DAWN: SCALABLE MATRIX-WHITENING OPTIMIZATION VIA SPECTRAL SELECTIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Matrix-whitening optimizers such as Muon improve the data efficiency of language-model pretraining, but their full-spectrum polar updates are expensive and treat all singular directions as equally reliable. We argue that this may be unnecessary: across training, gradient matrices have a stable low-rank head and a noisy residual tail. Polarizing the tail amplifies noise to unit magnitude and wastes the most expensive part of the per-step compute. We propose DAWN (Decoupled Adaptive Whitening of Noisy gradients), which polarizes only the empirical head and applies a row-wise RMS update on the residual—a block-Schatten partition with one coarse spectral block on the head and singleton row blocks on the tail. DAWN reduces the per-layer whitening cost from $O(d^3)$ to $O(d^2k)$, matches Muon on GPT-2XL-scale (1.5B) pretraining, and lowers optimizer wall-clock time.

1 INTRODUCTION

Matrix-whitening optimizers such as Muon (Jordan et al., 2024), SOAP (Vyas et al., 2024), and Shampoo (Gupta et al., 2018) have been shown to improve token efficiency on large language model pretraining, with carefully scaled hyperparameter transfer studies confirming a consistent $\sim 1.4\times$ speedup over AdamW from 190M to 1.4B parameters (Qiu et al., 2025). The cost of these gains, however, has two faces. *Computationally*, the per-layer polar/whitening operation scales with layer width and is not reduced by data parallelism the way forward–backward compute is: as cluster size grows, micro-batches are distributed (FB time falls), but per-GPU optimizer cost is constant. Team et al. (2025) report that at trillion-parameter scale, Muon's optimizer step occupies up to $\sim 20\%$ of forward–backward time in their training stack and motivates a custom distributed Muon implementation; smaller deployments (Zeng et al., 2025) report similar concerns. *Statistically*, full polar trusts the noisy gradient $Y = G_\star + E$ as if it were $G_\star$ itself, but the second-moment proxy that drives whitening is biased:

$$\mathbb{E}[\,Y^\top Y\,] \;=\; G_\star^\top G_\star \;+\; \Sigma_E, \qquad \Sigma_E := \mathbb{E}[\,E^\top E\,], \tag{1}$$

so the population $G_\star^\top G_\star$ and its empirical proxy $Y^\top Y$ differ by an additive noise covariance whose eigenvalues do not vanish with sample size at fixed B. By the Bai-Ben-Arous-Péché transition (Baik et al., 2005; Paul, 2007), this gap inflates every observed singular value of Y above its population value and—more importantly—makes any observed singular vector below the noise edge $\sigma_E(\sqrt{m}+\sqrt{n})$ asymptotically Haar-random, carrying no information about the population direction. Polarizing such directions lifts noise to unit magnitude and wastes the most expensive part of the per-step compute.

The Muon update base on the matrix sign $\mathrm{Polar}(Y) = \widehat{U}\widehat{V}^\top$, where $Y = \widehat{U}\widehat{\Sigma}\widehat{V}^\top$ is the SVD of the noisy observation. This sets every observed singular value of Y to one—a uniform operation that does not distinguish whether $\widehat{\sigma}_i$ comes from a true population spike or from the inflated tail of equation 1. The two faces of the cost meet here: every direction contributes equally to the $O(d^3)$ compute, and every direction is lifted to unit polar magnitude regardless of whether $\widehat{u}_i, \widehat{v}_i$ are reliable estimates of $u_i^\star, v_i^\star$.

Two empirical and theoretical lines suggest that this head/tail asymmetry is large in practice. Empirically, gradient matrices in language model training concentrate 80–95% of their Frobenius energy in the top ~ 100 singular directions (Sagun et al., 2017; Ghorbani et al., 2019; Papyan, 2018), even at width $d \gg 1000$ (Section 4). Theoretically, in an idealized spiked-gradient model, singular-vector alignment undergoes a phase transition: spikes above the noise edge retain a positive cosine with the

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.

population direction, and those below the edge are asymptotically Haar-random (Baik et al., 2005; Paul, 2007; Benaych-Georges & Nadakuditi, 2011). We do not assume real training exactly matches this model; rather, the model motivates a conservative design rule: *polarize only the empirically stable spectral head*.

We propose **DAWN** (**D**ecoupled **A**daptive **W**hitening of **N**oisy gradients), a Muon variant that polarizes only an empirical low-rank head and applies a bounded first-order update on the residual tail, with a spectral firewall keeping the two updates orthogonal. The whitening cost drops from $O(d^3)$ to $O(d^2k)$ for $k \ll d$. Concretely, our contributions are:

1. We propose DAWN, a head-tail decomposition with head polar whitening on a randomized-SVD-tracked subspace and a row-wise RMS update on the residual, separated by a spectral firewall (Section 2; Theorems 2.1 and 2.2).
2. Empirically, DAWN matches Muon on GPT-2 (124M) and GPT-2XL (1.5B) and a 4B-parameter wide variant, with measurably lower optimizer wall-clock time (Section 4).
3. Post-hoc, we frame DAWN's head-tail split as a *block-Schatten partition refinement*: matrix sign and row-wise RMS are two endpoints of one norm family, and DAWN refines the partition by reliability (Section 3.1).

2 The DAWN Algorithm

Notation. $G_\star \in \mathbb{R}^{m\times n}$ is the population matrix gradient ($m \geq n$); $Y = G_\star + E$ is the observed stochastic gradient. We write $Y = \widehat{U}\widehat{\Sigma}\widehat{V}^\top$ for the empirical SVD and $G_\star = U^\star\Sigma^\star V^{\star\top}$ for the population SVD. Denote $\beta := m/n$ and $d := \min(m,n) = n$. The polar (matrix-sign) factor is $\mathrm{Polar}(M) := UV^\top$ for any $M = U\Sigma V^\top$. Full notation reference in Section B.

