
LLaMA-134M	MLP	1.590 / 1.176	1.154 / 1.062	1.082 / 1.029

Attention shows the clearest spike-and-flat structure. Its ρ_0 medians range from 8 to 12, while the tails become substantially flatter by $r{=}32$, and every ρ_{128} median is at most 1.025. MLP layers have smaller ρ_0 values and less uniform tails. The least flat case is GPT-2 124M MLP, with median $\rho_{128} \approx 1.19$; the LLaMA-134M MLP medians are at most 1.082. For each matrix, we also measure how much of the full-spectrum log-AM/GM gap is removed after separating the leading 128 directions from the tail; this fraction is $1 - \log \rho_{128} / \log \rho_0$, computed before taking medians. For GPT-2 124M MLP, the least flat case, $r{=}128$ removes approximately 78%–79% of the full-spectrum $\log(\mathrm{AM}/\mathrm{GM})$ gap, indicating that most of the dominant spectral variation is captured. This also helps explain why even the small $r{=}32$, which covers only 3–4% of the width, already performs well: the update does not lose the uncovered directions, because the orthogonalized complement (§3.3) updates them.

Three points support the spike-and-flat structure beyond the visual evidence of Figures 1 and 7. First, the measurement is systematic: we compute the tail ratio for every layer and measured step across two architectures (GPT-2 124M and LLaMA-134M) and report compact summaries by model, layer type, and training stage. Second, the structure is also supported theoretically: under its rank-ρ signal-plus-uncorrelated-noise model, Lemma 1 shows that the KL-stationary factors have at least $n - \rho$ equal bottom eigenvalues (here ρ denotes the signal rank of Lemma 1, not a tail ratio). Third, the same spike–tail anisotropy is reported independently for the gradients themselves by Huang et al. (2026).

F Proof of Claim 1 (approximation gap)

Proof. For the lower bound: for any $U \in \mathrm{St}(n,r)$, $S \in \mathbb{S}^r_{++}$, and $\mu_\perp > 0$, the matrix $USU^\top + \mu_\perp P_\perp$ belongs to $\mathbb{S}^n_{++}$. Therefore the feasible set of (4) is contained in that of (1), which gives $\mathcal{J}^{\mathrm{restr}} \geq \mathcal{J}^{\mathrm{full}}$.

For the upper bound, we exhibit a feasible candidate. Let $R^* = Q\,\mathrm{Diag}(\mu_1^*, \ldots, \mu_n^*)\,Q^\top$ be the eigendecomposition of the full KL optimum, with eigenvalues ordered $\mu_1^* \geq \cdots \geq \mu_n^* > 0$ and $Q \in \mathbb{R}^{n\times n}$ orthogonal. Define:

$$U_0 \coloneqq Q_{:,\,1:r} \in \mathrm{St}(n,r), \tag{15}$$

$$S_0 \coloneqq \mathrm{Diag}(\mu_1^*, \ldots, \mu_r^*) \in \mathbb{S}^r_{++}, \tag{16}$$

$$\bar{\mu} \coloneqq \frac{1}{n-r} \sum_{i=r+1}^{n} \mu_i^* > 0, \tag{17}$$

and set

$$\hat{R} \coloneqq U_0\, S_0\, U_0^\top + \bar{\mu}\,(I_n - U_0 U_0^\top). \tag{18}$$

Since $U_0 \in \mathrm{St}(n,r)$, $S_0 \in \mathbb{S}^r_{++}$, and $\bar{\mu} > 0$, the tuple $(L^*, U_0, S_0, \bar{\mu})$ is feasible for (4). Let $\mathcal{J}(L,R)$ denote the KL objective in (1) as a function of the pair (L,R), so that $\mathcal{J}^{\mathrm{full}} = \mathcal{J}(L^*, R^*)$. Hence

$$\mathcal{J}^{\mathrm{restr}} - \mathcal{J}^{\mathrm{full}} \;\leq\; \mathcal{J}(L^*, \hat{R}) - \mathcal{J}(L^*, R^*). \tag{19}$$

We next expand the difference $\mathcal{J}(L^*, \hat{R}) - \mathcal{J}(L^*, R^*)$. Using standard Kronecker identities (see, e.g., Lin et al. 2025, §2.2), the KL objective expands to

$$\mathcal{J}(L,R) = \tfrac{1}{2}\left[n \log\det L + m \log\det R + \mathrm{Tr}\big(R^{-1}\,\Phi_L\big) - \log\det\Sigma - mn\right], \tag{20}$$

where $\Phi_L \coloneqq \mathbb{E}[G^\top L^{-1} G] \in \mathbb{S}^n_{++}$ is the L-whitened gradient column second moment. Since both $\mathcal{J}(L^*, \hat{R})$ and $\mathcal{J}(L^*, R^*)$ share $L = L^*$, the terms $\frac{n}{2}\log\det L^*$, $\frac{1}{2}\log\det\Sigma$, and $\frac{mn}{2}$ in (20) cancel, and the difference reduces to

$$\mathcal{J}(L^*, \hat{R}) - \mathcal{J}(L^*, R^*) = \underbrace{\frac{m}{2}\big(\log\det\hat{R} - \log\det R^*\big)}_{\text{(I)}} + \underbrace{\frac{1}{2}\big(\mathrm{Tr}(\hat{R}^{-1}\Phi_{L^*}) - \mathrm{Tr}(R^{*-1}\Phi_{L^*})\big)}_{\text{(II)}}. \tag{21}$$

We show that (II) $= 0$ and (I) $= \frac{m(n-r)}{2}\log\frac{\mathrm{AM}(\mu^*_{r+1},\ldots,\mu^*_n)}{\mathrm{GM}(\mu^*_{r+1},\ldots,\mu^*_n)}$, which gives the desired bound.

We compute term (II), the trace difference, first. By the full stationarity condition (2), $R^* = \frac{1}{m}\,\Phi_{L^*}$, so $\Phi_{L^*} = m\,R^*$. Substituting:

$$\mathrm{Tr}(\hat{R}^{-1}\Phi_{L^*}) = m\,\mathrm{Tr}(\hat{R}^{-1}R^*), \qquad \mathrm{Tr}(R^{*-1}\Phi_{L^*}) = m\,\mathrm{Tr}(R^{*-1}R^*) = m\,\mathrm{Tr}(I_n) = mn. \tag{22}$$

It remains to compute $\mathrm{Tr}(\hat{R}^{-1}R^*)$. By construction (15)–(18), U_0 consists of the first r columns of Q. Since $\hat{R}$ is constructed from R^*'s eigenbasis, both $Q^\top R^* Q = \mathrm{Diag}(\mu_1^*, \ldots, \mu_n^*)$ and $Q^\top \hat{R}\, Q = \mathrm{Diag}(\mu_1^*, \ldots, \mu_r^*, \bar{\mu}, \ldots, \bar{\mu})$ are diagonal. Hence $\hat{R}^{-1}R^*$ has eigenvalues $\mu_i^*/\mu_i^* = 1$ for $i \leq r$ and $\mu_i^*/\bar{\mu}$ for $i > r$, and:

$$\mathrm{Tr}(\hat{R}^{-1}R^*) = \mathrm{Tr}(Q^\top \hat{R}^{-1} R^*\, Q) = r + \frac{1}{\bar{\mu}} \sum_{i=r+1}^{n} \mu_i^* = r + \frac{(n-r)\bar{\mu}}{\bar{\mu}} = n, \tag{23}$$

where we use the definition of $\bar{\mu}$ in (17). Substituting into (22): $\mathrm{Tr}(\hat{R}^{-1}\Phi_{L^*}) = mn$. Combined with $\mathrm{Tr}(R^{*-1}\Phi_{L^*}) = mn$, the trace difference (II) equals zero.

We turn to term (I), the log-det difference. By the same diagonalization,

$$\begin{aligned}\log\det\hat{R} - \log\det R^* &= \left[\sum_{i=1}^{r}\log\mu_i^* + (n{-}r)\log\bar{\mu}\right] - \sum_{i=1}^{n}\log\mu_i^* \\ &= (n{-}r)\log\bar{\mu} - \sum_{i=r+1}^{n}\log\mu_i^* \\ &= (n{-}r)\log\bar{\mu} - (n{-}r)\log\mathrm{GM}(\mu_{r+1}^*,\dots,\mu_n^*) \\ &= (n{-}r)\log\frac{\mathrm{AM}(\mu_{r+1}^*,\dots,\mu_n^*)}{\mathrm{GM}(\mu_{r+1}^*,\dots,\mu_n^*)},\end{aligned} \tag{24}$$

where we use the definition of $\bar{\mu}$ in (17) and the geometric mean.

Substituting (I) from (24) and (II) $= 0$ into (21):

$$\mathcal{J}(L^*,\hat{R}) - \mathcal{J}(L^*,R^*) = \frac{m}{2}\cdot(n{-}r)\log\frac{\mathrm{AM}}{\mathrm{GM}} + 0 = \frac{m(n{-}r)}{2}\log\frac{\mathrm{AM}}{\mathrm{GM}}. \tag{25}$$

By the AM–GM inequality, $\mathrm{AM}(\mu_{r+1}^*,\dots,\mu_n^*) \geq \mathrm{GM}(\mu_{r+1}^*,\dots,\mu_n^*)$ with equality if and only if $\mu_{r+1}^* = \cdots = \mu_n^*$. Therefore (25) is non-negative, and combined with the lower bound shown above this yields (5). □

G Proof of Claim 2 (restricted stationarity)

Proof. We first compute $\hat{R}^{-1}$. Since $\mathrm{range}(U)$ and $\mathrm{range}(P_\perp)$ are orthogonal and span $\mathbb{R}^n$, the inverse of $\hat{R} = USU^\top + \mu_\perp P_\perp$ can be written blockwise:

$$\hat{R}^{-1} = US^{-1}U^\top + \mu_\perp^{-1}P_\perp. \tag{26}$$

We now expand the KL objective. Substituting (26) into the trace term of (20) and using trace cyclicity, with $\Phi_L \coloneqq \mathbb{E}[G^\top L^{-1}G] \in \mathbb{S}_{++}^n$:

$$\begin{aligned}\mathrm{Tr}\big(\hat{R}^{-1}\,\Phi_L\big) &= \mathrm{Tr}\big(US^{-1}U^\top\,\Phi_L\big) + \mu_\perp^{-1}\,\mathrm{Tr}\big(P_\perp\,\Phi_L\big) \\ &= \mathrm{Tr}\big(S^{-1}\,U^\top\Phi_L\,U\big) + \mu_\perp^{-1}\,\mathrm{Tr}\big(P_\perp\,\Phi_L\big).\end{aligned} \tag{27}$$

For the log-det term, since $\hat{R}$ has eigenvalues $\{\text{eigenvalues of } S\} \cup \{\mu_\perp \text{ with multiplicity } n{-}r\}$:

$$\log\det\hat{R} = \log\det S + (n{-}r)\log\mu_\perp. \tag{28}$$

Isolating from (20) the terms involving $(S,\mu_\perp)$:

$$g(S,\mu_\perp) = \frac{m}{2}\big[\log\det S + (n{-}r)\log\mu_\perp\big] + \frac{1}{2}\big[\mathrm{Tr}(S^{-1}\,U^\top\Phi_L\,U) + \mu_\perp^{-1}\,\mathrm{Tr}(P_\perp\,\Phi_L)\big]. \tag{29}$$

The remaining terms ($n\log\det L$, $\log\det\Sigma$, mn) do not depend on S or $\mu_\perp$. The objective separates into a matrix part in S and a scalar part in $\mu_\perp$. Setting derivatives to zero:

$$\begin{aligned}\frac{\partial g}{\partial S^{-1}} &= -\frac{m}{2}\,S + \frac{1}{2}\,U^\top\Phi_L\,U = 0 \\ \Longrightarrow\quad S^* &= \frac{1}{m}\,U^\top\Phi_L\,U = \frac{1}{m}\,\mathbb{E}[\widetilde{G}^\top L^{-1}\widetilde{G}].\end{aligned} \tag{30}$$

$$\begin{aligned}\frac{\partial g}{\partial\mu_\perp} &= \frac{m(n-r)}{2\mu_\perp} - \frac{\mathrm{Tr}(P_\perp\,\Phi_L)}{2\mu_\perp^2} = 0 \\ \Longrightarrow\quad \mu_\perp^* &= \frac{\mathrm{Tr}(P_\perp\,\Phi_L)}{m(n-r)} = \frac{\mathrm{Tr}\big(\mathbb{E}[G_\perp^\top L^{-1}G_\perp]\big)}{m(n-r)}.\end{aligned} \tag{31}$$

For the optimality of L, the terms in $\mathcal{J}$ depending on L are $\frac{n}{2}\log\det L + \frac{1}{2}\mathrm{Tr}(\hat{R}^{-1}\mathbb{E}[G^\top L^{-1}G])$. Using $\mathrm{Tr}(\hat{R}^{-1}\mathbb{E}[G^\top L^{-1}G]) = \mathbb{E}[\mathrm{Tr}(L^{-1}G\hat{R}^{-1}G^\top)]$ (cyclic property), we differentiate with respect to L^{-1}:

$$\frac{\partial \mathcal{J}}{\partial L^{-1}} = -\frac{n}{2}L + \frac{1}{2}\mathbb{E}[G\hat{R}^{-1}G^\top] = 0$$
$$\implies \quad L^*_{\mathrm{restr}} = \frac{1}{n}\mathbb{E}[G\,\hat{R}^{-1}\,G^\top]. \tag{32}$$

□

H Proof of Claim 3 (optimal subspace)

Proof. The proof has two parts: we first show that every global minimizer of the reduced objective must be an eigenspace of Φ_L (using first- and second-order optimality), then evaluate the objective on eigenspaces to obtain the AM/GM characterization. Recall $\Phi_L = \mathbb{E}[G^\top L^{-1}G] \in \mathbb{S}^n_{++}$.

Reduced objective. We substitute $S^* = \frac{1}{m}U^\top\Phi_L U$ and $\mu^*_\perp = \frac{\mathrm{Tr}(P_\perp\Phi_L)}{m(n-r)}$ into $\mathcal{J}$, where these are the Claim 2 optima evaluated at the fixed L. Using the block decompositions (27)–(28) from the proof of Claim 2, the trace term becomes

$$\mathrm{Tr}(\hat{R}^{*-1}\Phi_L) = \mathrm{Tr}(S^{*-1}U^\top\Phi_L U) + \mu_\perp^{*-1}\mathrm{Tr}(P_\perp\Phi_L) = mr + m(n-r) = mn,$$

where the second equality uses $S^{*-1}U^\top\Phi_L U = m\,I_r$ and $\mu_\perp^{*-1}\mathrm{Tr}(P_\perp\Phi_L) = m(n-r)$. The log-det term $m\log\det\hat{R}^*$ expands via (28):

$$m\log\det\hat{R}^* = m\big[\log\det S^* + (n-r)\log\mu^*_\perp\big].$$

Substituting $S^* = \frac{1}{m}U^\top\Phi_L U$:

$$\log\det S^* = \log\det\big(\tfrac{1}{m}U^\top\Phi_L U\big) = \log\det(U^\top\Phi_L U) - r\log m.$$

Substituting $\mu^*_\perp = \frac{\mathrm{Tr}(P_\perp\Phi_L)}{m(n-r)}$:

$$(n-r)\log\mu^*_\perp = (n-r)\log\mathrm{Tr}(P_\perp\Phi_L) - (n-r)\log m - (n-r)\log(n-r).$$

Combining:

$$m\log\det\hat{R}^* = m\big[\log\det(U^\top\Phi_L U) - r\log m + (n-r)\log\mathrm{Tr}(P_\perp\Phi_L) - (n-r)\log m - (n-r)\log(n-r)\big].$$

The U-dependent part of $\mathcal{J}$ (up to the positive factor $m/2$) is therefore

$$f(U) \;:=\; \log\det\big(U^\top\Phi_L U\big) \;+\; (n-r)\log\mathrm{Tr}(P_\perp\Phi_L). \tag{33}$$

The claim reduces to solving

$$\min_{U\in\mathrm{St}(n,r)} \; f(U). \tag{34}$$

First-order condition. The Stiefel manifold $\mathrm{St}(n,r) = \{U\in\mathbb{R}^{n\times r} : U^\top U = I_r\}$ is a closed and bounded subset of $\mathbb{R}^{n\times r}$, hence compact. The function f is well-defined and continuous on $\mathrm{St}(n,r)$: the matrix $U^\top\Phi_L U$ is positive definite (since $\Phi_L \succ 0$ and U has orthonormal columns, hence full column rank), so $\det(U^\top\Phi_L U) > 0$; and $\mathrm{Tr}(P_\perp\Phi_L) > 0$ because $P_\perp$ is a nonzero projection ($r < n$) and $\Phi_L \succ 0$. Therefore the minimum in (34) is attained.

We derive the first-order optimality condition via the Lagrangian for the constraint $U^\top U = I_r$:

$$\mathcal{L}(U,\Lambda) \;=\; f(U) \;-\; \tfrac{1}{2}\mathrm{Tr}\big(\Lambda(U^\top U - I_r)\big), \qquad \Lambda\in\mathrm{Sym}(r).$$

Setting $\nabla_U\mathcal{L} = 0$ gives $\nabla_U f(U) = U\Lambda$. Computing the Euclidean gradient:

$$\nabla_U f(U) = 2\,\Phi_L U\big(U^\top\Phi_L U\big)^{-1} - \frac{2(n-r)}{\mathrm{Tr}(P_\perp\Phi_L)}\,\Phi_L U.$$

For brevity, write $A \coloneqq U^{*\top}\Phi_L\, U^* \in \mathbb{S}^r_{++}$ and $\beta \coloneqq (n-r)/\mathrm{Tr}(P_\perp \Phi_L) > 0$ at a critical point U^*. The KKT system reads

$$2\,\Phi_L\, U^*\big(A^{-1} - \beta\, I_r\big) \;=\; U^*\Lambda. \tag{35}$$

Left-multiplying by $U^{*\top}$ gives $\Lambda = 2(I_r - \beta A)$, which is indeed symmetric.

If $A^{-1} - \beta I_r$ is invertible, right-multiplying (35) by $(A^{-1} - \beta I_r)^{-1}$ and using $\Lambda = 2(I_r - \beta A) = 2A(A^{-1} - \beta I_r)$ gives

$$\Phi_L\, U^* \;=\; U^*\, A.$$

Diagonalize $A = V\mathrm{Diag}(\lambda_1, \ldots, \lambda_r)V^\top$ with orthogonal $V \in \mathbb{R}^{r\times r}$. Then

$$\Phi_L\,(U^*V) \;=\; U^*\,A\,V \;=\; (U^*V)\,\mathrm{Diag}(\lambda_1, \ldots, \lambda_r),$$

so each column of U^*V is an eigenvector of Φ_L. Since U^*V has the same column span as U^*, the minimizer U^* spans an eigenspace of Φ_L.