Design principle. Matrix sign (Muon) and row-wise RMS (AdamW-style) are two endpoints of a single update-geometry family parameterized by how the gradient matrix is partitioned: one coupled spectral block at one end, m singleton row blocks at the other. The design question is therefore not "coupled matrix update vs. first-order" but *which directions deserve a coupled update*. Intuitively, coupling pays off where the singular structure of the observed Y is statistically reliable; the safer choice is to refine to row-wise singletons where it is not. DAWN instantiates this principle with a hard binary head/tail split: a coarse spectral block on a low-rank head subspace (head polar at $O(d^2k)$) and singleton row blocks on the residual (row-wise RMS). The formal block-Schatten partition and the statistical justification (per-mode and total MSE arguments) are deferred to Section 3.

2.1 Algorithm

Given $\tilde{G} \in \mathbb{R}^{m\times n}$ and rank k, compute the top-k subspace $Q \in \mathbb{R}^{m\times k}$ (e.g., randomized SVD) and let $B = Q^\top\tilde{G}$, $R = (I - QQ^\top)\tilde{G}$. The head update applies polar to the $k \times n$ projected matrix:

$$\Delta_{\mathrm{head}} \;=\; \sqrt{k}\,\frac{Q\ \mathrm{NS}(B)}{\|Q\ \mathrm{NS}(B)\|_F}, \qquad \mathrm{NS} \approx \mathrm{Polar}\,. \tag{2}$$

The tail uses a row-wise RMS-normalized residual with running second-moment $v_i \leftarrow \beta_2 v_i + (1-\beta_2)\|R_i\|_2^2/n$:

$$\widetilde{R}_i = R_i/\sqrt{v_i+\epsilon}, \quad \widetilde{R} \leftarrow (I - QQ^\top)\widetilde{R}, \quad \Delta_{\mathrm{tail}} = \sqrt{d-k}\,\frac{\widetilde{R}}{\|\widetilde{R}\|_F}. \tag{3}$$

The re-projection $(I - QQ^\top)\widetilde{R}$ is the *spectral firewall*: row-wise RMS scaling can reintroduce components into $\mathcal{H} = \mathrm{colsp}(Q)$, which we project out so the head and tail updates remain Frobenius-orthogonal. Combined: $\Delta = \Delta_{\mathrm{head}} + \Delta_{\mathrm{tail}}$, optionally followed by global RMS calibration $\Delta \leftarrow \mathrm{target_rms}\sqrt{mn}\,\Delta/\|\Delta\|_F$.

Theory–implementation note. The RowRMS operation in Eq. equation 3 is a practical coordinate-wise preconditioning step, in the same spirit as adaptive normalization tricks used broadly in stochastic optimization and recent practical Muon variants. Because row-wise rescaling can rotate the singular basis of the raw residual, the mode-wise MSE analysis in Section 3 is stated for the corresponding theory-aligned raw-residual tail, before this additional preconditioning. Thus the theory motivates the head–tail separation and the decision not to polarize the tail; RowRMS should be viewed as an implementation refinement rather than a consequence of the MSE theorem. Appendix E.6.4 verifies that removing RowRMS leaves performance comparable in the 124M setting.

Theorem 2.1 (Spectral firewall). *After re-projection $\widetilde{R} \leftarrow (I - QQ^\top)\widetilde{R}$, $\langle \Delta_{\text{head}}, \Delta_{\text{tail}} \rangle_F = 0$. Tail noise cannot leak into head, and vice versa.*

Proof sketch. $\Delta_{\text{head}} \in \text{colsp}(Q)$ and $\widetilde{R} \in \text{colsp}(Q)^\perp$ after re-projection; hence $Q^\top \widetilde{R} = 0$ and the inner product vanishes. Without the re-projection step, RMS rescaling can violate orthogonality (Appendix C).

Theorem 2.2 (Energy control). *$\|\Delta\|_F^2 = k + (d-k) = d$ before global calibration; with calibration $\|\Delta\|_F = \text{target_rms}\sqrt{mn}$. Equivalently, head and tail energies are dimension-balanced.*

DAWN applies polar updates only to reliable head directions and uses a first-order RMS update on the noise-dominated tail. The tail is not discarded, but its singular values are not artificially lifted to one. This yields $O(d^2k)$ cost versus Muon's $O(d^3)$ while preserving useful tail gradient signal. The theory, ablations, pseudocode and complexity details are provided in Section 3 and Appendices D.

3 THEORY

This section analyses the design choices of Section 2 post-hoc, in three pieces. Section 3.1 formalizes the partition framing of the design principle: matrix sign and row-wise RMS are endpoints of a block-Schatten norm family, and DAWN is one specific partition refinement in this family. Section 3.2 quantifies the per-mode MSE cost ($(1-\gamma_i)^2$) of polarizing weak directions in the BBP-style spiked-gradient model. Section 3.3 studies the theory-aligned raw-residual tail and shows when its energy-preserving spectral weighting improves over uniform polar. The practical RowRMS step can rotate this singular basis and is therefore not claimed to be derived by the mode-wise MSE result; it is an additional preconditioning refinement. None of these statements depends on training dynamics specific to deep networks; they are statements about update geometry and the spectrum–reliability relationship.

3.1 BLOCK-SCHATTEN GEOMETRY: DAWN AS A PARTITION REFINEMENT

Take a partition $\mathcal{P} = \{S_1, \ldots, S_L\}$ of the row index set $[m]$ and the norm

$$\|\Delta\|_{\mathcal{P},2,\infty} \;:=\; \max_{S \in \mathcal{P}} \|\Delta_{S,:}\|_2, \tag{4}$$

the largest spectral norm across blocks. Two limits recover familiar geometries:

- *Single block* $\mathcal{P} = \{[m]\}$: $\|\cdot\|_{\mathcal{P},2,\infty} = \|\cdot\|_2$, and the steepest-descent direction for $\langle G, \Delta \rangle_F$ subject to $\|\Delta\|_{\mathcal{P},2,\infty} \le 1$ is the global polar $UV^\top$. This is Muon.
- *Singleton blocks* $\mathcal{P} = \{\{1\}, \ldots, \{m\}\}$: $\|\cdot\|_{\mathcal{P},2,\infty} = \max_i \|\Delta_{i,:}\|_2$, whose steepest-descent direction normalizes each row independently: $\Delta^\star_{i,:} = G_{i,:}/\|G_{i,:}\|_2$. This is the row-wise group sign / RMS-style update.

DAWN's choice is the partition that groups high-reliability modes into one coarse spectral block (the head subspace $\text{colsp}(Q)$) and treats the unreliable residual via fine row-wise blocks. At the level of the reference geometry, DAWN can be viewed as a steepest-descent direction under a hybrid norm: spectral on $\text{colsp}(Q)$, row-wise ℓ_2 on $(I - QQ^\top)$. The formal block-Schatten generalization (with Schatten-p instead of spectral norm in each block) is in Appendix B.5.

3.2 PER-MODE EXCESS MSE OF HARD POLAR

The post-hoc analysis below quantifies what hard polar costs at each individual tail direction whose observed singular structure is unreliable. The setup is a noisy-observation model: we observe $Y = G_\star + E$ and want to recover $\text{Polar}(G_\star)$. The following theorem decomposes the per-mode MSE of any rank-one estimator $a_i \widehat{u}_i \widehat{v}_i^\top$ targeting the population mode $u_i^\star v_i^{\star\top}$, identifies the per-mode optimal weight $a_i^\star = \gamma_i$, and gives the closed-form excess MSE Muon pays by lifting every retained direction to unit magnitude.

Theorem 3.1 (Per-mode excess MSE of matrix Muon). *Let $Y = G_\star + E$, with empirical SVD $Y = \widehat{U}\widehat{\Sigma}\widehat{V}^\top$ and population SVD $G_\star = U^\star \Sigma^\star V^{\star\top}$. Let $\gamma_i := \mathbb{E}\big[\langle \widehat{u}_i, u_i^\star \rangle \langle \widehat{v}_i, v_i^\star \rangle \;\big|\; Y\big] \in [0,1]$ be the reliability of the i-th observed singular direction. For estimators of the form $a_i\, \widehat{u}_i \widehat{v}_i^\top$ targeting the population rank-one component $u_i^\star v_i^{\star\top}$, the per-mode mean-squared error is $\mathcal{R}_i(a_i) = a_i^2 + 1 - 2a_i\gamma_i$, minimized at $a_i^\star = \gamma_i$. The matrix Muon update applies $a_i = 1$ to every retained singular mode and therefore pays excess MSE*

$$\mathcal{R}_i(1) - \mathcal{R}_i(\gamma_i) \;=\; (1-\gamma_i)^2, \tag{5}$$

relative to the per-mode optimal $a_i = \gamma_i$. Furthermore, in an idealized isotropic Gaussian spiked model with noise $E_{ij} \sim \mathcal{N}(0, \sigma_E^2)$ and aspect ratio $m/n \to \beta$, BBP phase transition gives $\gamma_i \to 0$ for any singular direction below the noise edge $\widehat{\sigma}_i \le \sigma_E(\sqrt{m} + \sqrt{n})$, in which case the excess MSE approaches 1. *Matrix Muon lifts asymptotically uninformative noise to unit magnitude.*

Proof sketch. The Frobenius expansion of $\|a_i \widehat{u}_i \widehat{v}_i^\top - u_i^\star v_i^{\star\top}\|^2$ gives a quadratic-in-a_i risk; $a_i^\star = \gamma_i$ minimizes it; substituting $a_i = 1$ vs γ_i gives $(1-\gamma_i)^2$. The BBP statement is the singular-value version of Paul (2007); the explicit reliability formula is in Section B.6. Full derivation in Sections B.4 and B.6. □

Theorem 3.1 is an oracle motivation for treating low-γ directions differently from high-γ ones, not a derivation of DAWN. Two further oracle analyses—matrix sign as spectral-norm steepest descent and the MMSE matrix-sign estimator as a shrunken polar (Theorem B.7) in the appendix.

3.3 Total-MSE Comparison

The per-mode formula equation 5 treats singular modes one at a time. Sum over the tail subspace $\mathcal{T}$ to get the total MSE of either estimator against the population $G_\star$ (cross-mode terms vanish by orthogonality of singular bases; for details see Section B.4): $\mathrm{MSE}_{\mathrm{Muon}} = \sum_{i\in\mathcal{T}}(2 - 2\gamma_i)$,and $\mathrm{MSE}_{\mathrm{DAWN}} = \sum_{i\in\mathcal{T}}\big(a_i^2 + 1 - 2a_i\gamma_i\big)$, where a_i is the per-mode weight DAWN places on the i-th tail direction. The difference factors:

$$\mathrm{MSE}_{\mathrm{DAWN}} - \mathrm{MSE}_{\mathrm{Muon}} = \sum_{i\in\mathcal{T}}(a_i - 1)\big(a_i + 1 - 2\gamma_i\big). \tag{6}$$

For the theory-aligned reference tail, we normalize the raw residual to fixed Frobenius energy: $\Delta_{\mathrm{tail}}^{\mathrm{ref}} = \sqrt{r}\, R/\|R\|_F$ with $r = |\mathcal{T}|$, so $\sum_i a_i^2 = r = \sum_i 1^2$ matches Muon's tail energy exactly. Under this constraint, $\mathrm{MSE}_{\mathrm{Muon}} = 2r - 2\sum_i \gamma_i$ and $\mathrm{MSE}_{\mathrm{ref}} = 2r - 2\sum_i a_i\gamma_i$, so

$$\boxed{\mathrm{MSE}_{\mathrm{ref}} < \mathrm{MSE}_{\mathrm{Muon}} \iff \sum_{i\in\mathcal{T}} a_i\gamma_i > \sum_{i\in\mathcal{T}} \gamma_i.} \tag{7}$$

The reference raw-residual estimator beats Muon *exactly* when its tail weights are sufficiently aligned with reliability. This is not a free lunch from changing shape: under uniform reliability $\gamma_i \equiv \bar{\gamma}$, Cauchy–Schwarz gives $\sum_i a_i \le \sqrt{r \cdot \sum_i a_i^2} = r$ with equality iff $a_i \equiv 1$, so Muon's uniform polar is in fact MSE-optimal among all energy-preserving estimators when reliability is flat. The potential advantage comes from the spectrum, not the algebra.