The same reasoning applies if $A^{-1} - \beta I_r$ is singular but $\mathrm{range}(U^*)$ happens to be Φ_L-invariant.

It remains to rule out the possibility that a global minimizer U^* has $A^{-1} - \beta I_r$ singular and $\mathrm{range}(U^*)$ **not** Φ_L-invariant. We do this by exhibiting a feasible descent direction, contradicting second-order optimality.

Second-order argument. Suppose for contradiction that U^* is a global minimizer of (34), that $A^{-1} - \beta I_r$ is singular, and that $\mathrm{range}(U^*)$ is *not* Φ_L-invariant.

Since $A^{-1} - \beta I_r$ is singular, there exists a unit vector $w \in \mathbb{R}^r$ with $Aw = w/\beta$. From $\Lambda = 2(I_r - \beta A)$ we get

$$\Lambda w = 2(I_r - \beta A)w = 2\big(w - \beta \cdot w/\beta\big) = 0.$$

Since $\mathrm{range}(U^*)$ is not Φ_L-invariant, there exists a unit vector $u_\perp \perp \mathrm{range}(U^*)$ whose image $\Phi_L u_\perp$ has a nonzero component in $\mathrm{range}(U^*)$:

$$\alpha \;\coloneqq\; U^{*\top}\Phi_L\, u_\perp \;\neq\; 0 \quad \in \mathbb{R}^r.$$

(If no such $u_\perp$ existed, Φ_L would map $\mathrm{range}(U^*)^\perp$ into itself, and by symmetry of Φ_L also $\mathrm{range}(U^*)$ into itself—contradicting non-invariance.)

Define $Z \coloneqq u_\perp\, w^\top \in \mathbb{R}^{n\times r}$. We claim Z is a feasible perturbation direction at U^*: setting $U(t) \coloneqq U^* + tZ$ and expanding $(U^* + tZ)^\top(U^* + tZ) = I_r + t(U^{*\top}Z + Z^\top U^*) + t^2 Z^\top Z$, the constraint $U(t)^\top U(t) = I_r$ is preserved to first order iff $U^{*\top}Z + Z^\top U^* = 0$. This holds since $U^{*\top}u_\perp = 0$ (because $u_\perp \perp \mathrm{range}(U^*)$).

Recall that $f(U) = \log\det(U^\top\Phi_L\, U) + (n-r)\log\mathrm{Tr}(P_\perp\Phi_L)$ is the reduced objective from (33). Along the path $U(t)$, define

$$A(t) \;\coloneqq\; U(t)^\top\,\Phi_L\, U(t), \qquad s(t) \;\coloneqq\; \mathrm{Tr}(\Phi_L) - \mathrm{Tr}A(t),$$

so that $f(U(t)) = \log\det A(t) + (n-r)\log s(t)$. At $t = 0$ we have $A(0) = A = U^{*\top}\Phi_L\, U^*$ and $s(0) = (n-r)/\beta$. We will show that $\frac{d^2}{dt^2} f(U(t))\big|_{t=0} < 0$, meaning f has strictly negative curvature at U^* along Z.

We compute $\frac{dA}{dt}\big|_{t=0}$ and $\frac{d^2A}{dt^2}\big|_{t=0}$. Expanding $A(t) = (U^* + tZ)^\top\Phi_L(U^* + tZ) = U^{*\top}\Phi_L U^* + t(Z^\top\Phi_L U^* + U^{*\top}\Phi_L Z) + t^2 Z^\top\Phi_L Z$:

$$\begin{aligned}
\left.\frac{dA}{dt}\right|_{t=0} &= Z^\top\Phi_L\, U^* + U^{*\top}\Phi_L\, Z \\
&= (u_\perp w^\top)^\top\Phi_L\, U^* + U^{*\top}\Phi_L\,(u_\perp w^\top) \\
&= w\,\underbrace{(u_\perp^\top\Phi_L\, U^*)}_{=\alpha^\top} + \underbrace{(U^{*\top}\Phi_L\, u_\perp)}_{=\alpha}\, w^\top \\
&= w\alpha^\top + \alpha w^\top.
\end{aligned} \tag{36}$$

$$\begin{aligned}
\left.\frac{d^2A}{dt^2}\right|_{t=0} &= 2\,Z^\top\Phi_L\, Z = 2\,(u_\perp w^\top)^\top\Phi_L\,(u_\perp w^\top) \\
&= 2\,w\,(u_\perp^\top\Phi_L\, u_\perp)\, w^\top = 2(u_\perp^\top\Phi_L\, u_\perp)\, ww^\top.
\end{aligned} \tag{37}$$

Note $u_\perp^\top \Phi_L\, u_\perp > 0$ since $\Phi_L \succ 0$.

We now compute $\frac{d^2}{dt^2} \log\det A(t)\big|_{t=0}$ and $\frac{d^2}{dt^2}(n-r)\log s(t)\big|_{t=0}$ separately.

The $\log\det A$ *term.* The standard matrix identity gives

$$\frac{d^2}{dt^2}\log\det A(t)\bigg|_{t=0} = \mathrm{Tr}\left(A^{-1}\,\frac{d^2A}{dt^2}\bigg|_{t=0}\right) - \mathrm{Tr}\left(\left(A^{-1}\,\frac{dA}{dt}\bigg|_{t=0}\right)^2\right).$$

For the first trace, substituting (37) and using $A^{-1}w = \beta w$ (recall $Aw = w/\beta$):

$$\begin{aligned}\mathrm{Tr}\big(A^{-1}\cdot 2(u_\perp^\top \Phi_L\, u_\perp)\, ww^\top\big) &= 2(u_\perp^\top \Phi_L\, u_\perp)\,\mathrm{Tr}(A^{-1}ww^\top)\\ &= 2(u_\perp^\top \Phi_L\, u_\perp)\, w^\top A^{-1} w\\ &= 2(u_\perp^\top \Phi_L\, u_\perp)\,\beta.\end{aligned}$$

For the second trace, substituting (36):

$$A^{-1}\frac{dA}{dt}\bigg|_{t=0} = A^{-1}(w\alpha^\top + \alpha w^\top) = \underbrace{(A^{-1}w)}_{=\beta w}\alpha^\top + (A^{-1}\alpha)\, w^\top = \beta\, w\alpha^\top + (A^{-1}\alpha)\, w^\top.$$

We expand the square of this $r\times r$ matrix and take the trace. There are four terms:

$$\begin{aligned}&(\beta\, w\alpha^\top)(\beta\, w\alpha^\top) = \beta^2\, w\,(\alpha^\top w)\,\alpha^\top, && \mathrm{Tr} = \beta^2\,(\alpha^\top w)^2,\\ &(\beta\, w\alpha^\top)(A^{-1}\alpha\, w^\top) = \beta\, w\,(\alpha^\top A^{-1}\alpha)\, w^\top, && \mathrm{Tr} = \beta\,\alpha^\top A^{-1}\alpha,\\ &(A^{-1}\alpha\, w^\top)(\beta\, w\alpha^\top) = \beta\,(A^{-1}\alpha)(w^\top w)\alpha^\top = \beta\,(A^{-1}\alpha)\alpha^\top, && \mathrm{Tr} = \beta\,\alpha^\top A^{-1}\alpha,\\ &(A^{-1}\alpha\, w^\top)(A^{-1}\alpha\, w^\top) = (A^{-1}\alpha)\underbrace{(w^\top A^{-1}\alpha)}_{=\beta\, w^\top\alpha} w^\top, && \mathrm{Tr} = \beta^2\,(\alpha^\top w)^2,\end{aligned}$$

where we used $w^\top w = 1$ and $w^\top A^{-1}\alpha = (A^{-1}w)^\top\alpha = \beta\, w^\top\alpha$. Summing:

$$\mathrm{Tr}\left(\left(A^{-1}\frac{dA}{dt}\bigg|_{t=0}\right)^2\right) = 2\beta^2(\alpha^\top w)^2 + 2\beta\,\alpha^\top A^{-1}\alpha. \tag{38}$$

The $\log\det$ contribution to $\frac{d^2 f}{dt^2}\big|_{t=0}$ is therefore

$$2(u_\perp^\top \Phi_L\, u_\perp)\,\beta \;-\; 2\beta^2(\alpha^\top w)^2 \;-\; 2\beta\,\alpha^\top A^{-1}\alpha. \tag{39}$$

The $(n-r)\log s$ *term.* Differentiating $s(t) = \mathrm{Tr}(\Phi_L) - \mathrm{Tr}A(t)$:

$$\begin{aligned}\frac{ds}{dt}\bigg|_{t=0} &= -\mathrm{Tr}\left(\frac{dA}{dt}\bigg|_{t=0}\right) = -\mathrm{Tr}(w\alpha^\top + \alpha w^\top) = -2\,\mathrm{Tr}(w\alpha^\top) = -2\,\alpha^\top w,\\ \frac{d^2 s}{dt^2}\bigg|_{t=0} &= -\mathrm{Tr}\left(\frac{d^2A}{dt^2}\bigg|_{t=0}\right) = -2(u_\perp^\top \Phi_L\, u_\perp).\end{aligned}$$

For a scalar function $\frac{d^2}{dt^2}\log s = \frac{1}{s^2}\big(s\,\frac{d^2 s}{dt^2} - \big(\frac{ds}{dt}\big)^2\big)$. At $t=0$, $s(0) = (n-r)/\beta$, so

$$\begin{aligned}\frac{d^2}{dt^2}(n-r)\log s\bigg|_{t=0} &= (n-r)\cdot\frac{\frac{n-r}{\beta}\cdot\big(-2(u_\perp^\top \Phi_L\, u_\perp)\big) - \big(-2\alpha^\top w\big)^2}{\big(\frac{n-r}{\beta}\big)^2}\\ &= (n-r)\cdot\frac{-\frac{2(n-r)}{\beta}(u_\perp^\top \Phi_L\, u_\perp) - 4(\alpha^\top w)^2}{\frac{(n-r)^2}{\beta^2}}\\ &= -2(u_\perp^\top \Phi_L\, u_\perp)\,\beta \;-\; \frac{4\beta^2(\alpha^\top w)^2}{n-r}.\end{aligned} \tag{40}$$

Combining and concluding. Adding (39) and (40), the $2(u_\perp^\top \Phi_L\, u_\perp)\,\beta$ terms cancel ($+$ from log-det, $-$ from $\log s$):

$$\left.\frac{d^2 f}{dt^2}\right|_{t=0} \;=\; -2\beta\,\alpha^\top A^{-1}\alpha \;-\; 2\beta^2(\alpha^\top w)^2\left(1+\frac{2}{n-r}\right) \;<\; 0, \tag{41}$$

since $\alpha \neq 0$, $A^{-1} \succ 0$, and $\beta > 0$.

By the second-order necessary condition for equality-constrained optimization (Nocedal & Wright, 2006, Theorem 12.5), the second derivative of the Lagrangian $\mathcal{L}(U) = f(U) - \frac{1}{2}\mathrm{Tr}(\Lambda(U^\top U - I_r))$ along the path $U(t)$ must be non-negative at a minimizer. Differentiating $\mathcal{L}(U(t))$ twice:

$$\left.\frac{d^2\mathcal{L}}{dt^2}\right|_{t=0} = \left.\frac{d^2 f}{dt^2}\right|_{t=0} - \frac{1}{2}\mathrm{Tr}\left(\Lambda\left.\frac{d^2(U(t)^\top U(t))}{dt^2}\right|_{t=0}\right) = \left.\frac{d^2 f}{dt^2}\right|_{t=0} - \mathrm{Tr}\big(\Lambda\, Z^\top Z\big)\,,$$

where we used $\frac{d^2}{dt^2}(U^*+tZ)^\top(U^*+tZ)\big|_{t=0} = 2Z^\top Z$. The necessary condition requires $\frac{d^2\mathcal{L}}{dt^2}\big|_{t=0} \geq 0$. The correction term vanishes: $Z^\top Z = (u_\perp w^\top)^\top(u_\perp w^\top) = ww^\top$, so $\mathrm{Tr}(\Lambda\, ww^\top) = w^\top\Lambda\, w = 0$ (since $\Lambda w = 0$). The condition reduces to $\frac{d^2 f}{dt^2}\big|_{t=0} \geq 0$, contradicting (41). Hence U^* cannot be a global minimizer, and every minimizer of (34) is an eigenspace of Φ_L.

Evaluating f on eigenspaces. Since every global minimizer is an eigenspace of Φ_L, it remains to compare f across the finitely many ($\binom{n}{r}$) eigenspace choices and find the best one. Denote the eigenvalues of Φ_L by $\phi_1 \geq \cdots \geq \phi_n > 0$. Let $I \subset \{1,\ldots,n\}$ with $|I| = r$ be an eigenspace choice and $J \coloneqq I^c = \{1,\ldots,n\} \setminus I$ its complement ($|J| = n{-}r$). Abusing notation, we write $f(I)$ for the value of f at any U whose columns span the eigenspace indexed by I. At such U, $U^\top\Phi_L U = \mathrm{Diag}(\phi_i : i \in I)$ and $\mathrm{Tr}(P_\perp \Phi_L) = \sum_{j\in J}\phi_j$, so

$$f(I) \;=\; \log\prod_{i\in I}\phi_i \;+\; (n{-}r)\log\sum_{j\in J}\phi_j.$$

Recall the arithmetic and geometric means of the complement eigenvalues:

$$\mathrm{AM}_J \coloneqq \frac{1}{n-r}\sum_{j\in J}\phi_j, \qquad \mathrm{GM}_J \coloneqq \Big(\prod_{j\in J}\phi_j\Big)^{1/(n-r)}.$$

We rewrite $f(I)$ in terms of these. For the first term, split the product over all indices:

$$\log\prod_{i\in I}\phi_i = \log\prod_{i=1}^{n}\phi_i - \log\prod_{j\in J}\phi_j = \sum_{i=1}^{n}\log\phi_i - (n{-}r)\log\mathrm{GM}_J.$$

For the second term, substitute $\sum_{j\in J}\phi_j = (n{-}r)\,\mathrm{AM}_J$:

$$(n{-}r)\log\sum_{j\in J}\phi_j = (n{-}r)\log\big((n{-}r)\,\mathrm{AM}_J\big) = (n{-}r)\log(n{-}r) + (n{-}r)\log\mathrm{AM}_J.$$

Combining, with $C_0 \coloneqq \sum_{i=1}^{n}\log\phi_i + (n{-}r)\log(n{-}r)$ independent of the choice I:

$$f(I) \;=\; C_0 \;+\; (n{-}r)\big(\log\mathrm{AM}_J - \log\mathrm{GM}_J\big) \;=\; C_0 \;+\; (n{-}r)\log\frac{\mathrm{AM}_J}{\mathrm{GM}_J}. \tag{42}$$

Since C_0 does not depend on I, the minimizer over eigenspace choices is

$$I^* \;\in\; \operatorname*{arg\,min}_{|I|=r}\; \frac{\mathrm{AM}_J}{\mathrm{GM}_J},$$

which is the characterization stated in Claim 3. In particular, when $\phi_{r+1} = \cdots = \phi_n$, the complement $J = \{r{+}1,\ldots,n\}$ achieves $\mathrm{AM}_J/\mathrm{GM}_J = 1$ (the minimum possible value), giving $I^* = \{1,\ldots,r\}$—the top-r eigenspace. More generally, whenever the bottom-$(n{-}r)$ eigenvalues have the smallest AM/GM ratio among all $(n{-}r)$-subsets (this is the spike-and-flat condition in Claim 3), the same conclusion $I^* = \{1,\ldots,r\}$ follows directly from the $\arg\min$ characterization above. □

I SPIKE-AND-FLAT EXACTNESS

Lemma 1 (Exactness of spike-and-flat under rank-ρ signal plus uncorrelated noise)**.** *Suppose $G = AB^\top + \xi$ where $A \in \mathbb{R}^{m\times\rho}$, $B \in \mathbb{R}^{n\times\rho}$ are deterministic with $\rho < \min(m,n)$, ξ has i.i.d. mean-zero entries with variance $\sigma^2 > 0$. Then any KL stationary point* (2) *satisfies $L^* = S_L^* + f_L^* I_m$ and $R^* = S_R^* + f_R^* I_n$ where $S_L^* \succeq 0$, $S_R^* \succeq 0$, $\mathrm{rank}(S_L^*) \le \rho$, $\mathrm{rank}(S_R^*) \le \rho$, and $f_L^*, f_R^* > 0$.*

Corollary 1 (Zero KL gap)**.** *Under the model of Lemma 1, $\mathcal{J}^{\mathrm{restr}} = \mathcal{J}^{\mathrm{full}}$ for any $r \ge \rho$.*

Remark (Fixed gradient row space). Lemma 1 concerns the population stationary point. The equal bottom eigenvalues also hold exactly along the EMA trajectory of idealized KL-Shampoo's right factor, $R_{t+1} = \beta_2 R_t + (1-\beta_2)\frac{1}{m} G_t^\top L_t^{-1} G_t$ started from $R_0 = \lambda_0 I_n$ (any positive definite L_t), whenever every gradient's row space lies in a fixed r-dimensional subspace $V \subseteq \mathbb{R}^n$: then $G_t v = 0$ for all $v \in V^\perp$, so the accumulation $G_t^\top L_t^{-1} G_t$ vanishes on $V^\perp$ and is block-diagonal with respect to $V \oplus V^\perp$, and by induction R_t acts on $V^\perp$ as $\beta_2^t \lambda_0\, I$ while its block on V is $\beta_2^t \lambda_0\, I$ plus a positive semidefinite matrix. The bottom $n-r$ eigenvalues of R_t therefore stay equal to $\beta_2^t \lambda_0$ at every t. This is the sense in which the spike-and-flat structure holds exactly under a low-rank gradient model.

Proof of Lemma 1. Substitute $G = AB^\top + \xi$ into the L stationarity (2) and use $\mathbb{E}[\xi] = 0$ to eliminate the cross-product terms:

$$\mathbb{E}[G\,(R^*)^{-1}\,G^\top] \;=\; AB^\top (R^*)^{-1} BA^\top \;+\; \mathbb{E}[\xi (R^*)^{-1} \xi^\top].$$

For the noise contribution, write the (i,i') entry explicitly:

$$\big(\xi (R^*)^{-1} \xi^\top\big)_{i,i'} \;=\; \sum_{j,j'} \xi_{i,j}\,(R^*)^{-1}_{j,j'}\,\xi_{i',j'}.$$

Taking expectation, the uncorrelated-noise assumption $\mathbb{E}[\xi_{ij}\xi_{i'j'}] = \sigma^2 \mathbf{1}_{(i,j)=(i',j')}$ collapses the double sum to its diagonal:

$$\mathbb{E}\big[(\xi (R^*)^{-1} \xi^\top)_{i,i'}\big] \;=\; \sigma^2\, \mathbf{1}_{i=i'} \sum_{j=1}^{n} (R^*)^{-1}_{j,j} \;=\; \sigma^2\, \mathrm{Tr}((R^*)^{-1})\, \mathbf{1}_{i=i'}.$$

Hence $\mathbb{E}[\xi (R^*)^{-1} \xi^\top] = \sigma^2\, \mathrm{Tr}((R^*)^{-1})\, I_m$, and the L stationarity becomes

$$L^* \;=\; \underbrace{\frac{1}{n} A \left(B^\top (R^*)^{-1} B\right) A^\top}_{=:\, S_L^*} \;+\; \underbrace{\frac{\sigma^2\, \mathrm{Tr}((R^*)^{-1})}{n}}_{=:\, f_L^*} I_m. \tag{43}$$

S_L^* is PSD because $B^\top (R^*)^{-1} B \succeq 0$ (sandwich of $(R^*)^{-1} \succ 0$); its column space lies in $\mathrm{range}(A)$ by construction; and $\mathrm{rank}(S_L^*) \le \mathrm{rank}(A) \le \rho$. The scalar f_L^* is strictly positive because $R^* \succ 0$ implies $\mathrm{Tr}((R^*)^{-1}) > 0$. This establishes the decomposition $L^* = S_L^* + f_L^*\, I_m$ with $\mathrm{rank}(S_L^*) \le \rho$ and $f_L^* > 0$.

The same calculation applied to the R stationarity yields

$$R^* \;=\; \underbrace{\frac{1}{m} B \left(A^\top (L^*)^{-1} A\right) B^\top}_{=:\, S_R^*} \;+\; \underbrace{\frac{\sigma^2\, \mathrm{Tr}((L^*)^{-1})}{m}}_{=:\, f_R^*} I_n, \tag{44}$$

which establishes the analogous decomposition $R^* = S_R^* + f_R^*\, I_n$ with $S_R^* \succeq 0$, $\mathrm{rank}(S_R^*) \le \rho$, and $f_R^* > 0$. □

Proof of Corollary 1. By Lemma 1, $R^* = S_R^* + f_R^*\, I_n$ with $S_R^* \succeq 0$ and $\mathrm{rank}(S_R^*) \le \rho$. Since S_R^* has at least $n - \rho$ zero eigenvalues, R^* has at least $n - \rho$ eigenvalues equal to f_R^*. For any $r \ge \rho$, choose $U \in \mathrm{St}(n, r)$ so that $\mathrm{range}(S_R^*) \subseteq \mathrm{range}(U)$ (possible since $\mathrm{rank}(S_R^*) \le \rho \le r$). Then $P_\perp S_R^* = 0$, so on the complement $P_\perp R^* P_\perp = f_R^*\, P_\perp$. Setting $S^* := U^\top R^* U$ and $\mu_\perp^* := f_R^*$, the decomposition (3) satisfies $U S^* U^\top + \mu_\perp^*\, P_\perp = R^*$ exactly, so the restricted family contains the full KL optimum. Hence $\mathcal{J}^{\mathrm{restr}} \le \mathcal{J}(L^*, R^*) = \mathcal{J}^{\mathrm{full}}$, and the reverse inequality is trivial. □

Why exact, not just asymptotic. The decomposition is exact at any finite m, n, and $\sigma > 0$: the identity $\mathbb{E}[\xi(R^*)^{-1}\xi^\top] = \sigma^2 \mathrm{Tr}((R^*)^{-1})\, I_m$ used in the proof holds at the population level. This contrasts with spiked random matrix theory (Johnstone, 2001; Paul, 2007), where the spike-plus-bulk structure of the *empirical* covariance is recovered only asymptotically as $m, n \to \infty$ at fixed ratio. Lemma 1 therefore says: under signal-plus-uncorrelated-noise gradients, the population-level KL stationary point is *precisely* of spike-and-flat structure regardless of dimension.

Beyond i.i.d. noise. Lemma 1 establishes spike-and-flat exactly under rank-ρ signal plus i.i.d. noise. We examine here what changes under more general noise. The flat floor in Lemma 1 originates from the identity $\mathbb{E}[\xi R^{-1}\xi^\top] = \sigma^2 \mathrm{Tr}(R^{-1})\, I_m$, which is a scalar multiple of identity exactly because $\mathbb{E}[\xi_{ij}\, \xi_{i'j'}] = \sigma^2\, \mathbf{1}_{(i,j)=(i',j')}$. Under separable noise covariance $\mathbb{E}[\mathrm{vec}(\xi)\mathrm{vec}(\xi)^\top] = \Sigma_L^\xi \otimes \Sigma_R^\xi$ — equivalently, $\mathbb{E}[\xi_{ij}\, \xi_{i'j'}] = (\Sigma_L^\xi)_{ii'}\, (\Sigma_R^\xi)_{jj'}$ — the same calculation gives

$$\begin{aligned}\mathbb{E}\big[(\xi R^{-1}\xi^\top)_{i,i'}\big] \;&=\; \sum_{j,j'} (R^{-1})_{j,j'}\, \mathbb{E}[\xi_{ij}\, \xi_{i'j'}] \;=\; (\Sigma_L^\xi)_{ii'} \sum_{j,j'} (R^{-1})_{j,j'}\, (\Sigma_R^\xi)_{jj'} \\ &=\; (\Sigma_L^\xi)_{ii'}\, \mathrm{Tr}(R^{-1}\, \Sigma_R^\xi),\end{aligned}$$

where the last step uses symmetry of Σ_R^ξ. Hence $\mathbb{E}[\xi R^{-1}\xi^\top] = \mathrm{Tr}(R^{-1}\Sigma_R^\xi)\, \Sigma_L^\xi$, and the L stationarity becomes $L^* = S_L^* + f_L^*\, \Sigma_L^\xi$ with $f_L^* \coloneqq \mathrm{Tr}((R^*)^{-1}\Sigma_R^\xi)/n$. The low-rank spike S_L^* persists, but the floor $f_L^*\, I_m$ becomes $f_L^*\, \Sigma_L^\xi$ and inherits the spectrum of Σ_L^ξ. The flat-tail structure in Lemma 1 thus relies on noise isotropy; under more general noise the structure becomes spike-plus-shaped-floor.

J PROOFS AND ADDITIONAL MATERIAL FOR §3.4

This appendix states and proves the convergence result and the supporting lemmas referenced in the main text.

Notation reminder. Throughout §J, the algorithm state at step t comprises $L_t \in \mathbb{S}^m_{++}$, $S_t \in \mathbb{S}^r_{++}$, $U_t \in \mathrm{St}(n,r)$ (subspace basis with orthonormal columns), and $\mu_{\perp,t} > 0$. We write $P_{\perp,t} \coloneqq I_n - U_t U_t^\top$ for the orthogonal projection onto $\mathrm{range}(U_t)^\perp$, and $\hat{R}_t \coloneqq U_t S_t U_t^\top + \mu_{\perp,t} P_{\perp,t}$ for the right-side preconditioner. The state evolves as in Algorithm 1: the targets ΔL_t, ΔS_t, $\delta_{\perp,t}$ and the EMA outputs L_{t+1}, $\bar{S}_{t+1}$, $\mu_{\perp,t+1}$ are formed with U_t (ΔS_t and $\bar{S}_{t+1}$ in the U_t coordinates); $T_t \coloneqq U_t^\top U_{t+1} \in \mathbb{R}^{r\times r}$ transports $\bar{S}_{t+1}$ to the U_{t+1} coordinates; and $L_{t+1}^{-1/2}$, $S_{t+1}^{-1/2}$ are computed from exact eigendecompositions of L_{t+1} and $S_{t+1} = T_t^\top \bar{S}_{t+1} T_t$. The stochastic gradient is G_t with mean $\nabla f(W_t)$. The polar factor of a matrix $M = U_M \Sigma_M V_M^\top$ (compact SVD) is $\mathrm{polar}(M) \coloneqq U_M V_M^\top$; it satisfies $\|\mathrm{polar}(M)\|_{\mathrm{op}} \le 1$ (it is a *partial isometry*: an SVD with all nonzero singular values equal to 1). We use $\|\cdot\|_F$ for the Frobenius norm, $\|\cdot\|_{\mathrm{op}}$ for the operator (spectral) norm, and $\|\cdot\|_* = \sum_i \sigma_i$ for the nuclear (trace) norm.

Let $\mathcal{F}_t \coloneqq \sigma(G_0, \dots, G_{t-1})$ denote the natural filtration (all randomness up to but not including step t); in particular, $W_t, L_t, S_t, U_t, \mu_{\perp,t}$ are $\mathcal{F}_t$-measurable. We assume:

(i) Lower boundedness: $f^* \coloneqq \inf_W f(W) > -\infty$.
(ii) L_{op}-operator-norm smoothness: $f(W') \le f(W) + \langle \nabla f(W), W' - W\rangle + \frac{L_{\mathrm{op}}}{2}\|W' - W\|^2_{\mathrm{op}}$.
(iii) Unbiased stochastic gradients with bounded Frobenius variance: $\mathbb{E}[G_t \mid \mathcal{F}_t] = \nabla f(W_t)$ and $\mathbb{E}[\|G_t - \nabla f(W_t)\|_F^2 \mid \mathcal{F}_t] \le \sigma_{F,t}^2$ for a deterministic sequence $(\sigma_{F,t})_{t\ge 0}$. The constant case is recovered by $\sigma_{F,t} \equiv \sigma_F$; where a single constant is used (Theorem 2 and §J.2.2) we write $\sigma_F \coloneqq \sup_{t\ge 0} \sigma_{F,t}$.
(iv) Bounded stochastic gradient: $\|G_t\|_{\mathrm{op}} \le G_{\max}$ a.s.
(v) Spectral lower bound: there is a constant $C > 0$ such that $\lambda_{\min}(L_t),\ \lambda_{\min}(S_t),\ \mu_{\perp,t} \ge 1/C^2$ for all $t \ge 0$. This is an explicit assumption on the trajectory of Algorithm 1, which has no clipping; Algorithm 2 enforces the corresponding bound on its decoupled eigenvalue estimates, whose inverses enter its update, by eigenvalue clipping (Appendix K). In particular, (v) requires T_t to be nonsingular at every step, since $\lambda_{\min}(S_{t+1}) \le \sigma_{\min}(T_t)^2\, \lambda_{\max}(\bar{S}_{t+1})$; Algorithm 2 needs no such requirement, since only its clipped eigenvalue estimates are inverted.

Operator-norm smoothness is the natural smoothness model for matrix-parameter layers, where updates act as operators on activations (Bernstein & Newhouse, 2024b; Large et al., 2024).

Invariants of Algorithm 1. Lemma 2 and Proposition 4 rely on the following properties of the state, each guaranteed by the indicated step of Algorithm 1 or assumption. (a) U_t has orthonormal columns: $U_0 \in \mathrm{St}(n,r)$, and Step 3 returns the orthonormal factor of a QR decomposition. (b) $Q_{L,t}$ and $Q_{S,t}$ are orthogonal: Step 4 eigendecomposes the symmetric matrices L_t and S_t. (c) L_t and S_t are symmetric positive definite with $\lambda_{\min}(L_t), \lambda_{\min}(S_t) \ge 1/C^2$, $\|L_t\|_{\mathrm{op}} \le \Theta_L$, and $\|S_t\|_{\mathrm{op}} \le \Theta_S$: symmetry and positive semidefiniteness follow from Step 2, since ΔL_t and ΔS_t are symmetric positive semidefinite and L_{t+1}, $\bar{S}_{t+1}$ are convex combinations, and from Step 3, since $S_{t+1} = T_t^\top \bar{S}_{t+1} T_t$ is a congruence; the lower bounds are Assumption (v), and the upper bounds are Lemma 2. (d) $1/C^2 \le \mu_{\perp,t} \le \Theta_\mu$: the lower bound is Assumption (v), the upper bound is Lemma 2 ($\delta_{\perp,t} \ge 0$ by Step 2). (e) The state $(W_t, L_t, S_t, U_t, \mu_{\perp,t})$ entering D_t is $\mathcal{F}_t$-measurable: Step 1 uses only this state and G_t, and Steps 2–4 form the step-$(t+1)$ state from the step-t state and G_t, so the claim follows by induction from the initial state, which is fixed before G_0 is drawn. (f) $T_t = U_t^\top U_{t+1}$ is nonsingular with $\|T_t\|_{\mathrm{op}} \le 1$: the norm bound follows from (a), and nonsingularity is required by Assumption (v) through $\lambda_{\min}(S_{t+1}) \le \sigma_{\min}(T_t)^2\, \lambda_{\max}(\bar{S}_{t+1})$.

Lemma 2 (Preconditioner bounds)**.** *Let Assumptions (iv) and (v) hold along the trajectory of Algorithm 1, and set*

$$\Theta_L \coloneqq \max\Big(\|L_0\|_{\mathrm{op}},\ \tfrac{2C^2G_{\max}^2}{n}\Big), \qquad \Theta_S \coloneqq \max\Big(\|S_0\|_{\mathrm{op}},\ \tfrac{C^2G_{\max}^2}{m}\Big),$$
$$\Theta_\mu \coloneqq \max\Big(\mu_{\perp,0},\ \tfrac{C^2G_{\max}^2}{m}\Big), \qquad \Theta \coloneqq \max(\Theta_L, \Theta_S, \Theta_\mu).$$

Then for all $t \geq 0$, $\|L_t\|_{\mathrm{op}} \leq \Theta_L$, $\|S_t\|_{\mathrm{op}} \leq \Theta_S$, and $\mu_{\perp,t} \leq \Theta_\mu$; in particular,

$$\tfrac{1}{C^2} \;\leq\; \lambda_i(L_t),\; \lambda_j(S_t),\; \mu_{\perp,t} \;\leq\; \Theta \qquad \text{for all } i \in \{1,\dots,m\},\; j \in \{1,\dots,r\}. \tag{45}$$

The lower bounds are Assumption (v); the upper bounds are proved below. Since $n \geq 2$, $\Theta \leq \max(\|L_0\|_{\mathrm{op}}, \|S_0\|_{\mathrm{op}}, \mu_{\perp,0}, C^2 G_{\max}^2)$.

Proof. The lower bounds hold by Assumption (v). For the upper bounds we proceed by induction on t. At $t = 0$ they hold by the definition of $\Theta_L, \Theta_S, \Theta_\mu$. Assume $\|L_t\|_{\mathrm{op}} \leq \Theta_L$, $\|S_t\|_{\mathrm{op}} \leq \Theta_S$, and $\mu_{\perp,t} \leq \Theta_\mu$.

The three targets are uniformly bounded. By Assumption (v), $S_t^{-1} \preceq C^2 I_r$, $L_t^{-1} \preceq C^2 I_m$, and $\mu_{\perp,t}^{-1} \leq C^2$. Since U_t has orthonormal columns and $P_{\perp,t}$ is an orthogonal projection, Assumption (iv) gives $\|G_t U_t\|_{\mathrm{op}} \leq G_{\max}$ and $\|G_t P_{\perp,t}\|_{\mathrm{op}} \leq G_{\max}$; moreover $\|G_t P_{\perp,t}\|_F^2 = \mathrm{Tr}(P_{\perp,t} G_t^\top G_t P_{\perp,t}) \leq \|G_t\|_{\mathrm{op}}^2 \mathrm{Tr}(P_{\perp,t}) \leq G_{\max}^2 (n - r)$, hence $\|L_t^{-1/2} G_t P_{\perp,t}\|_F^2 \leq \|L_t^{-1/2}\|_{\mathrm{op}}^2 \|G_t P_{\perp,t}\|_F^2 \leq C^2 G_{\max}^2 (n - r)$. For the targets of Step 2 of Algorithm 1, writing $G_t P_{\perp,t} G_t^\top = (G_t P_{\perp,t})(G_t P_{\perp,t})^\top$,

$$\begin{aligned}
\|\Delta L_t\|_{\mathrm{op}} &\leq \tfrac{1}{n}\big(\|G_t U_t\|_{\mathrm{op}}^2 \|S_t^{-1}\|_{\mathrm{op}} + \mu_{\perp,t}^{-1} \|G_t P_{\perp,t}\|_{\mathrm{op}}^2\big) \;\leq\; \tfrac{2C^2 G_{\max}^2}{n} \;\leq\; \Theta_L, \\
\|\Delta S_t\|_{\mathrm{op}} &\leq \tfrac{1}{m} \|G_t U_t\|_{\mathrm{op}}^2 \|L_t^{-1}\|_{\mathrm{op}} \;\leq\; \tfrac{C^2 G_{\max}^2}{m} \;\leq\; \Theta_S, \\
\delta_{\perp,t} &= \tfrac{1}{m(n-r)} \|L_t^{-1/2} G_t P_{\perp,t}\|_F^2 \;\leq\; \tfrac{C^2 G_{\max}^2}{m} \;\leq\; \Theta_\mu.
\end{aligned}$$

The EMAs preserve the caps. Each EMA output is a convex combination of two matrices (scalars) whose operator norms (values) are at most the corresponding cap: $\|L_{t+1}\|_{\mathrm{op}} \leq \beta_2 \|L_t\|_{\mathrm{op}} + (1 - \beta_2)\|\Delta L_t\|_{\mathrm{op}} \leq \Theta_L$, and likewise $\|\bar{S}_{t+1}\|_{\mathrm{op}} \leq \Theta_S$ and $\mu_{\perp,t+1} \leq \Theta_\mu$.

The transport preserves the cap. Since U_t and U_{t+1} have orthonormal columns, $\|T_t\|_{\mathrm{op}} = \|U_t^\top U_{t+1}\|_{\mathrm{op}} \leq \|U_t\|_{\mathrm{op}} \|U_{t+1}\|_{\mathrm{op}} = 1$, hence

$$\|S_{t+1}\|_{\mathrm{op}} \;=\; \|T_t^\top \bar{S}_{t+1} T_t\|_{\mathrm{op}} \;\leq\; \|T_t\|_{\mathrm{op}}^2 \|\bar{S}_{t+1}\|_{\mathrm{op}} \;\leq\; \Theta_S.$$

This completes the induction. The inverse square roots entering D_t are computed from exact eigendecompositions of L_t and S_t (Step 4 of Algorithm 1), so (45) applies to the operators that appear in the update. □

J.1 Main convergence theorem

This subsection contains the convergence result for Pro-KLShampoo (Algorithm 1): Theorem 1, stated in §3.4, is the stationarity bound up to a noise-dependent neighborhood for the idealized update with exact eigendecompositions and is proved in §J.1.4; Lemma 3 shows that its measure vanishes only at stationary points; the comparison paragraph and Corollary 2 relate it to worst-case constants and to Frobenius-norm stationarity. Smok-Hop (without polar orthogonalization) is analyzed separately as a technical companion in §J.2 below.