For the raw residual, the spectral weights satisfy $a_i \propto s_i$. In the BBP regime, larger singular values correspond to more reliable directions, while weak directions near the noise edge become progressively less identifiable. This gives the theory-aligned mechanism for retaining, rather than polarizing, the tail. The practical RowRMS preconditioner in Eq. equation 3 need not preserve these singular vectors exactly, so we do not identify its implemented per-mode weights with a_i in this argument. Instead, it is an implementation-level conditioning trick layered on top of the same head–tail decomposition. The no-RowRMS ablation in Appendix E.6.4 gives comparable three-seed performance, supporting that the core mechanism does not depend on this extra normalization. The detailed per-mode and total-MSE analyses are given in Equations (5) and (7).

4 Experiments

Synthetic Validation. We validate the head-tail framework on controlled synthetic data: $d \times d$ matrices with rank-r signal embedded in isotropic noise. We test six spectral decay patterns (flat $\kappa = 1$, linear $\kappa = 10$, exponential, power-law $\beta \in \{0.5, 1.0\}$, step) across SNR levels from 0.5 to 20 and matrix sizes from 512 to 16K.

Key findings (Fig. 1): (1) **DAWN excels at low SNR.** Panel (a) shows that when signal-to-noise ratio drops below 1, Muon's full-spectrum whitening amplifies tail noise, causing subspace alignment to collapse. DAWN's selective head whitening preserves signal: at SNR=0.5, DAWN achieves 3–5$\times$ better alignment than Muon. (2) **Fast-decaying spectra favor DAWN.** The gap is largest for exponential and power-law spectra where signal concentrates in few directions—exactly the head-tail regime our theory predicts. (3) **Speedup scales with model width.** Panel (b) shows DAWN achieves

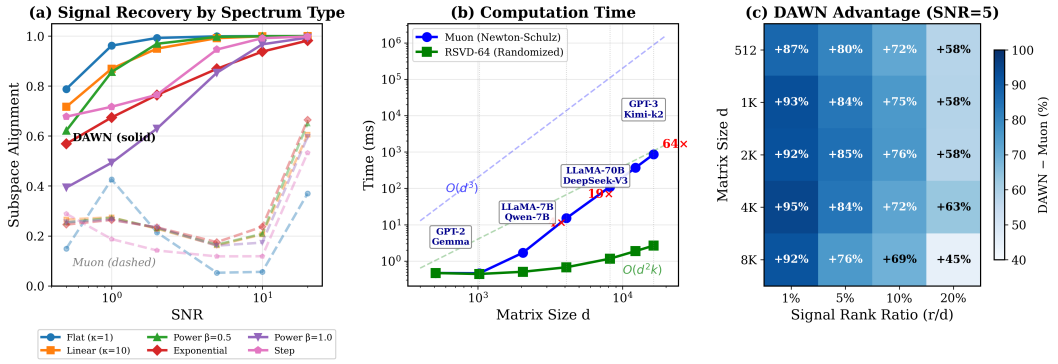

Figure 1: **Synthetic validation of the head-tail framework. (a)** DAWN consistently outperforms Muon in signal-subspace recovery, particularly at low SNR. **(b)** Computational scaling comparison between Muon and DAWN. **(c)** DAWN advantage increases as signal spectra become more concentrated and low-rank. See Section H for setup details and metric definitions.

Table 1: Ablation on GPT-2 (124M) at 8B tokens across major DAWN design choices. Val-loss uncertainty is estimated from the 3-seed. Best-performing configurations are highlighted in green; substantially degraded variants in red.

Configuration	Val Loss	Δ
Muon (NS)	**3.024 ±0.003**	—
AdamW	3.069 ±0.004	+0.045
Alternative Optimizers		
Shampoo (Gupta et al., 2018)	3.050 ±0.004	+0.026
Sophia-H	3.039 ±0.004	+0.015
Efficient Head Computation		
Periodic SVD (p=40)	**3.020 ±0.003**	-0.004
Periodic SVD (p=20)	3.042 ±0.003	+0.018
Adaptive (SVD on drift)	3.035 ±0.005	+0.011
Anneal (rank 128→4)	3.065 ±0.006	+0.041

Configuration	Val Loss	Δ
Every-Step SVD Variants		
Every-Step SVD (k=32)	3.072 ±0.004	+0.048
Every-Step SVD (k=64)	3.042 ±0.004	+0.018
Every-Step SVD (k=128)	3.057 ±0.004	+0.033
Tail Processing Variants		
Bounded tail	3.036 ±0.004	+0.012
Normalized tail	3.028 ±0.003	+0.004
Random tail (Gaussian)	3.20 ±0.02	+0.18
Component Ablations		
Head-only (k=128, no tail)	3.28 ±0.02	+0.26
Head-only (k=64, no tail)	3.47 ±0.03	+0.45
No orthogonalization	4.21 ±0.09	+1.19

19× speedup at LLaMA-7B scale ($d = 4096$) and 64× at GPT-3/Kimi-k2 scale (d = 12–16K). For frontier models with $d > 16$K, the gap widens further. (4) **Consistent across tested conditions.** Panel (c) shows DAWN better align with true gradient matrix $G \cdot G^T$ (+45% to +95%) holds across all tested matrix sizes and rank ratios in this synthetic setup, with larger gains when signal is more concentrated (low rank ratio). Full experimental setup in Section H.

4.1 Mechanism Ablation: What Makes Split Training Work?

Before scaling up, we identify which mechanisms are essential. We decompose the gradient G into head (top-k subspace) and tail (residual): $G = QQ^\top G + (I - QQ^\top)G$, where $Q \in \mathbb{R}^{m\times k}$ contains top-k left singular vectors. We then vary three components: (1) **Q computation**—how to find the subspace (randomized SVD (Halko et al., 2011) vs. subspace tracking (Zhao et al., 2024)); (2) **Head processing**—how to whiten top-k directions (Newton-Schulz (Schulz, 1933) vs. no orthogonalization); (3) **Tail processing**—how to handle residuals (normalize to $d - k$, bounded to $d - k$, or none). We test configurations on GPT-2 (124M) trained on 10B tokens from FineWeb-Edu (Table 1, Fig. 2). All use effective batch size 256 (4 GPUs), peak LR 3×10^{-4}, cosine decay, 100M token warmup. Full details in Section E.