Remark (Order of the state updates). Theorem 1 also holds if the statistics update is performed after the transport, in the U_{t+1} coordinates (an option in our implementation; see Appendix K), or if the subspace step uses $\bar{S}_{t+1}$ in place of S_t. The proof uses only the predictability, orthonormality, and spectral bounds of the state entering D_t, and these still hold; in the first case, Assumption (v) is also imposed on $T_t^\top S_t T_t$, whose inverse then enters ΔL_t.

Remark (Clipped variant). Suppose Step 4 of Algorithm 1 also raises every eigenvalue of L_{t+1} and S_{t+1}, and the scalar $\mu_{\perp,t+1}$, to at least $1/C^2$, and the initial state satisfies Assumption (v). Then Assumption (v) holds by construction, and Theorem 1 holds for this variant with the same proof and constants: the clipping is a deterministic function of the state, so the state entering D_t stays predictable; it keeps L_{t+1} and S_{t+1} symmetric positive definite; and it keeps the upper bounds of Lemma 2, because $\Theta_L, \Theta_S, \Theta_\mu \geq 1/C^2$ when the initial state satisfies Assumption (v).

Lemma 3. *For any $W \in \mathbb{R}^{m \times n}$, $L \in \mathbb{S}_{++}^m$, $S \in \mathbb{S}_{++}^r$, $U \in \mathrm{St}(n, r)$, and $P_\perp := I_n - UU^\top$,*

$$c_a \sqrt{\lambda_{\min}(L)}\, \big\|L^{-1/2} \nabla f(W)\, P_\perp\big\|_* + \alpha_{\mathrm{kl}} \big\|L^{-1/4} \nabla f(W)\, U\, S^{-1/4}\big\|_F^2 \;=\; 0$$

if and only if $\nabla f(W) = 0$. In particular, $M_t = 0$ if and only if $\nabla f(W_t) = 0$.

The right-hand side of (14) gives $O(1/\sqrt{T})$ decay to a noise-dependent neighborhood of stationarity of width $2c_a\sqrt{k}\,\frac{1}{T}\sum_{t<T}\sigma_{F,t}$ (a constant floor $2c_a\sqrt{k}\,\sigma_F$ when $\sigma_{F,t}\equiv\sigma_F$). The measure is *state-dependent*: it splits the gradient $\nabla f(W_t)$ via the algorithm's running $(U_t, P_{\perp,t})$ and weights the two pieces by the per-step spectral quantities of (L_t, S_t) rather than by global worst-case constants. Lemma 3 shows the left-hand side vanishes if and only if $\nabla f(W_t) = 0$.

Comparison with worst-case constants. Define $\kappa_t \coloneqq \kappa(L_t) = \lambda_{\max}(L_t)/\lambda_{\min}(L_t)$ and $b_t \coloneqq \lambda_{\max}(L_t)\,\lambda_{\max}(S_t)$. By Lemma 4,

$$M_t \;\geq\; \frac{c_a}{\sqrt{\kappa_t}}\big\|\nabla f(W_t)\,P_{\perp,t}\big\|_* \;+\; \frac{\alpha_{\mathrm{kl}}}{\sqrt{b_t}}\big\|\nabla f(W_t)\,U_t\big\|_F^2. \tag{46}$$

Under Lemma 2, $\sqrt{\kappa_t} \leq C\sqrt{\Theta_L}$ and $\sqrt{b_t} \leq \sqrt{\Theta_L\Theta_S}$; with $\Theta = \max(\Theta_L, \Theta_S, \Theta_\mu)$ these give $\sqrt{\kappa_t} \leq C\sqrt{\Theta}$ and $\sqrt{b_t} \leq \Theta$, so Theorem 1 implies the weaker bound with the global constants $c_a/(C\sqrt{\Theta})$ and $\alpha_{\mathrm{kl}}/\Theta$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\tfrac{c_a}{C\sqrt{\Theta}}\big\|\nabla f(W_t)\,P_{\perp,t}\big\|_* + \tfrac{\alpha_{\mathrm{kl}}}{\Theta}\big\|\nabla f(W_t)\,U_t\big\|_F^2\Big] \;\leq\; 2\sqrt{\frac{\Delta_0\,L_{\mathrm{op}}\,K}{T}} + 2\,c_a\sqrt{k}\,\frac{1}{T}\sum_{t=0}^{T-1}\sigma_{F,t}. \tag{47}$$

The bound (14) retains the trajectory-dependent spectral factors before these global worst-case relaxations are applied and is therefore pointwise tighter.

Corollary 2 (Frobenius-norm stationarity)**.** *Under the assumptions of Theorem 1, suppose that almost surely $\kappa_t \leq K_L$ and $b_t \leq B_{LS}$ for all $t \geq 0$. Then, writing ε_T for the right-hand side of* (14)*,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(W_t)\|_F^2 \;\leq\; \max\left\{\frac{G_{\max}\sqrt{K_L}}{c_a},\;\frac{\sqrt{B_{LS}}}{\alpha_{\mathrm{kl}}}\right\}\varepsilon_T. \tag{48}$$

Proof. Since $\nabla f(W_t) = \nabla f(W_t)\,U_tU_t^\top + \nabla f(W_t)\,P_{\perp,t}$ with $P_{\perp,t}U_t = 0$, the two pieces have orthogonal row spaces and $\|\nabla f(W_t)\|_F^2 = \|\nabla f(W_t)\,U_t\|_F^2 + \|\nabla f(W_t)\,P_{\perp,t}\|_F^2$. For the complement piece, $\|X\|_F^2 \leq \|X\|_{\mathrm{op}}\|X\|_*$ with $X = \nabla f(W_t)\,P_{\perp,t}$, and $\|X\|_{\mathrm{op}} \leq \|\nabla f(W_t)\|_{\mathrm{op}} \leq G_{\max}$, where the last step is Jensen's inequality applied to $\nabla f(W_t) = \mathbb{E}[G_t \mid \mathcal{F}_t]$ together with Assumption (iv). Hence, by (46) and the assumed bounds $\kappa_t \leq K_L$, $b_t \leq B_{LS}$,

$$\begin{aligned}\|\nabla f(W_t)\|_F^2 \;&\leq\; G_{\max}\|\nabla f(W_t)\,P_{\perp,t}\|_* + \|\nabla f(W_t)\,U_t\|_F^2\\ &\leq\; \frac{G_{\max}\sqrt{K_L}}{c_a}\cdot\frac{c_a}{\sqrt{\kappa_t}}\|\nabla f(W_t)\,P_{\perp,t}\|_* + \frac{\sqrt{B_{LS}}}{\alpha_{\mathrm{kl}}}\cdot\frac{\alpha_{\mathrm{kl}}}{\sqrt{b_t}}\|\nabla f(W_t)\,U_t\|_F^2\\ &\leq\; \max\Big\{\tfrac{G_{\max}\sqrt{K_L}}{c_a},\;\tfrac{\sqrt{B_{LS}}}{\alpha_{\mathrm{kl}}}\Big\}\,M_t.\end{aligned}$$

Taking expectations, averaging over $t = 0, \ldots, T-1$, and applying Theorem 1 gives (48). □

J.1.1 Single-step descent inequality

We first state the per-step descent inequality from which Theorem 1 is obtained by telescoping. The constant σ_{kl}^2, defined in Theorem 1 of the main text, satisfies $\mathbb{E}[\|L_t^{-1/2}G_tU_tS_t^{-1/2}\|_{\mathrm{op}}^2 \mid \mathcal{F}_t] \leq \sigma_{kl}^2$ a.s. for all $t \geq 0$ by construction (see §J.1.6 for reference values). Throughout, we write $W_{t+1} = W_t + \eta\,\Delta W_t$ for the Pro-KLShampoo update from Step 1 of Algorithm 1, i.e., $\Delta W_t \coloneqq -D_t$. By the scale invariance $\mathrm{polar}(\alpha M) = \mathrm{polar}(M)$ for $\alpha > 0$ and $\mu_{\perp,t} > 0$, the complement term in (7) simplifies to $\mathrm{polar}(L_t^{-1/2}\,G_t\,P_{\perp,t})$, so ΔW_t takes the form

$$\Delta W_t \;=\; -\alpha_{\mathrm{kl}}\,L_t^{-1/2}\,G_t\,U_t\,S_t^{-1/2}\,U_t^\top \;-\; c_a\,\mathrm{polar}\big(L_t^{-1/2}\,G_t\,P_{\perp,t}\big).$$

Proposition 4 (Descent inequality)**.** *Under Assumptions (i)–(v), Algorithm 1 satisfies for every $t \geq 0$, with M_t as in* (13)*:*

$$\mathbb{E}\big[f(W_{t+1}) - f(W_t)\,\big|\,\mathcal{F}_t\big] \;\leq\; -\eta\,M_t \;+\; 2\eta\,c_a\sqrt{k}\,\sigma_{F,t} \;+\; \eta^2 L_{\mathrm{op}}\big(c_a^2 + \alpha_{\mathrm{kl}}^2\,\sigma_{kl}^2\big), \tag{49}$$

where $c_a = \sqrt{\max(1, m/n)}$ and $k = \min(m, n-r)$. The complement descent is measured by the nuclear norm $\|M\|_* \coloneqq \sum_i \sigma_i(M)$ (from the identity $\langle M, \mathrm{polar}(M)\rangle = \|M\|_*$); the subspace descent is measured by the Frobenius norm squared.

J.1.2 Proof of Proposition 4

Proof of Proposition 4. We define several shorthand quantities. The noise is $N_t \coloneqq G_t - \nabla f(W_t)$, and:

$$A_t \coloneqq L_t^{-1/2}\,\nabla f(W_t)\,P_{\perp,t} \quad \text{(complement gradient, with } L_t^{-1/2} \text{ on the left)},$$
$$Z_t \coloneqq L_t^{-1/2}\,N_t\,P_{\perp,t} \quad \text{(complement noise, with } L_t^{-1/2} \text{ on the left)},$$
$$V_t \coloneqq L_t^{-1/4}\,\nabla f(W_t)\,U_t \quad \text{(subspace gradient, with } L_t^{-1/4} \text{ on the left)}.$$

A_t and V_t are $\mathcal{F}_t$-measurable (they depend on the randomness only through W_t, L_t, U_t); N_t and Z_t are not. The inequalities below that involve Z_t hold pathwise, so its measurability is never used; the conditional-expectation steps use only the $\mathcal{F}_t$-measurability of (W_t, L_t, S_t, U_t) and Assumption (iii). Note that $L_t^{-1/2}G_tP_{\perp,t} = A_t + Z_t$. We write $\langle X, Y\rangle \coloneqq \mathrm{Tr}(X^\top Y)$ for the Frobenius (trace) inner product on matrices.

Step 0. By operator-norm smoothness (Assumption (ii)) and $W_{t+1} = W_t + \eta\,\Delta W_t$:

$$f(W_{t+1}) - f(W_t) \;\le\; \eta\,\langle \nabla f(W_t),\,\Delta W_t\rangle \;+\; \tfrac{\eta^2 L_{\mathrm{op}}}{2}\,\|\Delta W_t\|_{\mathrm{op}}^2. \tag{50}$$

From (7),

$$\Delta W_t = -\alpha_{\mathrm{kl}}\,L_t^{-1/2}\,G_t\,U_t\,S_t^{-1/2}\,U_t^\top \;-\; c_a\,\mathrm{polar}\big(L_t^{-1/2}\,G_t\,P_{\perp,t}\big),$$

so $\langle \nabla f(W_t), \Delta W_t\rangle$ splits into a subspace piece and a complement piece. Steps 1–3 below bound the conditional expectations of the two inner-product pieces and the second-order term $\|\Delta W_t\|_{\mathrm{op}}^2$.

Step 1. We prove the lower bound

$$\mathbb{E}\big[\langle \nabla f(W_t),\, c_a\mathrm{polar}(L_t^{-1/2}G_tP_{\perp,t})\rangle \;\big|\; \mathcal{F}_t\big] \;\ge\; c_a\sqrt{\lambda_{\min}(L_t)}\,\|A_t\|_* \;-\; 2c_a\sqrt{k}\,\sigma_{F,t}. \tag{51}$$

Write $M \coloneqq A_t + Z_t = L_t^{-1/2}G_tP_{\perp,t}$. Since $P_{\perp,t}$ is a projection, the compact SVD of M has right singular vectors in $\mathrm{range}(P_{\perp,t})$, so $\mathrm{polar}(M)\,P_{\perp,t} = \mathrm{polar}(M)$. Using $P_{\perp,t}^\top = P_{\perp,t}$ and trace cyclicity:

$$\begin{aligned}
\langle \nabla f(W_t),\,\mathrm{polar}(M)\rangle &= \langle \nabla f(W_t)\,P_{\perp,t},\,\mathrm{polar}(M)\rangle \\
&= \langle L_t^{1/2}\,A_t,\,\mathrm{polar}(M)\rangle \\
&= \underbrace{\langle L_t^{1/2}\,M,\,\mathrm{polar}(M)\rangle}_{=:\,\spadesuit_t} \;-\; \underbrace{\langle L_t^{1/2}\,Z_t,\,\mathrm{polar}(M)\rangle}_{=:\,\diamondsuit_t},
\end{aligned}$$

where the last step uses $A_t = M - Z_t$.

Bounding $\spadesuit_t$ from below. Let $M = \sum_i \sigma_i\,u_iv_i^\top$ be the SVD, so $\mathrm{polar}(M) = \sum_i u_iv_i^\top$. Then

$$\spadesuit_t \;=\; \sum_i \sigma_i\,(u_i^\top L_t^{1/2}u_i) \;\ge\; \sqrt{\lambda_{\min}(L_t)}\,\sum_i \sigma_i \;=\; \sqrt{\lambda_{\min}(L_t)}\,\|M\|_*,$$

where we used $u_i^\top L_t^{1/2}u_i \ge \sqrt{\lambda_{\min}(L_t)}$ for the unit vectors u_i. Since the target (51) is in terms of $\|A_t\|_*$ rather than $\|M\|_*$, we use the nuclear-norm triangle inequality $\|M\|_* = \|A_t + Z_t\|_* \ge \|A_t\|_* - \|Z_t\|_*$ and bound $\|Z_t\|_*$:

$$\begin{aligned}
\|Z_t\|_* &= \|L_t^{-1/2}\,N_t\,P_{\perp,t}\|_* \\
&\le \sqrt{\mathrm{rank}(Z_t)}\,\|Z_t\|_F \\
&\le \sqrt{\mathrm{rank}(Z_t)}\,\|L_t^{-1/2}\|_{\mathrm{op}}\,\|N_t\,P_{\perp,t}\|_F \\
&\le \sqrt{k}\,\|L_t^{-1/2}\|_{\mathrm{op}}\,\|N_t\|_F \\
&= \sqrt{k}\,\lambda_{\min}(L_t)^{-1/2}\,\|N_t\|_F.
\end{aligned}$$

The first inequality uses $\|X\|_* \le \sqrt{\mathrm{rank}(X)}\,\|X\|_F$. The second splits the Frobenius norm via sub-multiplicativity $\|MN\|_F \le \|M\|_{\mathrm{op}}\,\|N\|_F$. The third combines $\mathrm{rank}(Z_t) \le \min(m, n-r) = k$

and $\|N_t P_{\perp,t}\|_F \le \|N_t\|_F$ (since $\|P_{\perp,t}\|_{\mathrm{op}} \le 1$). The last step uses $\|L_t^{-1/2}\|_{\mathrm{op}} = \lambda_{\min}(L_t)^{-1/2}$. Combining, $\sqrt{\lambda_{\min}(L_t)}\,\|Z_t\|_* \le \sqrt{k}\,\|N_t\|_F$ and

$$\spadesuit_t \;\ge\; \sqrt{\lambda_{\min}(L_t)}\big(\|A_t\|_* - \|Z_t\|_*\big) \;\ge\; \sqrt{\lambda_{\min}(L_t)}\,\|A_t\|_* - \sqrt{k}\,\|N_t\|_F.$$

Bounding $|\diamondsuit_t|$ from above. Since $\|\mathrm{polar}(M)\|_{\mathrm{op}} \le 1$, the matrix Hölder inequality $|\langle X, Y\rangle| \le \|X\|_*\,\|Y\|_{\mathrm{op}}$ gives

$$|\diamondsuit_t| \;\le\; \|L_t^{1/2}\,Z_t\|_* \;=\; \|N_t\,P_{\perp,t}\|_* \;\le\; \sqrt{k}\,\|N_t\,P_{\perp,t}\|_F \;\le\; \sqrt{k}\,\|N_t\|_F,$$

where the equality uses $L_t^{1/2}\,Z_t = L_t^{1/2}\cdot L_t^{-1/2}\,N_t\,P_{\perp,t} = N_t\,P_{\perp,t}$, and the next step uses $\|X\|_* \le \sqrt{\mathrm{rank}(X)}\,\|X\|_F$ with $\mathrm{rank}(N_t P_{\perp,t}) \le \min(m, n-r) = k$.

Combining.

$$\langle \nabla f(W_t),\, \mathrm{polar}(M)\rangle \;=\; \spadesuit_t - \diamondsuit_t \;\ge\; \sqrt{\lambda_{\min}(L_t)}\,\|A_t\|_* - 2\sqrt{k}\,\|N_t\|_F,$$

which holds pathwise. Multiplying by c_a, taking conditional expectation (the first term on the right is $\mathcal{F}_t$-measurable), and applying Jensen's inequality $\mathbb{E}[\|N_t\|_F \mid \mathcal{F}_t] \le \sqrt{\mathbb{E}[\|N_t\|_F^2 \mid \mathcal{F}_t]} \le \sigma_{F,t}$ (Assumption (iii)) yields (51).