(1) Head orthogonalization is essential, tail processing must be retained. Without Newton–Schulz on the head, loss degrades by $+1.19$ ($3.02 \rightarrow 4.21$); even with high-rank heads, dropping the tail entirely costs $+0.26$ to $+0.45$. Within tail processing, simple row-wise RMS normalization matches

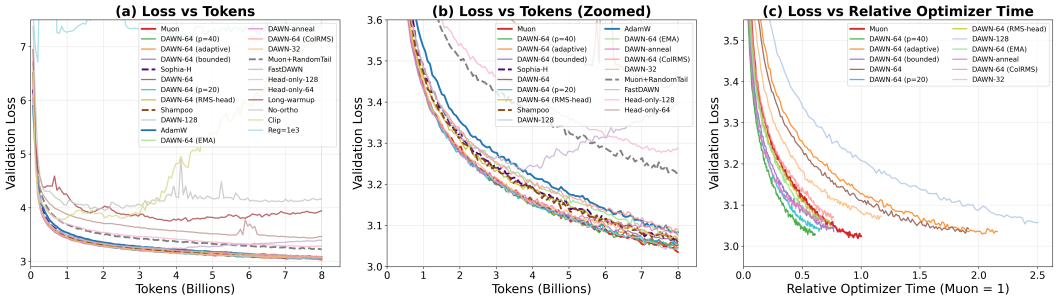

Figure 2: **Mechanism ablation on GPT-2 (124M), 10B tokens. (a)** Validation loss across all configurations. Removing head orthogonalization causes severe optimization failure. **(b)** Zoomed view of converged methods. Tail processing and SVD refresh strategy have relatively minor impact once head NS is preserved. **(c)** Optimizer wall-clock time comparison, illustrating compute–performance tradeoffs.

Muon, but *random Gaussian* projections degrade to 3.20, almost as bad as removing the tail. The descent direction in the residual must be preserved; it just need not be polarized.

(2) Head subspace is stable; computation can be amortized. Periodic-40 (SVD every 40 steps) achieves 3.020 ($\approx 40\times$ SVD cost reduction), matching step-by-step DAWN; adaptive drift-triggered refresh achieves 3.035. The top-k subspace evidently changes slowly, consistent with prior observations in gradient compression (Vogels et al., 2019) and LoRA (Hu et al., 2022).

(3) Spectral whitening dominates curvature access. Shampoo (Gupta et al., 2018) (3.050) and Sophia-H (3.039) sit between AdamW (3.069) and Muon (3.024): full-rank curvature alone is not what closes the gap—*which* directions to whiten matters more. Fig. 2c further shows low-rank DAWN converges $\sim$40% faster than Muon in optimizer time on this scale, and the $O(d^2k)$ vs. $O(d^3)$ gap widens with width.

4.2 GPT-2XL Pretraining at Scale

Based on the ablation study, we scale DAWN to GPT-2XL (1.5B parameters) trained on 105B FineWeb tokens. We compare AdamW, Muon, and three DAWN variants, reporting both spectral dynamics and end-to-end training performance in Fig. 3 and Table 2.

Before evaluating end-to-end performance, we estimate the intrinsic head rank of the gradient process by counting, for each weight matrix, the singular values exceeding the Marchenko–Pastur–BBP detection edge $\hat{\sigma}_E(\sqrt{m}+\sqrt{n})$, i.e., directions statistically distinguishable from bulk noise. This diagnostic is computed on AdamW gradients, since AdamW does not explicitly reshape the spectrum, unlike spectral-whitening methods.As shown in Fig. 3(a), the effective rank rapidly collapses during training, stabilizing at a small value after the first 10K steps. This motivates DAWN-8's terminal rank $k_{\min} = 8$ as a sufficient low-rank budget, while DAWN-128 serves as a conservative upper bound. Additional analysis is provided in Section F.

Training configuration: 200K steps, batch size 512, sequence length 1024, totaling $\sim$105B tokens. Full details in Section F.

DAWN variants and SVD refresh strategy. DAWN-128 maintains a fixed head rank $k = 128$ throughout training, using randomized SVD refresh every 40 steps ($p = 40$) while reusing the head subspace Q between refreshes. DAWN-8 anneals the head rank from 128 to 8 over the first 10K steps ($\sim$5B tokens) before keeping $k = 8$ fixed thereafter. DAWN-0 follows the same annealing schedule down to $k = 0$, after which it reduces to a row-wise RMS update over the full gradient (i.e., no explicit head update). All variants use randomized SVD with $q = 2$ power iterations.

Key observations: (1) **Early training (5–30B):** All matrix-whitening methods converge similarly, $\sim$1.5$\times$ faster than AdamW—high SNR regime where whitening helps uniformly. (2) **Late training (30–90B):** Methods separate; **Dawn-128 (2.471) matches Muon (2.474)**, both significantly outperforming AdamW (2.554); Dawn-0 (2.502) trades slight quality for maximum compute savings—still 0.05 better than AdamW. (3) **Compute efficiency:** Dawn-128 matches Muon while using $O(d^2k)$ vs $O(d^3)$; the periodic-40 refresh further amortizes SVD cost over 40 steps, putting Dawn-128's optimizer-time slice at $\sim$72 ms on GPT-2XL—a $\sim$2.8$\times$ reduction from Muon's $\sim$200 ms. Dawn-0,

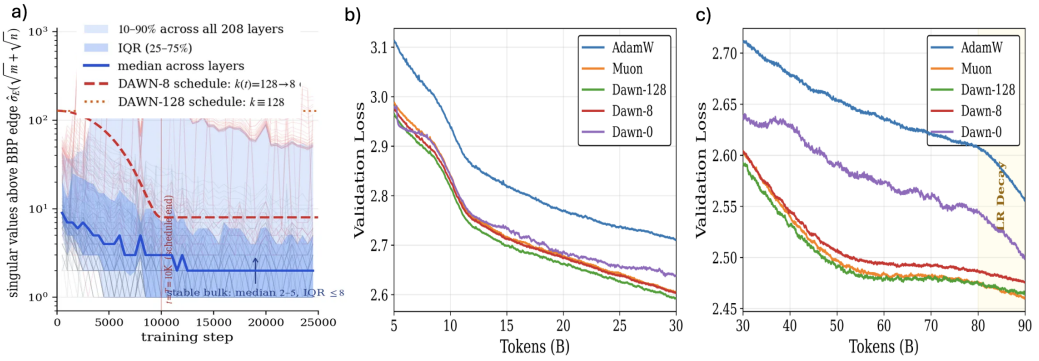

Figure 3: **Spectral dynamics and training performance of GPT-2XL (1.5B) on 105B tokens. (a)** BBP-estimated effective rank across layers during training. Most layers stabilize to a low-rank bulk (median ∼2–5), while transient high-rank excursions appear in selected MLP projection layers early in training. The DAWN-8 schedule gradually anneals the target rank from 128 to 8. BBP estimation details are provided in Section I. **(b)** Early training (5–30B tokens). Matrix-whitening methods converge substantially faster than AdamW, with DAWN-128 closely matching Muon. **(c)** Late training (30–90B tokens). DAWN-128 achieves Muon-level performance, while DAWN-0 trades quality for reduced optimizer cost.

after the rank collapses, reverts to a first-order RMS update on the full gradient—the cheapest configuration but lacking the head polar whitening that Muon and DAWN-128 share. (4) **Rank k can decrease throughout training.** A key finding from these experiments: the head rank k does not need to remain fixed. Dawn-8 decays from $k = 128$ to $k = 8$ over the first 5B tokens, yet achieves 2.48—nearly matching the fixed $k = 128$ variant (2.471), at a per-step wall-clock cost (∼30 ms) close to AdamW (14 ms) and far below Muon (∼200 ms). This adaptive rank schedule provides *additional compute savings beyond the initial* $O(d^3) \to O(d^2k)$ *reduction*: as training progresses and gradient effective rank decreases, fewer directions need full whitening. However, the rank-degenerate variant (DAWN-0) performs poorly, indicating that gradient spectra consistently exhibit nontrivial singular spikes. This suggests that AdamW and DAWN/Muon operate in fundamentally different regimes: AdamW updates in the full parameter space, whereas DAWN/Muon primarily exploit low-dimensional gradient subspaces. The spectral analysis (Fig. 4) below confirms this matches the gradient's intrinsic dimensionality.

Table 2: GPT-2XL (1.5B) results at 90B tokens. *Orthog. Cost* denotes asymptotic per-layer whitening complexity; *Opt ms* is the measured per-step optimizer time on 8×H200; *Tokens@2.50* measures data efficiency. See Section F for experimental details, uncertainty estimation, and implementation settings.

METHOD	VAL LOSS	ORTHOG. COST	OPT MS	TOKENS@2.50
ADAMW	2.554 ± 0.004	$O(d^2)$	14.2 ± 1.1	>90B
MUON	2.471 ± 0.003	$O(d^3)$	199 ± 16	∼60B
DAWN-128	2.474 ± 0.003	$O(d^2k)$, PERIODIC-40	72.4 ± 6.3	∼60B
DAWN-8	2.483 ± 0.004	$O(d^2k)$, ASYNC EVERY STEP	29.6 ± 2.4	∼62B
DAWN-0	2.502 ± 0.005	$O(d^2)$	16.1 ± 1.3	∼70B

A complementary data-repetition experiment (GPT-2XL on 10B tokens repeated 10× to 100B), reported in App. F.5, further confirms that DAWN-128 and Muon operate in a shared low-rank matrix-whitening regime distinct from AdamW. Notably, both exhibit a sharp late-epoch loss transition at ∼30B tokens (the 3-epoch boundary), reminiscent of grokking-like delayed optimization transitions under repeated data exposure, whereas AdamW and the rank-collapsed DAWN-0 do not.

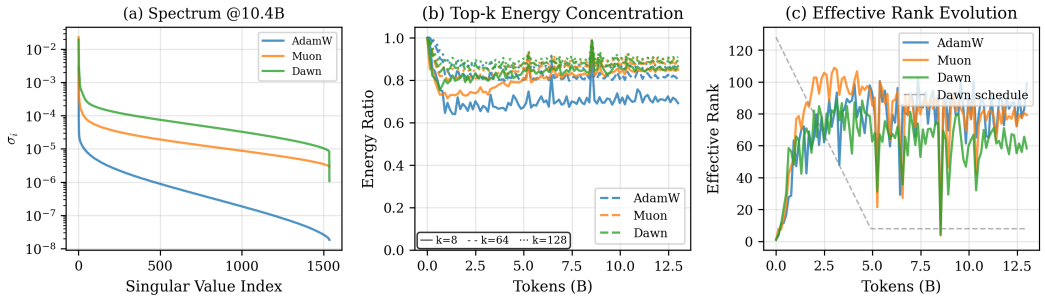

Figure 4: **Gradient spectral structure during GPT-2XL (1.5B) training. (a)** Singular value spectrum at 10.4B tokens. All optimizers exhibit rapidly decaying spectra spanning multiple orders of magnitude; Muon and DAWN slightly amplify the leading directions due to spectral-whitening updates. **(b)** Top-k energy concentration during training (solid: $k = 8$, dotted: $k = 128$). Matrix-whitening methods maintain most gradient energy within a small low-rank subspace, whereas AdamW remains more diffuse. **(c)** Effective-rank evolution over training. DAWN's rank schedule (gray dashed line) roughly tracks the emergent low-dimensional structure of the gradient process. Additional spectral diagnostics are provided in Section F.