Step 2. We prove the identity

$$\begin{aligned}\mathbb{E}\big[\langle \nabla f(W_t),\, \alpha_{\mathrm{kl}}\,L_t^{-1/2} G_t U_t S_t^{-1/2} U_t^\top\rangle \;\big|\; \mathcal{F}_t\big] \;&=\; \alpha_{\mathrm{kl}}\,\|V_t\,S_t^{-1/4}\|_F^2 \\ &=\; \alpha_{\mathrm{kl}}\,\big\|L_t^{-1/4}\nabla f(W_t)\,U_t\,S_t^{-1/4}\big\|_F^2. \qquad (52)\end{aligned}$$

Since L_t, S_t, U_t are $\mathcal{F}_t$-measurable, the integrand is linear in G_t, and $\mathbb{E}[G_t \mid \mathcal{F}_t] = \nabla f(W_t)$, the conditional expectation of the left-hand side equals

$$\alpha_{\mathrm{kl}}\,\mathrm{Tr}\big(U_t^\top \nabla f(W_t)^\top\, L_t^{-1/2}\,\nabla f(W_t)\,U_t\cdot S_t^{-1/2}\big) \;=\; \alpha_{\mathrm{kl}}\,\mathrm{Tr}\big(V_t^\top V_t\cdot S_t^{-1/2}\big),$$

where we used $V_t^\top V_t = U_t^\top \nabla f(W_t)^\top L_t^{-1/2}\,\nabla f(W_t)\,U_t$ (since $L_t^{-1/4}\cdot L_t^{-1/4} = L_t^{-1/2}$). Writing $S_t^{-1/2} = S_t^{-1/4}S_t^{-1/4}$ and using trace cyclicity, $\mathrm{Tr}(V_t^\top V_t\cdot S_t^{-1/2}) = \mathrm{Tr}\big(S_t^{-1/4}V_t^\top V_t S_t^{-1/4}\big) = \|V_t S_t^{-1/4}\|_F^2$, proving (52).

Step 3. We prove the upper bound

$$\mathbb{E}\big[\|\Delta W_t\|_{\mathrm{op}}^2 \;\big|\; \mathcal{F}_t\big] \;\le\; 2\alpha_{\mathrm{kl}}^2\,\sigma_{kl}^2 + 2c_a^2. \qquad (53)$$

By the triangle inequality for operator norm and $(a+b)^2 \le 2a^2 + 2b^2$:

$$\|\Delta W_t\|_{\mathrm{op}}^2 \;\le\; 2\,\|\alpha_{\mathrm{kl}}\,L_t^{-1/2}G_tU_tS_t^{-1/2}U_t^\top\|_{\mathrm{op}}^2 \;+\; 2\,\|c_a\,\mathrm{polar}(L_t^{-1/2}G_tP_{\perp,t})\|_{\mathrm{op}}^2.$$

For the first term, $\|XU_t^\top\|_{\mathrm{op}} = \|X\|_{\mathrm{op}}$ since U_t has orthonormal columns, so it equals $2\alpha_{\mathrm{kl}}^2\,\|L_t^{-1/2}G_tU_tS_t^{-1/2}\|_{\mathrm{op}}^2$. For the second term, $\|\mathrm{polar}(\cdot)\|_{\mathrm{op}} \le 1$ deterministically gives $2c_a^2$. Taking conditional expectation and applying $\mathbb{E}[\|L_t^{-1/2}G_tU_tS_t^{-1/2}\|_{\mathrm{op}}^2 \mid \mathcal{F}_t] \le \sigma_{kl}^2$ a.s. proves (53).

Step 4. Take conditional expectation of both sides of (50). The inner product $\langle \nabla f(W_t), \Delta W_t\rangle$ splits according to (7); applying (51) (with sign flipped) and (52) (with sign flipped):

$$\begin{aligned}\eta\,\mathbb{E}\big[\langle \nabla f(W_t),\,\Delta W_t\rangle \;\big|\; \mathcal{F}_t\big] \;&\le\; -\eta\,\alpha_{\mathrm{kl}}\,\|V_t\,S_t^{-1/4}\|_F^2 \;-\; \eta\,c_a\sqrt{\lambda_{\min}(L_t)}\,\|A_t\|_* \;+\; 2\eta\,c_a\sqrt{k}\,\sigma_{F,t} \\ &=\; -\eta\,M_t \;+\; 2\eta\,c_a\sqrt{k}\,\sigma_{F,t}.\end{aligned}$$

Applying (53) to the second-order piece:

$$\tfrac{\eta^2 L_{\mathrm{op}}}{2}\,\mathbb{E}\big[\|\Delta W_t\|_{\mathrm{op}}^2 \;\big|\; \mathcal{F}_t\big] \;\le\; \eta^2 L_{\mathrm{op}}\,(\alpha_{\mathrm{kl}}^2\,\sigma_{kl}^2 + c_a^2).$$

Substituting into (50) and recalling $A_t = L_t^{-1/2}\nabla f(W_t)\,P_{\perp,t}$ and $V_t = L_t^{-1/4}\nabla f(W_t)\,U_t$ yields (49). □

J.1.3 SCALING INEQUALITIES

Lemma 4 (Scaling inequalities). *Let $L \in \mathbb{R}^{m\times m}$ and $S \in \mathbb{R}^{r\times r}$ be SPD. For any $X \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{m\times r}$:*

(i) $\|L^{-1/2}X\|_* \geq \lambda_{\max}(L)^{-1/2}\,\|X\|_*$.

(ii) $\|L^{-1/4}\,Y\,S^{-1/4}\|_F^2 \geq \big(\lambda_{\max}(L)\,\lambda_{\max}(S)\big)^{-1/2}\,\|Y\|_F^2$.

Proof. (i) The key step is sub-multiplicativity of the nuclear norm against the operator norm: $\|MN\|_* \leq \|M\|_{\mathrm{op}}\,\|N\|_*$ for any matrices M, N. Apply this to $X = L^{1/2}\cdot L^{-1/2}X$:

$$\|X\|_* \;\leq\; \|L^{1/2}\|_{\mathrm{op}}\,\|L^{-1/2}X\|_* \;=\; \sqrt{\lambda_{\max}(L)}\,\|L^{-1/2}X\|_*.$$

Dividing by $\sqrt{\lambda_{\max}(L)}$ gives (i).

(ii) The key step is sub-multiplicativity of the Frobenius norm: $\|MN\|_F \leq \|M\|_{\mathrm{op}}\,\|N\|_F$ and $\|MN\|_F \leq \|M\|_F\,\|N\|_{\mathrm{op}}$ for any matrices M, N. Apply both to $Y = L^{1/4}\cdot\big(L^{-1/4}\,Y\,S^{-1/4}\big)\cdot S^{1/4}$:

$$\|Y\|_F \;\leq\; \|L^{1/4}\|_{\mathrm{op}}\,\|L^{-1/4}YS^{-1/4}\|_F\,\|S^{1/4}\|_{\mathrm{op}} \;=\; \lambda_{\max}(L)^{1/4}\lambda_{\max}(S)^{1/4}\,\|L^{-1/4}YS^{-1/4}\|_F.$$

Squaring and dividing by $\sqrt{\lambda_{\max}(L)\,\lambda_{\max}(S)}$ gives (ii). □

J.1.4 PROOF OF THEOREM 1

Proof of Theorem 1. From Proposition 4, taking total expectation and writing $K := c_a^2 + \alpha_{\mathrm{kl}}^2\,\sigma_{kl}^2$:

$$\mathbb{E}[f(W_{t+1}) - f(W_t)] \;\leq\; -\eta\,\mathbb{E}[M_t] + 2\eta c_a\sqrt{k}\,\sigma_{F,t} + \eta^2 L_{\mathrm{op}}K. \tag{54}$$

No worst-case relaxation of the spectral factors is applied here; the relaxed form (47) is derived after the theorem.

Telescoping. Summing over $t = 0, \ldots, T-1$, the left-hand side telescopes to $\mathbb{E}[f(W_T) - f(W_0)] \geq -\Delta_0$ (Assumption (i)). Rearranging and dividing by ηT:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[M_t] \;\leq\; \tfrac{\Delta_0}{\eta T} + 2c_a\sqrt{k}\,\frac{1}{T}\sum_{t=0}^{T-1}\sigma_{F,t} + \eta L_{\mathrm{op}}K.$$

Optimizing the step size. The η-dependent terms on the right are $\Delta_0/(\eta T) + \eta L_{\mathrm{op}}K$. Their product $\Delta_0 L_{\mathrm{op}}K/T$ is independent of η. The sum is minimized at $\eta = \sqrt{\Delta_0/(TL_{\mathrm{op}}K)}$, giving $\Delta_0/(\eta T) = \eta L_{\mathrm{op}}K = \sqrt{\Delta_0 L_{\mathrm{op}}K/T}$. Substituting yields (14). If $\Delta_0 = 0$, this choice gives $\eta = 0$; then $W_t = W_0$ is a global minimizer, $M_t = 0$, and (14) holds trivially. □

J.1.5 PROOF OF LEMMA 3

Proof of Lemma 3. ($\Leftarrow$) trivial. ($\Rightarrow$) both summands are non-negative, so both vanish. Since $c_a\sqrt{\lambda_{\min}(L)} > 0$ and $\alpha_{\mathrm{kl}} > 0$, $\|L^{-1/2}\nabla f(W)\,P_\perp\|_* = 0$ gives $L^{-1/2}\nabla f(W)\,P_\perp = 0$, hence $\nabla f(W)\,P_\perp = 0$ because $L^{-1/2}$ is invertible; and $\|L^{-1/4}\nabla f(W)\,U\,S^{-1/4}\|_F = 0$ gives $\nabla f(W)\,U = 0$ because $L^{-1/4}$ and $S^{-1/4}$ are invertible. Therefore $\nabla f(W) = \nabla f(W)(UU^\top + P_\perp) = \nabla f(W)\,U\,U^\top + \nabla f(W)\,P_\perp = 0$. □

J.1.6 INTERPRETING σ_{kl}^2

σ_{kl}^2 admits a worst-case upper bound and a stationary reference value for the integrand:

- **Worst case (via Assumption (v) and Assumption (iv)).** Submultiplicativity gives

$$\|L^{-1/2}GUS^{-1/2}\|_{\mathrm{op}} \;\leq\; \|L^{-1/2}\|_{\mathrm{op}}\,\|G\|_{\mathrm{op}}\,\|S^{-1/2}\|_{\mathrm{op}} \;\leq\; C\cdot G_{\max}\cdot C \;=\; C^2G_{\max},$$

  hence one may choose $\sigma_{kl}^2 \leq C^4G_{\max}^2$.

- **At restricted KL stationarity (Claim 2).** For any subspace U, the S-stationarity condition (8) reads $S^* = \frac{1}{m}\mathbb{E}[U^\top G^\top {L^*_{\mathrm{restr}}}^{-1} G\, U]$, where S^* and L^*_{restr} are the corresponding optimal preconditioners for that U. Evaluating σ^2_{kl} at $(L_t, S_t) = (L^*_{\mathrm{restr}}, S^*)$ for the algorithm's current $U_t = U$:

$$\mathbb{E}\Big[\|{L^*_{\mathrm{restr}}}^{-1/2} G U\, {S^*}^{-1/2}\|^2_{\mathrm{op}}\Big] \;\le\; \mathbb{E}\Big[\|{L^*_{\mathrm{restr}}}^{-1/2} G U\, {S^*}^{-1/2}\|^2_{F}\Big] \;=\; \mathrm{Tr}({S^*}^{-1} \cdot mS^*) \;=\; mr,$$

  where the inequality uses $\|X\|^2_{\mathrm{op}} \le \|X\|^2_F$, and the equality uses trace cyclicity together with the S-stationarity. Thus mr is a *stationary reference value* for the integrand—independent of C—rather than a bound on the constant σ^2_{kl}, which must dominate the integrand at every t. In our practical algorithm (§K), the inverse square roots of each preconditioner's eigenvalues are clipped at a dimension-aware ceiling of the form $\max(10, \min(\dim, 4000))$; uniformly across all preconditioners, C is therefore of order $\max(m, n)$ for the layer dimensions of interest. The worst-case bound $C^4 G^2_{\max}$ thus scales as $\max(m, n)^4$, while the stationary reference value mr is linear in m.

Because the EMA updates of Algorithm 1 have the form of KL-Shampoo's stochastic proximal-gradient steps (Lin et al., 2025), with the fixed point of Claim 2 (§3.1), we expect $\sigma^2_{kl} = O(mr)$ whenever the algorithm's state is close to the restricted KL stationary point; a quantitative bound along the trajectory requires a tracking analysis of the EMA and is left to future work.

J.1.7 Noise floor and rate asymmetry

The left-hand side of (14) is *state-dependent*: it reads the gradient through the algorithm's own subspace decomposition $(U_t, P_{\perp,t})$. It is nevertheless a stationarity measure: it vanishes only at stationary points (Lemma 3), and (46)–(47) relate it to norms of the gradient. The noise floor $2c_a\sqrt{k}\,\frac{1}{T}\sum_{t<T}\sigma_{F,t}$ is linear in the noise level, a feature of the sign-SGD–style descent step (Bernstein et al., 2018) for the polar update (Proposition 4); this differs from the quadratic-in-σ_F floor in analyses where the update is a positive-definite linear function of the gradient (cf. Theorem 2 below, which has no floor after step-size balancing). The Smok-Hop bound (Theorem 2) uses Frobenius smoothness and a second-moment bound, so the two bounds are not directly comparable.

Conversion to Frobenius-norm stationarity is given by Corollary 2, whose constant uses the trajectory bounds $\sqrt{K_L}$ and $\sqrt{B_{LS}}$ in place of the worst-case caps $C\sqrt{\Theta}$ and Θ of (47), each of which scales at least as C^2; the mixed-norm bound (14) is the result we establish.

Rate asymmetry and noise floor. When the average noise level is $O(T^{-1/2})$, the complement's nuclear-norm stationarity converges at $O(T^{-1/2})$, while the subspace's Frobenius-norm stationarity converges at $O(T^{-1/4})$ (taking the square root of the squared-Frobenius term). This does not imply smaller r is preferable: the noise floor $2c_a\sqrt{k}\,\frac{1}{T}\sum_{t<T}\sigma_{F,t}$ grows as r shrinks when $n - r < m$ ($k = \min(m, n-r)$), offsetting the faster complement rate. In this momentum-free analysis the bound retains a noise-dependent term arising from the nonlinearity of the polar map; we do not show that it is unavoidable. The subspace update, being linear in the gradient, contributes no such noise term to the bound (cf. Theorem 2). For a fixed noise level $\sigma_{F,t} \equiv \sigma_F > 0$, Theorem 1 therefore bounds the averaged measure only up to a noise-dependent neighborhood of stationarity: because the polar map is nonlinear, unbiasedness of G_t does not imply unbiasedness of the update direction, and the resulting term is analogous to that in signSGD-style analyses (Bernstein et al., 2018). The time-varying form of Assumption (iii) characterizes when the floor vanishes. The term $2c_a\sqrt{k}\,\frac{1}{T}\sum_{t<T}\sigma_{F,t}$ vanishes whenever the average noise level tends to zero; in particular, with exact gradients, the averaged M_t-stationarity measure converges at rate $O(T^{-1/2})$. Under conditionally independent mini-batch sampling with per-sample variance bounded by ν^2, a batch of size B_t satisfies $\sigma_{F,t} \le \nu/\sqrt{B_t}$. A fixed batch therefore leaves an $O(B^{-1/2})$ term in the bound, matching the batch-size dependence in signSGD (Bernstein et al., 2018), whereas a linearly growing batch $B_t \propto t + 1$ yields an overall $O(T^{-1/2})$ rate (the batch-growth route taken for Muon by Nagashima & Iiduka, 2026). These vanishing-noise statements presuppose that Assumption (v) holds along the whole run; since the EMA statistics of Algorithm 1 shrink with the gradients, this requires the eigenvalue lower bounds to be enforced by clipping, as in the clipped variant of Algorithm 1 (remark in Appendix J.1). Theorem 1 analyzes the momentum-free update; in this setting the bound does not exclude a noise

floor. Recent Muon analyses remove analogous fixed-batch terms using scheduled momentum or variance reduction (Shen et al., 2025; Chang et al., 2025; Sfyraki & Wang, 2025). We leave extending these mechanisms to Pro-KLShampoo to future work.

J.2 Technical companion: Smok-Hop

Smok-Hop (Eq. (6)) replaces the polar orthogonalization in Step 1 of Algorithm 1 with the scalar scaling $\mu_\perp^{-1/2} L^{-1/2} G_\perp$, giving the update direction $\Delta W = L^{-1/2} G \hat{R}^{-1/2}$. This is a linear function of G and admits a simpler analysis than Pro-KLShampoo's polar update. We include the convergence result here for comparison; it is *not* the algorithm used in practice.

This subsection replaces operator-norm smoothness with Frobenius smoothness, which is implied by Assumption (ii) with $L_F = L_{\mathrm{op}}$:

(ii)′ L_F-Frobenius smoothness: $f(W') \le f(W) + \langle \nabla f(W), W' - W \rangle + \frac{L_F}{2}\|W' - W\|_F^2$.

Theorem 2 (Convergence of Smok-Hop). *Let σ_P^2 be a deterministic constant, independent of η_0, that bounds the conditional second moment (in Frobenius norm) of the preconditioned stochastic gradient:*

$$\mathbb{E}\Big[\big\|L_t^{-1/2} G_t \hat{R}_t^{-1/2}\big\|_F^2 \;\big|\; \mathcal{F}_t\Big] \;\le\; \sigma_P^2 \qquad \text{a.s. for all } t \ge 0, \tag{55}$$

where $\hat{R}_t = U_t S_t U_t^\top + \mu_{\perp,t} P_{\perp,t}$ is the spike-and-flat right-side preconditioner at step t. Under Assumptions (i), (iii), (iv), (v) and (ii)′ above, with Assumption (v) imposed along the Smok-Hop trajectory, for any $T \ge 1$ and step size $\eta = \eta_0/\sqrt{T}$ with $\eta_0 > 0$:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(W_t)\|_F^2 \;\le\; \frac{\Theta\,\Delta_0}{\eta_0\sqrt{T}} \;+\; \frac{\eta_0\, L_F\, \Theta\, \sigma_P^2}{2\sqrt{T}}, \tag{56}$$

where $\Delta_0 \coloneqq f(W_0) - f^*$. The balanced choice $\eta_0^\star = \sqrt{2\Delta_0/(L_F\sigma_P^2)}$, for $\Delta_0 > 0$, gives the rate

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(W_t)\|_F^2 \;=\; O\!\left(\Theta\sqrt{\frac{L_F \Delta_0 \sigma_P^2}{T}}\right).$$

The constant C of Assumption (v) enters the rate through Θ (Lemma 2) and through σ_P^2 (see §J.2.2 for reference values).

J.2.1 Proof of Theorem 2

Throughout this subsubsection, the scalar-complement update direction at step t is

$$\Delta W_t \;\coloneqq\; L_t^{-1/2} G_t \hat{R}_t^{-1/2}, \qquad \hat{R}_t \coloneqq U_t S_t U_t^\top + \mu_{\perp,t} P_{\perp,t},$$

so that $W_{t+1} = W_t - \eta\,\Delta W_t$. (This is the scalar-complement update; the Pro-KLShampoo update direction in §J.1 is different.)

Linear structure. The map $G_t \mapsto \Delta W_t$ is linear in G_t; its vec form is multiplication by the matrix $P_t \coloneqq L_t^{-1/2} \otimes \hat{R}_t^{-1/2}$, which is symmetric positive definite with eigenvalues in $[\Theta^{-1}, C^2]$ (the lower bound is from Lemma 2, the upper bound from Assumption (v)). The lower bound on $\lambda_{\min}(P_t)$ implies, for any matrix M of the same shape as G_t,

$$\langle M,\, L_t^{-1/2} M \hat{R}_t^{-1/2} \rangle \;=\; \langle \mathrm{vec}(M),\, P_t\,\mathrm{vec}(M) \rangle \;\ge\; \Theta^{-1}\|M\|_F^2. \tag{57}$$

Single-step descent. By Frobenius smoothness (ii)′, with $W_{t+1} - W_t = -\eta\,\Delta W_t$:

$$f(W_{t+1}) \;\le\; f(W_t) - \eta\,\langle \nabla f(W_t),\, \Delta W_t \rangle + \tfrac{\eta^2 L_F}{2}\,\|\Delta W_t\|_F^2.$$

Take conditional expectation given $\mathcal{F}_t$ (so $L_t, S_t, U_t, \mu_{\perp,t}$ are measurable; only G_t is random). Since ΔW_t is linear in G_t and $\mathbb{E}[G_t \mid \mathcal{F}_t] = \nabla f(W_t)$ (Assumption (iii)),

$$\mathbb{E}[\Delta W_t \mid \mathcal{F}_t] \;=\; L_t^{-1/2}\,\nabla f(W_t)\,\hat{R}_t^{-1/2}.$$

Apply (57) with $M = \nabla f(W_t)$ to the first-order term, and bound the second-order term by $\mathbb{E}[\|\Delta W_t\|_F^2 \mid \mathcal{F}_t] \le \sigma_P^2$ from (55):

$$\mathbb{E}[f(W_{t+1}) \mid \mathcal{F}_t] \;\le\; f(W_t) - \tfrac{\eta}{\Theta}\,\|\nabla f(W_t)\|_F^2 + \tfrac{\eta^2 L_F}{2}\,\sigma_P^2. \tag{58}$$

Telescoping. Taking total expectation, summing over $t = 0, \dots, T-1$, and using $\mathbb{E}[f(W_T)] \ge f^*$:

$$\tfrac{\eta}{\Theta} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(W_t)\|_F^2 \;\le\; \Delta_0 + \tfrac{T\eta^2 L_F \sigma_P^2}{2}.$$

Substituting $\eta = \eta_0/\sqrt{T}$ and dividing by $T\eta/\Theta = \eta_0\sqrt{T}/\Theta$:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(W_t)\|_F^2 \;\le\; \frac{\Theta\Delta_0}{\eta_0\sqrt{T}} + \frac{\eta_0 L_F\,\Theta\,\sigma_P^2}{2\sqrt{T}},$$

which is (56). The balanced choice $\eta_0^\star = \sqrt{2\Delta_0/(L_F\sigma_P^2)}$ minimizes the right-hand side, giving the rate $\Theta\sqrt{2L_F\Delta_0\sigma_P^2/T}$. □

J.2.2 Interpreting σ_P^2

From its definition (55), σ_P^2 bounds the conditional second moment (in Frobenius norm) of the preconditioned stochastic gradient. We give two reference values:

- **Worst case (Assumption (v)).** The spectral bounds $\|L_t^{-1/2}\|_{\mathrm{op}}, \|\hat{R}_t^{-1/2}\|_{\mathrm{op}} \le C$ give $\|L_t^{-1/2} G_t \hat{R}_t^{-1/2}\|_F^2 \le C^4\|G_t\|_F^2$, and Assumptions (iv) and (iii) imply $\mathbb{E}\|G_t\|_F^2 \le \sigma_F^2 + \min(m,n)\,G_{\max}^2$ (with $\sigma_F = \sup_t \sigma_{F,t}$). Hence one may choose $\sigma_P^2 \le C^4(\sigma_F^2 + \min(m,n)G_{\max}^2)$.
- **Restricted KL stationarity (Claim 2).** At the restricted KL fixed point $(L^*_{\mathrm{restr}}, \hat{R}^*)$, the L-stationarity condition (10) reads $L^*_{\mathrm{restr}} = \frac{1}{n}\,\mathbb{E}[G\,\hat{R}^{*-1}\,G^\top]$. By trace cyclicity,

  $$\mathbb{E}\Big[\big\|{L^*_{\mathrm{restr}}}^{-1/2}\,G\,\hat{R}^{*-1/2}\big\|_F^2\Big] \;=\; \mathrm{Tr}\big({L^*_{\mathrm{restr}}}^{-1}\,\mathbb{E}[G\,\hat{R}^{*-1}\,G^\top]\big) \;=\; n\,\mathrm{Tr}(I_m) \;=\; mn. \tag{59}$$

  Hence the integrand of (55) has expectation exactly mn at the restricted KL fixed point—independent of C.

Because the EMA updates of Algorithm 1 have the form of KL-Shampoo's stochastic proximal-gradient steps (Lin et al., 2025), with the fixed point of Claim 2 (§3.1), we expect $\sigma_P^2 = O(mn)$ whenever the algorithm's state is close to the restricted KL stationary point. A quantitative bound on σ_P^2 along the entire trajectory requires a tracking analysis of the EMA and is left to future work.

J.3 Calibration of the mixing weight α_{kl}

This subsection derives the closed-form magnitude estimate of α_{kl} used in §3.3. The principle is to choose α_{kl} so that the relative scale between the subspace and complement components of the Pro-KLShampoo update (7) matches that of the scalar-complement update (6) at the restricted KL stationary state. This way, replacing the scalar-complement whitening on the complement with polar does not, in expectation, alter the relative scale between subspace and complement prescribed by the restricted-KL solution. Operator norm is the natural matching norm: the polar map yields a partial isometry, so its output has a deterministic operator norm. Claim 2 determines $\mathbb{E}\|\Delta_{\mathrm{res}}^{\mathrm{naive}}\|_F^2$ but not $\mathbb{E}\|\Delta_{\mathrm{res}}^{\mathrm{naive}}\|_{\mathrm{op}}^2$; combining this with the bound $\mathrm{rank}(\Delta_{\mathrm{res}}^{\mathrm{naive}}) \le k$ a.s. gives a bracketed range on α^*_{kl}.

Update components at the stationary state. At the restricted KL stationary state, write the subspace component (shared by both updates) and the two complement components as

$$\begin{aligned}
\Delta_{\mathrm{kl}} &\coloneqq (L^*_{\mathrm{restr}})^{-1/2}\,G\,U^*(S^*)^{-1/2}\,U^{*\top}, \\
\Delta_{\mathrm{res}}^{\mathrm{naive}} &\coloneqq (\mu^*_\perp)^{-1/2}\,(L^*_{\mathrm{restr}})^{-1/2}\,G_\perp, && \text{(scalar-complement, scalar-whitened)} \\
\Delta_{\mathrm{res}}^{\mathrm{polar}} &\coloneqq c_a\,\mathrm{polar}\big((L^*_{\mathrm{restr}})^{-1/2}\,G_\perp\big), && \text{(Pro-KLShampoo, orthogonalized)}
\end{aligned}$$

with $c_a = \sqrt{\max(1, m/n)}$. The scalar-complement update direction is $\Delta_{\text{kl}} + \Delta_{\text{res}}^{\text{naive}}$; the Pro-KLShampoo update direction is $\alpha_{\text{kl}}\,\Delta_{\text{kl}} + \Delta_{\text{res}}^{\text{polar}}$. Since Δ_{kl} is identical in both, the calibration only needs to match the complement components in scale (under a chosen unitarily invariant norm; in expectation where applicable).

Polar complement: deterministic norms. The polar map produces a partial isometry: $\text{polar}(X) = U_X V_X^\top$ where $X = U_X \Sigma_X V_X^\top$ is the compact SVD. Write $\bar{m} \coloneqq \min(m, n)$ and $\bar{n} \coloneqq \max(m, n)$. The input to polar is $(L_{\text{restr}}^*)^{-1/2} G_\perp \in \mathbb{R}^{m \times n}$ in the right projection case ($m \le n$) and $G_\perp\,(R_{\text{restr}}^*)^{-1/2} \in \mathbb{R}^{m \times n}$ in the left projection case ($m > n$, Appendix K); under any absolutely continuous gradient distribution it has rank

$$k \;\coloneqq\; \min(\bar{m},\, \bar{n} - r) \;=\; \begin{cases} \min(m,\, n - r), & m \le n, \\ \min(m - r,\, n), & m > n, \end{cases}$$

almost surely. Hence both norms of the polar update are deterministic:

$$\|\Delta_{\text{res}}^{\text{polar}}\|_F \;=\; c_a\sqrt{k}, \qquad \|\Delta_{\text{res}}^{\text{polar}}\|_{\text{op}} \;=\; c_a, \tag{60}$$

where $c_a = \sqrt{\max(1, m/n)}$ is computed from the original matrix dimensions (matching the implementation).

Scalar-complement: norms in expectation. For the scalar-complement,

$$\|\Delta_{\text{res}}^{\text{naive}}\|_F^2 \;=\; (\mu_\perp^*)^{-1}\,\text{Tr}\big(G_\perp^\top (L_{\text{restr}}^*)^{-1}\, G_\perp\big).$$

At the restricted KL stationary state, $\mu_\perp$-stationarity (9) gives $\text{Tr}\big(\mathbb{E}[G_\perp^\top (L_{\text{restr}}^*)^{-1} G_\perp]\big) = m(n - r)\,\mu_\perp^*$, hence $\mathbb{E}\|\Delta_{\text{res}}^{\text{naive}}\|_F^2 = m(n - r)$ for $m \le n$. The left projection case ($m > n$) is symmetric under $L \leftrightarrow R$ and gives $n(m - r)$. Uniformly,

$$\sqrt{\mathbb{E}\|\Delta_{\text{res}}^{\text{naive}}\|_F^2} \;=\; \sqrt{\bar{m}(\bar{n} - r)}. \tag{61}$$

For the operator norm, since $\text{rank}(\Delta_{\text{res}}^{\text{naive}}) \le k$ a.s., the inequalities $\|X\|_F^2/\text{rank}(X) \le \|X\|_{\text{op}}^2 \le \|X\|_F^2$ give

$$\frac{\bar{m}(\bar{n} - r)}{k} \;\le\; \mathbb{E}\|\Delta_{\text{res}}^{\text{naive}}\|_{\text{op}}^2 \;\le\; \bar{m}(\bar{n} - r), \tag{62}$$

with the lower (resp. upper) bound attained when the singular values of $\Delta_{\text{res}}^{\text{naive}}$ are uniformly spread (resp. concentrated in one direction).

Matching condition: bracketed range. We choose α_{kl}^* so that the subspace-to-complement norm ratio is preserved across the two update rules, with Δ_{kl} shared:

$$\frac{\|\alpha_{\text{kl}}^*\,\Delta_{\text{kl}}\|_\diamond}{\|\Delta_{\text{res}}^{\text{polar}}\|_\diamond} \;=\; \frac{\|\Delta_{\text{kl}}\|_\diamond}{\sqrt{\mathbb{E}\|\Delta_{\text{res}}^{\text{naive}}\|_\diamond^2}} \quad \text{if and only if} \quad \alpha_{\text{kl}}^* \;=\; \frac{\|\Delta_{\text{res}}^{\text{polar}}\|_\diamond}{\sqrt{\mathbb{E}\|\Delta_{\text{res}}^{\text{naive}}\|_\diamond^2}},$$

where $\|\cdot\|_\diamond$ is a unitarily invariant norm.

The natural matching norm is the operator norm: by (60), $\|\Delta_{\text{res}}^{\text{polar}}\|_{\text{op}} = c_a$ deterministically, regardless of the gradient's singular value distribution. Under operator-norm matching,

$$\alpha_{\text{kl}}^* \;=\; \frac{c_a}{\sqrt{\mathbb{E}\|\Delta_{\text{res}}^{\text{naive}}\|_{\text{op}}^2}}.$$

Claim 2 determines $\mathbb{E}\|\Delta_{\text{res}}^{\text{naive}}\|_F^2$ but not $\mathbb{E}\|\Delta_{\text{res}}^{\text{naive}}\|_{\text{op}}^2$. Substituting the bracket (62) into the operator-norm matching formula gives

$$\boxed{\frac{c_a}{\sqrt{\bar{m}(\bar{n} - r)}} \;\le\; \alpha_{\text{kl}}^* \;\le\; \frac{c_a\sqrt{k}}{\sqrt{\bar{m}(\bar{n} - r)}}.} \tag{63}$$

The lower bound corresponds to a maximally concentrated scalar-complement update (Frobenius mass in a single direction; op-norm equals Frobenius); the upper bound corresponds to a uniformly spread one (op-norm equals Frobenius$/\sqrt{k}$). The upper bound coincides with the simpler *Frobenius matching* formula $\alpha_{\text{kl}}^* = c_a\sqrt{k}/\sqrt{\bar{m}(\bar{n} - r)}$, recovered from (60)–(61) by setting $\|\cdot\|_\diamond = \|\cdot\|_F$. The bracket has multiplicative width $\sqrt{k}$. For square layers ($m = n$, so $\bar{m} = \bar{n} = n$, $c_a = 1$, $k = n - r$), it reduces to $1/\sqrt{n(n - r)} \le \alpha_{\text{kl}}^* \le 1/\sqrt{n}$.

Numerical values. Evaluating the bracket (63) at $r = 128$ for the four model configurations used in our experiments:

Layer	m	n	r	c_a	k	bracket on α^*_{kl}
GPT-2 124M (d=768, MLP intermediate $4d$, 12 layers)						
Attention QKV / out	768	768	128	1	640	$[0.0014,\ 0.036]$
MLP up ($4d \times d$)	3072	768	128	2	768	$[0.0013,\ 0.037]$
MLP down ($d \times 4d$)	768	3072	128	1	768	$[0.00067,\ 0.018]$
GPT-medium 350M (d=1024, MLP intermediate $4d$, 24 layers)						
Attention QKV / out	1024	1024	128	1	896	$[0.0010,\ 0.031]$
MLP up ($4d \times d$)	4096	1024	128	2	1024	$[0.0010,\ 0.032]$
MLP down ($d \times 4d$)	1024	4096	128	1	1024	$[0.00050,\ 0.016]$
LLaMA-134M (d=768, MLP intermediate e=2048, 12 layers, SwiGLU)						
Attention Q/K/V/O	768	768	128	1	640	$[0.0014,\ 0.036]$
MLP gate / up ($e \times d$)	2048	768	128	$\sqrt{8/3}$	768	$[0.0013,\ 0.037]$
MLP down ($d \times e$)	768	2048	128	1	768	$[0.00082,\ 0.023]$
LLaMA-450M (d=1024, MLP intermediate e=2816, 30 layers, SwiGLU)						
Attention Q/K/V/O	1024	1024	128	1	896	$[0.0010,\ 0.031]$
MLP gate / up ($e \times d$)	2816	1024	128	$\sqrt{11/4}$	1024	$[0.0010,\ 0.032]$
MLP down ($d \times e$)	1024	2816	128	1	1024	$[0.00060,\ 0.019]$

In our experiments we sweep $\alpha_{\mathrm{kl}} \in \{0.005,\ 0.01,\ 0.015\}$ as a single value shared across layers (§3.3). Intersecting the per-layer brackets in the table gives the all-layer-feasible interval $[\max_i \mathrm{lower}_i,\ \min_i \mathrm{upper}_i] \approx [1.4 \times 10^{-3},\ 1.6 \times 10^{-2}]$; the swept values $\{0.005, 0.01, 0.015\}$ all sit inside this intersection. A sweep over this interval on GPT-2 350M is reported in Appendix O. The best swept value lies between 0.005 and 0.01, varying across the four model configurations. Pinning α^*_{kl} to a point would require characterizing the singular value concentration of $\Delta^{\mathrm{naive}}_{\mathrm{res}}$ along the trajectory, which we leave to future work.

K PRACTICAL ALGORITHM

Algorithm 2 below is the implementation form used in our experiments. It follows Algorithm 1 step for step, with three practical additions in the parameter update, namely (P1) Nesterov momentum, (P2) damping and clipping of the inverse square roots, and (P3) Newton–Schulz approximation of the polar factor. It also adds decoupled weight decay and one change in the spectral layer, (P4): the exact eigendecompositions of Step 4 of Algorithm 1 are replaced by maintained spectral estimates, i.e., decoupled eigenvalue EMAs with clipping as in KL-Shampoo (Lin et al., 2025), an eigenbasis transported with the subspace, and a periodic QR refresh. We present the case $m \le n$; $m > n$ is symmetric (see the remark on left projection below). For numerical stability under bfloat16 we store $\lambda_L^{\odot -1/2}, \lambda_S^{\odot -1/2}, \mu_\perp^{-1/2}$ rather than $\lambda_L, \lambda_S, \mu_\perp$ (Lin et al., 2025).

Algorithm 2 Pro-KLShampoo (practical; one step $t \to t+1$, $m \le n$)

State at step t: $(W_t, H_t, L_t, S_t, U_t, \mu_{\perp,t})$, $(Q_{L,t}, \lambda_{L,t})$, $(Q_{S,t}, \lambda_{S,t})$ (spectral estimates of L_t, S_t).
Write $\hat{L}_t^{-p} := Q_{L,t}\mathrm{Diag}(\lambda_{L,t}^{\odot -p})Q_{L,t}^\top$ and $\hat{S}_t^{-p} := Q_{S,t}\mathrm{Diag}(\lambda_{S,t}^{\odot -p})Q_{S,t}^\top$ for $p \in \{1/2, 1\}$.
Set $P_{\perp,t} = I_n - U_t U_t^\top$.