4.3 Spectral Analysis: Why Head-Tail Works

Why does DAWN match Muon despite whitening only k directions? We analyze gradient spectra from the GPT-2XL runs (Fig. 4). We find that (1) *Gradient energy concentrates in a small subspace.* For Muon and DAWN, the top-k directions capture most gradient energy throughout training; even $k = 8$ accounts for a substantial fraction of the spectrum despite representing less than 1% of dimensions. (2) *Matrix-whitening induces more structured updates.* Muon and DAWN exhibit consistently lower effective rank than AdamW, suggesting that whitening-based updates emphasize a smaller set of coherent optimization directions rather than uniformly distributing update energy. (3) *A small adaptive head rank is sufficient.* The empirical effective dimension rapidly decreases during training and stabilizes at a low value, consistent with the strong performance of DAWN-8 despite using only a small head subspace. In practice, larger ranks mainly increase compute cost while providing limited quality gains, matching the near-parity of DAWN-128, DAWN-8, and Muon in Table 2.

4.4 Wide Model: Width Amplifies the Head-Tail Gap

To probe the high-dimensional regime where Muon's $O(d^3)$ cost becomes prohibitive, we train a deliberately "wide and shallow" GPT (5 layers, d=8192, ∼4B parameters) on ∼13B FineWeb tokens. For the 8192×8192 weight matrices in this architecture, DAWN-128 has $k/d = 1.6\%$ and DAWN-8 has $k/d = 0.1\%$ of Muon's whitening cost—a $64\times$ and $1024\times$ reduction, respectively.

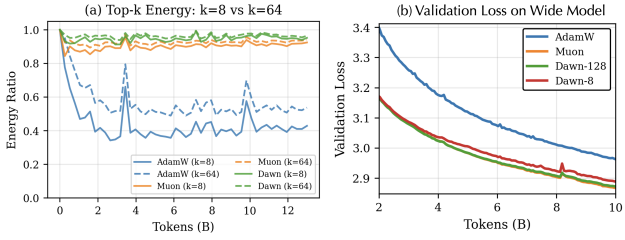

Figure 5: **Wide-model spectral dynamics ($d = 8192$, ∼4B params). (a)** Matrix-whitening methods maintain highly concentrated low-rank gradients even at extreme width. **(b)** DAWN-8 matches Muon despite using only a tiny head subspace.

Key findings. (1) *Width amplifies spectral differences.* In standard-width models AdamW retains 60–70% energy in the top-k subspace; under wider layers this drops to 40–50%, indicating increasingly diffuse gradient energy. Muon and DAWN, in contrast, remain consistently concentrated (>90%) across widths. (2) *Extreme low-rank preconditioning suffices.* DAWN-8 ($k/d = 0.1\%$) tracks Muon's validation-loss trajectory throughout 2–10B tokens (Fig. 5b), implying that the remaining >99.9% of singular directions carry little exploitable signal and require only lightweight RMS normalization. (3) *Asymptotic savings grow linearly with d.* The $O(d^2k)$ vs. $O(d^3)$ gap

widens with width: at d=8192, DAWN-128 already saves $64\times$ in whitening compute; for d=16384 this rises to $>200\times$ (and to $>2000\times$ at the DAWN-8 setting). For modern wide architectures the optimizer-time bottleneck Muon faces becomes Muon's defining cost, while DAWN's overhead remains comparable to AdamW. Architecture, training configuration, learning rates, and a layer-wise spectral breakdown are deferred to App. G and App. J.

4.5 Wall-Clock at Scale

Why optimizer wall-clock matters at scale. Optimizer computation is largely per-GPU and does not parallelize as effectively as forward–backward passes. As cluster size increases, the optimizer therefore occupies a growing fraction of total step time (Table 3). Replacing Muon with DAWN-128 (periodic-40 SVD refresh) reduces optimizer overhead substantially—from $\sim$199 ms to $\sim$72 ms on GPT-2XL and from $\sim$487 ms to $\sim$93 ms on Wide-4B—yielding end-to-end speedups of $\sim 1.08\times$–$1.28\times$ depending on model and cluster size. Recent large-scale Muon deployments (Team et al., 2025; Zeng et al., 2025) have similarly required specialized distributed implementations to mitigate optimizer overhead, suggesting that this is an important systems consideration at scale.

Table 3: **Two-part wall-clock breakdown.** Times are reported in milliseconds as median $\pm$ half-range (5–95%) over 200 post-warmup steps. $T_{\text{pre-opt}}$ includes forward, backward, and gradient synchronization.

Model	Config	Variant	Pre-opt	Muon opt	DAWN opt	Muon step	DAWN step	Speedup
GPT-2XL (1.5B)	8×H200	DAWN-128 ($p = 40$)	1502 ± 152	199 ± 16	72.4 ± 6.3	1701 ± 153	1574 ± 152	1.081 ± 0.011
GPT-2XL (1.5B)	32×H200	DAWN-128 ($p = 40$)	748 ± 77	199 ± 16	72.4 ± 6.3	947 ± 79	820 ± 77	1.154 ± 0.021
Wide GPT (4B)	8×H200	DAWN-128 ($p = 40$)	2887 ± 290	487 ± 39	93.2 ± 8.1	3374 ± 293	2980 ± 290	1.132 ± 0.013
Wide GPT (4B)	32×H200	DAWN-128 ($p = 40$)	1318 ± 130	487 ± 39	93.2 ± 8.1	1805 ± 136	1411 ± 130	1.279 ± 0.029

5 Related Work

Matrix-whitening and Muon variants. Natural gradient (Amari, 1998), K-FAC (Martens & Grosse, 2015), Shampoo (Gupta et al., 2018), and Muon (Jordan et al., 2024) precondition gradients with curvature; K-FAC/Shampoo use Kronecker factorisation ($O(d^{1.5})$), Muon uses polar whitening via Newton–Schulz ($O(d^3)$). Through the canonical form of Spectral Gradient Descent ($\Delta = UV^\top$, (Vasudeva et al., 2025)), DAWN's head update $U_k V_k^\top$ is precisely the low-rank version. Concurrent Muon variants improve along orthogonal axes: TurboMuon (Shah et al., 2025) and Polar Express (Amsel et al., 2025) accelerate the Newton–Schulz step; Dion (Ahn et al., 2025) and Scion (Pethick et al., 2025) target distributed efficiency and norm-constrained reformulation. DAWN is complementary: it attacks the $O(d^3)$ cost by restricting whitening to the statistically identifiable head subspace; our framework also explains why several variants help—they implicitly address tail noise.