1: Parameter update (P1–P3)

$H_{t+1} = \mu\, H_t + G_t, \quad \hat{G}_t = G_t + \mu\, H_{t+1}, \quad \hat{G}_{\parallel,t} = \hat{G}_t\, U_t, \quad \hat{G}_{\perp,t} = \hat{G}_t\, P_{\perp,t}$

$\Delta_{\mathrm{sub},t} = \hat{L}_t^{-1/2}\, \hat{G}_{\parallel,t}\, \hat{S}_t^{-1/2}\, U_t^\top, \quad \Delta_{\mathrm{res},t} = c_a\, \mathrm{NS}\big(\hat{L}_t^{-1/2}\, \hat{G}_{\perp,t}\big), \quad c_a = \sqrt{\max(1,\, m/n)}$

$W_{t+1} = (1 - \eta\, \lambda_{\mathrm{wd}})\, W_t - \eta\, \big(\Delta_{\mathrm{res},t} + \alpha_{\mathrm{kl}}\, \Delta_{\mathrm{sub},t}\big)$

2: Statistics update in the U_t coordinates (raw G_t)

$\widetilde{G}_t = G_t U_t, \quad G_{\perp,t} = G_t P_{\perp,t}, \quad \Delta L_t = \frac{1}{n}\big(\widetilde{G}_t\, \hat{S}_t^{-1}\, \widetilde{G}_t^\top + \mu_{\perp,t}^{-1}\, G_{\perp,t} G_{\perp,t}^\top\big)$

$\Delta S_t = \frac{1}{m}\, \widetilde{G}_t^\top\, \hat{L}_t^{-1}\, \widetilde{G}_t, \quad \delta_{\perp,t} = \frac{1}{m(n-r)} \big\| \hat{L}_t^{-1/2} G_{\perp,t} \big\|_F^2$

$L_{t+1} = \beta_2 L_t + (1-\beta_2)\Delta L_t, \quad \bar{S}_{t+1} = \beta_2 S_t + (1-\beta_2)\Delta S_t, \quad \mu_{\perp,t+1} = \beta_2\, \mu_{\perp,t} + (1-\beta_2)\, \delta_{\perp,t}$

Eigenvalue EMA: $\lambda_{L,t+1} = \beta_2 \lambda_{L,t} + (1-\beta_2)\, \mathrm{diag}(Q_{L,t}^\top \Delta L_t\, Q_{L,t})$

$\lambda_{S,t+1} = \beta_2 \lambda_{S,t} + (1-\beta_2)\, \mathrm{diag}(Q_{S,t}^\top \Delta S_t\, Q_{S,t})$

Clipping (64): $\lambda_{L,t+1}^{\odot -1/2} \leftarrow \min(\lambda_{L,t+1}^{\odot -1/2}, C_L)$; likewise for $\lambda_{S,t+1}^{\odot -1/2}$ (C_S) and $\mu_{\perp,t+1}^{-1/2}$ (C_μ)

3: Subspace and projected-state update

$U_{t+1} = \mathrm{qr}\big(\beta_2\, U_t S_t + \frac{1-\beta_2}{m}\, G_t^\top\, \hat{L}_t^{-1}\, G_t\, U_t\big), \quad T_t = U_t^\top U_{t+1}$

$S_{t+1} = T_t^\top \bar{S}_{t+1} T_t, \quad Q_{S,t+1} = T_t^\top Q_{S,t}$ (transported eigenbasis), $\quad Q_{L,t+1} = Q_{L,t}$

4: Periodic eigenbasis refresh (every τ steps, replacing Step 4 of Algorithm 1)

$Q_{L,t+1} \leftarrow \mathrm{qr}(L_{t+1}\, Q_{L,t+1}), \quad Q_{S,t+1} \leftarrow \mathrm{qr}(S_{t+1}\, Q_{S,t+1})$

State. Per parameter $W \in \mathbb{R}^{m\times n}$: momentum buffer H, subspace basis $U \in \mathbb{R}^{n\times r}$ with $U^\top U = I_r$, left factor $L \in \mathbb{R}^{m\times m}$ and subspace factor $S \in \mathbb{R}^{r\times r}$, both symmetric positive semidefinite, complement scalar $\mu_\perp > 0$, and maintained spectral estimates (Q_L, λ_L) of L and (Q_S, λ_S) of S. These estimates take the place of the exact eigendecompositions of Algorithm 1: the eigenvalues λ_L, λ_S are updated by EMA (Step 2), Q_S is transported with the subspace (Step 3), and both eigenbases are refreshed by the periodic QR of Step 4. The update direction uses $\hat{L}^{-1/2} := Q_L\, \mathrm{Diag}(\lambda_L^{\odot -1/2})\, Q_L^\top$ and $\hat{S}^{-1/2} := Q_S\, \mathrm{Diag}(\lambda_S^{\odot -1/2})\, Q_S^\top$ in place of $L^{-1/2}$ and $S^{-1/2}$.

Hyperparameter defaults. $\mu = \beta_2 = 0.95$ (momentum and EMA weight), $\varepsilon = 10^{-8}$ (damping), $\alpha_{\mathrm{kl}} = 0.01$ (mixing weight), $\tau = 10$ (QR refresh period), $T_{\mathrm{NS}} = 5$ (Newton–Schulz iterations), $r \in \{32, 64, 128\}$ (subspace rank), $\lambda_0 = 0.1$ (initial eigenvalue scale). Learning rate η and weight decay λ_{wd} are swept per configuration.

Initialization. $H \leftarrow 0$; U is the top-r right singular vectors of the first observed gradient $G^{(0)}$; $L \leftarrow \frac{1-\beta_2}{r}\widetilde{G}\widetilde{G}^\top$, $S \leftarrow \frac{1-\beta_2}{m}\widetilde{G}^\top\widetilde{G}$ with $\widetilde{G} = G^{(0)}U$; $\lambda_L, \lambda_S, \mu_\perp \leftarrow \lambda_0$ (fixed initialization, since single-sample eigenvalues are noisy); Q_L, Q_S are the eigenvectors of L, S sorted by descending eigenvalue.

(P1) Nesterov momentum. The update direction uses the Nesterov-lookahead gradient $\hat{G} := G + \mu H$ in place of the raw gradient G. The EMAs of $L, S, \mu_\perp$ continue to use the raw G.

(P2) Damping and clipping. We apply $\hat{L}^{-1/2} = Q_L \operatorname{Diag}(\lambda_L^{\odot -1/2})\, Q_L^\top$ (similarly $\hat{S}^{-1/2}$). In $\Delta_{\mathrm{sub},t}$, each product $\lambda_{L,i}^{-1/2}\lambda_{S,j}^{-1/2}$ is damped to $1/(\sqrt{\lambda_{L,i}\lambda_{S,j}} + \varepsilon)$; the complement term uses the undamped $\hat{L}^{-1/2}$. All inverse square roots are additionally clipped from above by a dimension-aware ceiling (Eigenvalue clipping below). Algorithm 2 writes the clipping (Step 2) but not the damping.

Eigenvalue clipping. After every per-eigenvalue EMA update we clip the inverse-square-roots:

$$\lambda_{L,i}^{-1/2} \leftarrow \min\big(\lambda_{L,i}^{-1/2},\, C_L\big), \quad \lambda_{S,j}^{-1/2} \leftarrow \min\big(\lambda_{S,j}^{-1/2},\, C_S\big), \quad \mu_\perp^{-1/2} \leftarrow \min\big(\mu_\perp^{-1/2},\, C_\mu\big), \tag{64}$$

with dimension-aware ceilings $C_L = \max(10, \min(m, 4000))$, $C_S = \max(10, \min(n, 4000))$, $C_\mu = \max(10, \min(\max(m, n), 4000))$, in particular flooring λ at $1/C^2$ where $C := \max(C_L, C_S, C_\mu)$. The clip enforces, on the decoupled eigenvalue estimates that enter the update, the lower bound that Assumption (v) of Appendix J imposes on L_t, S_t, $\mu_{\perp,t}$, and provides numerical stability for the inverse-square-roots. It also restricts the scalar gauge: $L \otimes \hat{R}$ is invariant under $(L, \hat{R}) \mapsto (cL,\, c^{-1}\hat{R})$ for any $c > 0$, so the KL stationarity system determines only the Kronecker product and not the individual factors; clipping all three preconditioners simultaneously breaks this rescaling freedom. We use $C \in [768, 4000]$.

(P3) Newton–Schulz orthogonalization. The exact polar factor in Algorithm 1 is approximated by T_{NS} iterations of the Muon (Jordan et al., 2024) polynomial $X \mapsto a\,X + b\,XX^\top X + c\,(XX^\top)^2 X$ with $(a, b, c) = (3.4445, -4.7750, 2.0315)$, starting from $X/\|X\|_F$.

(P4) Maintained spectral estimates. Algorithm 1 eigendecomposes L_{t+1} and S_{t+1} at every step (its Step 4). Algorithm 2 instead maintains (Q_L, λ_L) and (Q_S, λ_S): the eigenvalues follow the EMA of the diagonal of the rotated targets $Q_L^\top \Delta L_t\, Q_L$ and $Q_S^\top \Delta S_t\, Q_S$ (Step 2), Q_S is transported with the subspace through T_t (Step 3), and both eigenbases are refreshed every τ steps by one power-iteration step with QR orthonormalization (Step 4). Between refreshes the transported Q_S is not exactly orthonormal. Steps 1–3 otherwise follow Algorithm 1, and Theorem 1 is stated for Algorithm 1.

Order of Steps 2 and 3. The statistics of Step 2 can also be formed after the transport of Step 3, in the U_{t+1} coordinates; this is the alternative order of the Remark in Appendix J.1, and our implementation provides it as an option. Since U changes by a single subspace-iteration step with EMA weight β_2 at every step, consecutive subspaces are typically close, and we expect the two orders to behave similarly.

Smok-Hop (6) is recovered by replacing Δ_{res} with

$$\mu_\perp^{-1/2}\, Q_L \operatorname{Diag}(\lambda_L^{\odot -1/2})\, Q_L^\top\, \hat{G}_\perp,$$

setting $\alpha_{\mathrm{kl}} = 1$ and using EMA momentum in place of Nesterov.

Remark (Left projection, $m > n$). When $m > n$, we project the left-side preconditioner L rather than the right-side R. This is implemented by transposing all roles: the r-dimensional subspace $U \in \mathbb{R}^{m \times r}$ acts on the left, $\widetilde{G} := U^\top G \in \mathbb{R}^{r \times n}$, complement $G_\perp := G - U\widetilde{G}$, and the orthogonalized update becomes $\Delta_{\mathrm{res}} = \mathrm{NS}(\hat{G}_\perp\, R^{-1/2}) \cdot \sqrt{\max(1,\, m/n)}$. Claims 1–3, Lemma 2, Proposition 4 and Theorem 1 carry over by applying them to $W^\top$, which exchanges the roles of m and n and of L and R; the proofs use c_a only as a positive constant, and the noise term has $k = \min(m - r, n)$.

L MEMORY AND COMPUTATIONAL COST

We compare the memory and per-step computational cost of Pro-KLShampoo and KL-Shampoo for one weight $W \in \mathbb{R}^{m\times n}$ with $m \le n$, Newton–Schulz iteration count T_{NS}, subspace rank $r \ll n$, and eigenbasis-refresh period τ (Table 4, in the row layout of Lin et al. (2025)). Memory rows give exact element counts; compute rows give leading-order matmul cost with constants suppressed.

Table 4: Memory and per-step computational cost for one weight $W \in \mathbb{R}^{m\times n}$, $m \le n$, $r \ll n$.

	KL-Shampoo	Pro-KLShampoo
Memory (per parameter)		
Kronecker / subspace factors (L, R vs L, S)	$m^2 + n^2$	$m^2 + r^2$
Eigenbases (Q_L, Q_R vs Q_L, Q_S)	$m^2 + n^2$	$m^2 + r^2$
Eigenvalues (λ_L, λ_R vs λ_L, λ_S)	$m + n$	$m + r$
Subspace basis U	N/A	nr
Complement scalar $\mu_\perp$	N/A	1
Momentum buffer	mn	mn
Compute (per step, leading order)		
Matmul (preconditioning + statistics)	mn^2	$T_{\mathrm{NS}}\, m^2 n + mnr$
Eigenbasis QR (amortized over τ steps)	$(m^3 + n^3)/\tau$	$(m^3 + r^3)/\tau$

Memory. The right-side $n \times n$ Kronecker factor R and its eigenbasis Q_R are replaced by the $r \times r$ pair (S, Q_S), plus the $n \times r$ subspace basis U and the scalar $\mu_\perp$; for $r \ll n$ the right-side state shrinks by a factor of about $2n/r$.

Compute. KL-Shampoo's per-step cost is dominated by the right-side mn^2 matmul (applying $R^{-1/2}$) and the periodic n^3 QR refresh of Q_R. Pro-KLShampoo replaces the full preconditioning by a rank-r Kronecker step on the projection and T_{NS} Newton–Schulz iterations on the $m \times n$ residual, and the right-side QR shrinks to r^3; the subspace projection and basis tracking add mnr overhead (e.g., forming $\hat{G}U$). Newton–Schulz is dense matrix multiplication and runs efficiently on GPU, whereas QR is sequential and does not—so the $n^3 \to r^3$ QR shrink is the main per-step wallclock saving, partly offset by the subspace overhead when n is moderate.

M WALLCLOCK COMPARISON WITH MUON

The main results report Pro-KLShampoo's wallclock saving relative to KL-Shampoo (§4.1). Here we report the wallclock saving relative to Muon (Jordan et al., 2024) on LLaMA. Validation loss and memory against Muon at all four scales are reported in Table 1.

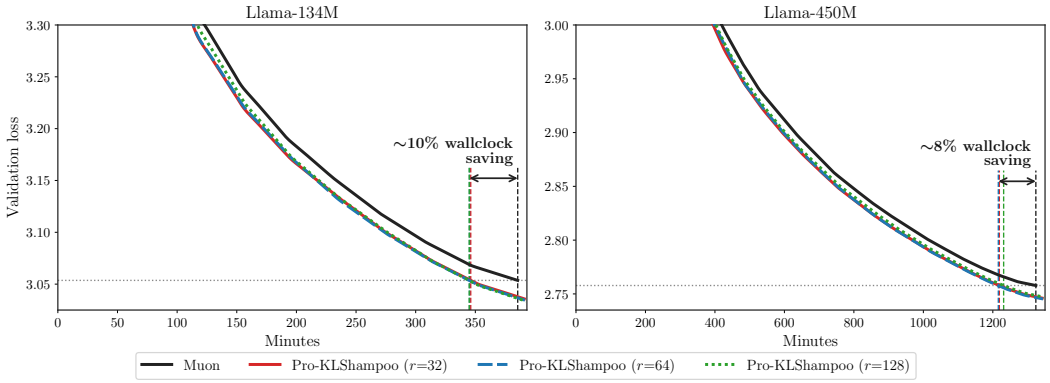

Figure 8: Validation loss versus wallclock time for Muon and Pro-KLShampoo at $r \in \{32, 64, 128\}$ on LLaMA (134M, left; 450M, right). Pro-KLShampoo reaches Muon's final validation loss in approximately 10% less wallclock time at 134M and 8% less at 450M.

On LLaMA (Figure 8), Pro-KLShampoo reaches Muon's final validation loss in less wallclock time at every rank: approximately 10% less at 134M and 8% less at 450M. On GPT-2, Pro-KLShampoo does not show a wallclock saving over Muon.

N COMPARISON WITH COSMOS

We provide a head-to-head comparison with COSMOS (Liu et al., 2025), which combines SOAP-style estimation inside a tracked subspace with Muon on the complement, at matched ranks across all four scales.

COSMOS configuration. We use the official COSMOS release with default hyperparameters (Nesterov momentum 0.95, 5-step Newton–Schulz, mixing weight $\gamma = 0.25$). Embedding and output weights are trained by AdamW, matching §4. The subspace rank r and learning rate are swept independently per scale (Appendix O).

Results. Table 5 reports the head-to-head. At matched rank $r \in \{64, 128\}$, Pro-KLShampoo reaches lower validation loss than COSMOS at three of the four scales (GPT-2 124M, GPT-2 350M, LLaMA 134M), with margins of 0.004–0.026. On LLaMA 450M the gap is small (0.005 in favor of COSMOS at WD$=0$) and reverses under weight decay (below). Pro-KLShampoo also uses lower peak GPU memory at every entry (largest reduction 0.76 GiB on LLaMA 450M).

Weight decay ablation on LLaMA 450M. The LLaMA experiments above follow the GaLore convention of WD$=0$ (§4). We additionally evaluate both methods with weight decay on LLaMA 450M. COSMOS is swept over WD $\in \{0.1, 0.3, 0.5\}$ at its previously-selected learning rate, reaching 2.7264 at $r=64$. Pro-KLShampoo at WD$=0.03$ (no further sweep) reaches 2.7142 at $r=64$—a gap of -0.0122 over COSMOS.

O HYPERPARAMETER SWEEPS

This appendix lists the hyperparameter sweep ranges and the selected values for every method and every model scale used in the experiments of §4. The tuning protocol is that of §4: AdamW

Table 5: Head-to-head comparison with COSMOS at matched rank, across all four scales. Validation loss is the best across hyperparameter sweeps; memory (mem, in GiB) is the corresponding peak per-GPU memory of that run. Memory is comparable only within each scale. Pro-KLShampoo maintains the optimizer state in half precision, while COSMOS uses the full-precision optimizer state of its official release.

Scale	r	COSMOS		Pro-KLShampoo		Δ (Pro $-$ COSMOS)	
		val loss	mem	val loss	mem	val loss	mem
GPT-2 124M	64	3.3014	30.24	3.2754	30.16	-0.0260	-0.08
	128	3.2985	30.25	3.2745	30.17	-0.0240	-0.08
GPT-2 350M	64	3.0492	72.39	3.0348	72.24	-0.0144	-0.15
	128	3.0474	72.41	3.0344	72.25	-0.0130	-0.16
LLaMA 134M	64	3.0383	31.90	3.0345	31.75	-0.0038	-0.15
	128	3.0386	31.91	3.0335	31.76	-0.0051	-0.15
LLaMA 450M	64	2.7404	72.62	2.7453	71.86	$+0.0049$	-0.76
	128	2.7416	72.64	2.7469	71.89	$+0.0053$	-0.75

is tuned first; its selected learning rate and weight decay are then fixed for the embedding and output layers under all matrix optimizers; the matrix-side hyperparameters of each method are then swept independently per method. For Pro-KLShampoo, the mixing weight α_{kl} is swept jointly with the learning rate. On LLaMA, weight decay is fixed to 0 throughout, following the GaLore convention (Zhao et al., 2024); on GPT-2, weight decay is swept jointly with the learning rate. Other hyperparameters follow the source releases of Lin et al. (2025); Jordan et al. (2024); Liu et al. (2025): AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$; Muon and Pro-KLShampoo use Nesterov momentum with coefficient 0.95; KL-Shampoo uses momentum coefficient 0.95; KL-Shampoo and Pro-KLShampoo use $\beta_2 = 0.95$ and refresh the preconditioner every 10 steps; Muon, COSMOS, and Pro-KLShampoo use 5-step Newton–Schulz iteration for orthogonalization.

Architectures. GPT-2 124M: d=768, MLP intermediate $4d$, 12 layers. GPT-2 350M: d=1024, MLP intermediate $4d$, 24 layers. LLaMA-134M: d=768, MLP intermediate 2048, 12 layers, SwiGLU. LLaMA-450M: d=1024, MLP intermediate 2816, 30 layers, SwiGLU.

Batch sizes and sequence lengths. The sequence length is 1024 at every scale. The batch sizes, in sequences per step, are 512 for GPT-2 124M, 512 for GPT-2 350M, 960 for LLaMA-134M, and 1620 for LLaMA-450M.

Hardware and measurement. Each run uses 2 or 4 A100 GPUs; at a given scale, all methods use the same GPUs, per-GPU batch and gradient accumulation. Wallclock counts training time only, measured with CUDA synchronization and excluding validation and the first 10 steps, which include one-time start-up costs such as `torch.compile` compilation; peak memory is the maximum allocated CUDA memory per GPU.

Reporting protocol and seed variation. Best validation loss over the hyperparameter sweeps is reported; for each method and hyperparameter configuration we run two seeds and report the best validation loss. As a significance check, we ran two additional seeds at the selected hyperparameters of each method on GPT-2 350M (Pro-KLShampoo at r=128) and report all four outcomes, the two sweep seeds and the two additional ones (Table 6). We chose GPT-2 350M because the gains on GPT-2 are smaller, making it the conservative setting. The ranges do not overlap (exact one-sided rank-sum test over the four seeds, $p = 0.014$), and they also do not overlap for the two additional seeds alone. Table 1 and all figures keep the sweep runs; one additional KL-Shampoo seed (3.0384) is below its Table 1 value (3.0397), by less than one standard deviation, while Pro-KLShampoo's best run is its Table 1 value (3.0344). For reference, the seed standard deviation at 350M (at most 0.0014) is an order of magnitude below the gaps between Pro-KLShampoo and KL-Shampoo on LLaMA in Table 1 (3.0335 vs. 3.0504 at 134M; 2.7453 vs. 2.7650 at 450M).

Table 6: Validation loss on GPT-2 350M over four seeds at the selected hyperparameters of each method (Pro-KLShampoo at r=128; two sweep seeds and two additional seeds).

Method	Validation loss on 4 seeds	Mean	Std
Pro-KLShampoo	3.0344, 3.0346, 3.0354, 3.0364	3.0352	0.0009
KL-Shampoo	3.0384, 3.0397, 3.0406, 3.0418	3.0401	0.0014

Table 7: Sweep ranges and selected values, GPT-2 124M on FineWeb-10B. "–" indicates the hyperparameter is not used by the method. Selected values are given in parentheses.

Method	lr	wd	α_{kl}
AdamW	{1e−3, 2e−3, 3e−3} (2e−3)	{2e−2, 5e−2, 1e−1} (5e−2)	–
Muon	{1e−2, 2e−2, 3e−2} (2e−2)	{0, 2e−2, 3e−2, 5e−2, 1e−1} (2e−2)	–
KL-Shampoo	{1e−3, 2e−3, 3e−3} (2e−3)	{5e−2, 1e−1, 1.5e−1, 2e−1, 2.5e−1, 3e−1} (2.5e−1)	–
COSMOS (r=64)	{2e−4, 5e−4, 1e−3, 2e−3} (5e−4)	{1e−1, 2.5e−1, 5e−1} (2.5e−1)	–
COSMOS (r=128)	{2e−4, 5e−4, 1e−3, 2e−3} (5e−4)	{1e−1, 2.5e−1, 5e−1} (2.5e−1)	–
Pro-KLShampoo (r=32)	{1e−2, 2e−2, 3e−2} (2e−2)	{1e−2, 2e−2, 3e−2, 5e−2, 1e−1} (3e−2)	{5e−3, 1e−2, 1.5e−2} (1e−2)
Pro-KLShampoo (r=64)	{1e−2, 2e−2, 3e−2} (2e−2)	{1e−2, 2e−2, 3e−2, 5e−2, 1e−1} (3e−2)	{5e−3, 1e−2, 1.5e−2} (1e−2)
Pro-KLShampoo (r=128)	{1e−2, 2e−2, 3e−2} (2e−2)	{1e−2, 2e−2, 3e−2, 5e−2, 1e−1} (3e−2)	{5e−3, 1e−2, 1.5e−2} (1e−2)
Smok-Hop (r=128)	uses KL-Shampoo's selected lr (2e−3)	2.5e−1	–

Table 8: Sweep ranges and selected values, GPT-2 350M on FineWeb-10B. Selected values are given in parentheses.

Method	lr	wd	α_{kl}
AdamW	{1e−3, 2e−3, 3e−3} (2e−3)	{5e−2, 1e−1, 1.5e−1} (5e−2)	–
Muon	{1e−2, 2e−2, 3e−2, 5e−2} (2e−2)	{0, 2e−2, 3e−2, 5e−2, 1e−1} (3e−2)	–
KL-Shampoo	{1e−3, 2e−3, 3e−3} (2e−3)	{5e−2, 1e−1, 1.5e−1, 2e−1, 2.5e−1, 3e−1} (2.5e−1)	–
COSMOS (r=64)	{2e−4, 5e−4, 1e−3, 2e−3} (1e−3)	{2e−1, 3e−1, 5e−1} (5e−1)	–
COSMOS (r=128)	{2e−4, 5e−4, 1e−3, 2e−3} (1e−3)	{2e−1, 3e−1, 5e−1} (5e−1)	–
Pro-KLShampoo (r=32)	{1e−2, 2e−2, 3e−2} (2e−2)	{1e−2, 2e−2, 3e−2, 5e−2, 1e−1} (3e−2)	{5e−3, 1e−2, 1.5e−2} (1e−2)
Pro-KLShampoo (r=64)	{1e−2, 2e−2, 3e−2} (2e−2)	{1e−2, 2e−2, 3e−2, 5e−2, 1e−1} (3e−2)	{5e−3, 1e−2, 1.5e−2} (1e−2)
Pro-KLShampoo (r=128)	{1e−2, 2e−2, 3e−2} (2e−2)	{1e−2, 2e−2, 3e−2, 5e−2, 1e−1} (3e−2)	{5e−3, 1e−2, 1.5e−2} (1e−2)

Table 9: Sweep ranges and selected values, LLaMA 134M on C4. Weight decay is 0 for all methods, following the GaLore convention. Selected values are given in parentheses.

Method	lr	α_{kl}
AdamW	{1e−3, 2e−3, 3e−3, 5e−3} (3e−3)	–
Muon	{5e−3, 1e−2, 2e−2, 3e−2} (2e−2)	–
KL-Shampoo	{2e−3, 3e−3, 4e−3, 5e−3, 8e−3} (5e−3)	–
COSMOS (r=64)	{2e−4, 5e−4, 1e−3, 2e−3} (1e−3)	–
COSMOS (r=128)	{2e−4, 5e−4, 1e−3, 2e−3} (1e−3)	–
Pro-KLShampoo (r=32)	{1e−2, 2e−2, 3e−2} (2e−2)	{5e−3, 1e−2, 1.5e−2} (1e−2)
Pro-KLShampoo (r=64)	{1e−2, 2e−2, 3e−2} (2e−2)	{5e−3, 1e−2, 1.5e−2} (1e−2)
Pro-KLShampoo (r=128)	{1e−2, 2e−2, 3e−2} (3e−2)	{5e−3, 1e−2, 1.5e−2} (1e−2)
Smok-Hop (r=128)	{2e−3, 3e−3, 5e−3} (3e−3)	–

The subspace-only and complement-only variants on LLaMA 134M use the selected hyperparameters of Pro-KLShampoo at r=128 with the learning rate scaled by 1.5 as on GPT-2 124M (same rationale).

Table 10: Sweep ranges and selected values, LLaMA 450M on C4. Weight decay is 0 for all methods. Selected values are given in parentheses.

Method	lr	α_{kl}
AdamW	$\{2e{-}3, 3e{-}3, 5e{-}3\}$ (2e−3)	–
Muon	$\{2e{-}2, 3e{-}2, 5e{-}2\}$ (2e−2)	–
KL-Shampoo	$\{2e{-}3, 3e{-}3, 5e{-}3\}$ (3e−3)	–
COSMOS (r=64)	$\{2e{-}4, 5e{-}4, 1e{-}3, 2e{-}3\}$ (5e−4)	–
COSMOS (r=128)	$\{2e{-}4, 5e{-}4, 1e{-}3, 2e{-}3\}$ (5e−4)	–
Pro-KLShampoo (r=32)	$\{2e{-}2, 3e{-}2\}$ (2e−2)	$\{5e{-}3, 1e{-}2, 1.5e{-}2\}$ (5e−3)
Pro-KLShampoo (r=64)	$\{2e{-}2, 3e{-}2\}$ (2e−2)	$\{5e{-}3, 1e{-}2, 1.5e{-}2\}$ (5e−3)
Pro-KLShampoo (r=128)	$\{2e{-}2, 3e{-}2\}$ (2e−2)	$\{5e{-}3, 1e{-}2, 1.5e{-}2\}$ (5e−3)
Smok-Hop (r=128)	uses KL-Shampoo's selected lr (3e−3)	–

Ablation hyperparameters. The subspace-only variant and the complement-only variant reuse the selected hyperparameters from the corresponding Pro-KLShampoo r=128 runs, except with a learning rate scaled by 1.5 to compensate for the smaller update after one component is removed (the two components have orthogonal right supports, so removing either one reduces the Frobenius norm of the update). The factor is a heuristic and was not tuned for either variant.

The 10,000-iteration protocol. We additionally train KL-Shampoo and Pro-KLShampoo on GPT-2 124M under the budget protocol of Lin et al. (2025) (10,000 iterations at 0.5M tokens per iteration; batch 512× sequence length 1024), sweeping momentum and β_2 for both methods (the main experiments fix both to the release defaults of 0.95). Learning rate and weight decay are fixed at each method's selected values from the main experiments (Table 7). Table 11 reports the final validation loss of every cell.

Table 11: Final validation loss on GPT-2 124M under the 10,000-iteration protocol of Lin et al. (2025), with momentum and β_2 swept for both methods (Pro-KLShampoo at r=128).

Momentum	β_2	KL-Shampoo	Pro-KLShampoo
0.95	0.95 (release default)	3.2198	3.2160
0.95	0.90	3.2221	3.2146
0.95	0.99	3.2336	3.2307
0.90	0.90	3.2255	3.2159
0.90	0.95	3.2335	3.2214
0.90	0.99	3.5203	3.2302

Sensitivity to α_{kl} on GPT-2 350M. Appendix J.3 derives model- and layer-dependent brackets for α_{kl} by operator-norm matching at the restricted-KL stationary state. For GPT-2 350M, intersecting the per-layer brackets gives the all-layer feasible interval $[0.001, 0.016]$. The sweep in Table 12 samples this interval and performs best at $\alpha_{\text{kl}} \in \{0.005, 0.01\}$, while the smaller and larger tested values perform worse. For reference, the KL-Shampoo result reported in Table 1 at this scale is 3.0397; every tested cell with $\alpha_{\text{kl}} \in \{0.005, 0.01\}$ lies below that reference.

Table 12: Validation loss of Pro-KLShampoo on GPT-2 350M over α_{kl} and r. "—" indicates a cell that was not run.

α_{kl}	0.001	0.003	0.005	0.01	0.015
r=32	—	3.0390	3.0382	3.0364	3.0390
r=64	—	3.0392	3.0357	3.0348	3.0407
r=128	3.0650	3.0398	3.0349	3.0344	3.0406

Choosing r. The validation loss varies by at most 0.003 over $r \in \{32, 64, 128\}$ at all four scales, and every cell lies below KL-Shampoo's (Table 1); for example, at GPT-2 350M the loss is 3.0364, 3.0348, and 3.0344 at ratios of r to the attention width of 3.1%, 6.3%, and 12.5%. Over the tested

range we observe no need for r to grow with model size. Our guideline for choosing r has three parts. *Default:* use r=128, which required no adjustment over the tested 124M to 450M range; within this range, every tested rank reaches KL-Shampoo's final validation loss in less wallclock time. *Verification:* when a tuning budget allows, run a short pilot (a few thousand steps) of KL-Shampoo, compute the per-factor tail ratio ρ_r (Appendix E), and take the smallest r for which ρ_r is close to one or has dropped sharply from ρ_0. *Extrapolation:* when no pilot budget is available at a substantially larger width, extrapolate by ratio rather than by the absolute r=128: pick r in the 3.1%–16.7% band realized in our experiments by $r \in \{32, 64, 128\}$ at widths 768 and 1024. Both a fixed spike size and a proportionally growing spike are consistent with our measurements, since even the proportional case stays below 128 at every tested width. At substantially larger widths, the two cases may separate. Extrapolating within the observed ratio band is therefore a conservative choice, whereas fixing r=128 assumes that the spike size remains approximately constant. Adapting r over time is not a drop-in schedule: changing r during training changes the dimensions of the maintained states U, S and their EMAs, so a dynamic-rank version would require explicit rules for expanding or truncating these states and for initializing newly added directions. Given this state-transfer complexity, the observed insensitivity to the tested fixed ranks, and tail ratios that change little between the early and late stages at the two measured scales (Appendix E), we leave dynamic ranks to future work; different fixed ranks across layers, given their different matrix dimensions and spectral structures, are a promising direction.

Ranks beyond 128**.** For the rank ablation of §4.2, we also evaluate r=256 and r=512 on both GPT-2 124M and LLaMA-450M, using the training protocol of each model. For each new rank, we swept learning rate and weight decay over a 2×2 grid on GPT-2 124M (Table 13), and learning rate over $\{0.02, 0.03\}$ on LLaMA-450M with $\alpha_{\text{kl}} = 0.005$ (Table 14). The r=128 references from Table 1 are 3.2745 (GPT-2 124M) and 2.7469 (LLaMA-450M).

Table 13: Validation loss of Pro-KLShampoo at $r \in \{256, 512\}$ on GPT-2 124M over the 2×2 learning-rate and weight-decay grid.

r	lr 0.02, wd 0.02	lr 0.02, wd 0.03	lr 0.03, wd 0.02	lr 0.03, wd 0.03
256	3.2731	3.2718	3.2772	3.2786
512	3.2752	3.2755	3.2788	3.2792

Table 14: Validation loss of Pro-KLShampoo at $r \in \{256, 512\}$ on LLaMA-450M ($\alpha_{\text{kl}} = 0.005$) over the learning rate.

r	lr 0.02	lr 0.03
256	2.7425	2.7492
512	2.7495	2.7486

On GPT-2 124M, the best r=256 result (3.2718) improves on r=128 (3.2745), at the cost of higher per-step time; r=512 is worse than r=256 in every cell and returns to the r=128 level (best 3.2752), so there is no further gain at r=512. The same ordering is observed on LLaMA-450M, where the r=128 reference is 2.7469 (Table 1): r=256 reaches 2.7425 and r=512 reaches 2.7486. One possible explanation for the loss of the gain at r=512 is drift toward full KL-Shampoo: as r grows, a larger share of the update is governed by the tracked subspace, while the orthogonalized complement shrinks. The state to estimate also grows with r, including the $n \times r$ tracked basis U and the $r \times r$ projected factor S, which increases the estimation burden from the same gradient stream. The endpoint of this drift is itself weaker: full KL-Shampoo reaches 3.2796, behind 3.2718 at r=256. Under this explanation, moving toward a weaker endpoint naturally loses part of the gain. The small gain at r=256 is consistent with the spectrum: GPT-2 124M has the least-flat MLP tails (Appendix E), so extending the tracked subspace to 256 captures part of the remaining non-flat tail.

Truncated pseudo-inverse KL-Shampoo. For the truncated variant of §4.2 (GPT-2 124M, r=128; the top-r eigenpairs retained, the tail zeroed, and the pseudo-inverse used in both preconditioning and

factor estimation), we fixed the weight decay at KL-Shampoo's selected value (Table 7) and swept the learning rate over five points; Table 15 reports the final validation loss.

Table 15: Final validation loss of truncated pseudo-inverse KL-Shampoo on GPT-2 124M (r=128) over the learning rate.

Learning rate	0.001	0.002	0.003	0.005	0.008
Truncated KL-Shampoo	3.7498	3.7085	3.7090	3.7734	3.8101

The best validation loss of the truncated variant is 3.7085; for reference, at this scale Pro-KLShampoo reaches 3.2745, KL-Shampoo 3.2796, and Smok-Hop, which keeps the complement update but preconditions it by a single shared scalar, 3.2791. At the restricted-KL stationary state, $\mathbb{E}\|\mu_\perp^{-1/2} L^{-1/2} G\, P_\perp\|_F^2 = m(n-r)$ versus mr for the subspace, so the complement carries $(n-r)/r$ times the subspace's expected energy (from $5\times$ to $31\times$ across the tested layers); zeroing it removes most of the update. The result is consistent with the subspace-only ablation of Table 2, whose update likewise zeroes the complement, though its factor estimation differs from exact truncation (3.7803 vs. 3.2745 on GPT-2 124M and 3.4598 vs. 3.0335 on LLaMA-134M). Truncation also does not reduce cost: identifying the small eigenvalues still requires forming the full $n \times n$ factor and eigendecomposing it, so memory remains $O(n^2)$ and the eigendecomposition cost remains $O(n^3)$. Taken together, the three treatments of the complement compare as follows. On GPT-2 124M, zeroing the complement worsens validation loss by about 0.43, whereas retaining it with a single shared scalar stays within 0.005 of both Pro-KLShampoo and KL-Shampoo; at the LLaMA scales the scalar-tail variant trails Pro-KLShampoo by 0.014 (134M) and 0.024 (450M), and at 450M it also trails KL-Shampoo itself (2.7704 vs. 2.7650); the orthogonalized complement is best at every scale (Table 2). A single shared scalar is exact only when the complement tail is perfectly flat. When the tail is only approximately flat, a single scalar cannot represent direction-specific scaling, whereas the polar step recovers the algebraic form of the complement update on the current minibatch (§3.3). The ablation in Table 2 suggests that this distinction becomes more important with scale.