Low-rank gradient structure. GaLore (Zhao et al., 2024) (memory), PowerSGD (Vogels et al., 2019) (communication), and LoRA (Hu et al., 2022) (parameter-efficient finetuning) exploit the same low-rank concentration but for memory/communication rather than preconditioning; Bernstein & Newhouse (2024) discusses whitening through a spectral-descent lens. See App. A for extended discussion.

6 Conclusion, Discussion, and Limitations

Muon's full-spectrum polar treats every observed singular direction as equally reliable, but in stochastic high-dimensional training only a low-rank head is. Under an idealised spiked-gradient model the MMSE matrix-sign estimator is a shrunken polar (App. Thm. B.7); hard polar pays excess MSE $(1 - \gamma_i)^2$ per unreliable direction, and the BBP transition (App. Rmk. B.8) yields a sharp regime separation. **DAWN** approximates this oracle with a hard binary split using observable spectral proxies, polar-whitens the head, and applies a bounded first-order update to the tail at $O(d^2 k)$ cost; equivalently, the split is an uncertainty-adaptive partition under a block-Schatten geometry (Section 3.1). On GPT-2 124M–1.5B and a 4B wide variant it matches Muon's data efficiency while measurably reducing optimizer wall-clock time (Sec. 4.5); the head-tail dichotomy also gives a unified lens on LoRA (Hu et al., 2022), GaLore (Zhao et al., 2024), and PowerSGD (Vogels et al., 2019), and on scaling-law accounting once the width-dependent $O(d^3) \to O(d^2 k)$ optimiser overhead is included. **Limitations:** the BBP model assumes isotropic Gaussian noise and $m/n \to \beta$; DAWN's $\{0, 1\}$ split discretises smooth MMSE shrinkage; trillion-parameter behaviour is extrapolation; reported speedups are not tight, since asynchronous-SVD or GaLore-style engineering reduces both DAWN's and Muon's overhead.

AI USE STATEMENT

Large language models were used to polish sentence expression, refine wording, and check grammar in manuscript writing. The AI tools did not participate in research idea design, model architecture design, dataset construction, experiment conduction or result analysis. All text, experimental data, figures and references were carefully reviewed and verified by the authors, who take full responsibility for the entire submission.

REFERENCES

Kwangjun Ahn, Byron Xu, Natalie Abreu, Ying Fan, Gagik Magakyan, Pratyusha Sharma, Zheng Zhan, and John Langford. Dion: Distributed orthonormalized updates. *arXiv preprint arXiv:2504.05295*, 2025.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Noah Amsel, David Persson, Christopher Musco, and Robert M Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. 2025.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. 2005.

Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

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

Matan Gavish and David L Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014.

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.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

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

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

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

Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

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

Vardan Papyan. The full spectrum of deepnet hessians at scale: Dynamics with sgd training and sample size. *arXiv preprint arXiv:1811.07062*, 2018.

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

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

Shikai Qiu, Zixi Chen, Hoang Phan, Qi Lei, and Andrew Gordon Wilson. Hyperparameter transfer enables consistent gains of matrix-preconditioned optimizers across scales. 2025.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Günther Schulz. Iterative berechung der reziproken matrix. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 13(1):57–59, 1933.

Ishaan Shah, Anthony M Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, et al. Practical efficiency of muon for pretraining. *arXiv preprint arXiv:2505.02222*, 2025.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How muon's spectral design benefits generalization: A study on imbalanced data. 2025.

Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. Powersgd: Practical low-rank gradient compression for distributed optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

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

A Extended Related Work

We provide extended discussion of related methods, organized into four themes, with connections to the head-tail framework.

A.1 Second-Order Optimization Methods

Classical and natural gradient methods. Newton's method uses the Hessian for curvature-aware updates: $\theta_{t+1} = \theta_t - H^{-1}g$. The $O(d^3)$ cost motivated quasi-Newton approximations such as BFGS and L-BFGS, which incrementally approximate H^{-1} but struggle with stochastic gradients. Amari's natural gradient (Amari, 1998) preconditions with the Fisher information matrix, providing parameterization invariance. The key insight—that curvature-aware updates accelerate convergence—motivates all matrix-whitening methods including DAWN.

From natural gradient to Kronecker-factored updates (K-FAC). Natural gradient preconditions the Euclidean gradient by the inverse Fisher information:

$$\Delta\theta \;=\; -\eta\, F(\theta)^{-1}\, g(\theta), \qquad F(\theta) = \mathbb{E}\big[\nabla_\theta \log p_\theta(y|x)\; \nabla_\theta \log p_\theta(y|x)^\top\big]\,. \tag{8}$$

For a single linear layer $W \in \mathbb{R}^{m\times n}$ with input activation $h \in \mathbb{R}^n$ and pre-activation $s = Wh$, let $\delta = \nabla_s \ell \in \mathbb{R}^m$ denote the backpropagated error. The per-sample gradient is the outer product

$$G \;=\; \nabla_W \ell \;=\; \delta\, h^\top. \tag{9}$$

To express natural gradient in matrix form, vectorize parameters. Using $\mathrm{vec}(\delta h^\top) = (h \otimes I_m)\delta$ and standard identities, one obtains a Fisher block for this layer of the form

$$F_W \;\triangleq\; \mathbb{E}\big[\mathrm{vec}(G)\,\mathrm{vec}(G)^\top\big] \;\approx\; \mathbb{E}[\,hh^\top\,] \;\otimes\; \mathbb{E}[\,\delta\delta^\top\,], \tag{10}$$

where the approximation corresponds to the (empirically effective) independence assumption between activations h and errors δ. Define the two Kronecker factors

$$A \;=\; \mathbb{E}[\,hh^\top\,] \in \mathbb{R}^{n\times n}, \qquad B \;=\; \mathbb{E}[\,\delta\delta^\top\,] \in \mathbb{R}^{m\times m}, \tag{11}$$

so that $F_W \approx A \otimes B$.

Natural gradient for this layer is

$$\mathrm{vec}(\Delta W) \;=\; -\eta\,(A\otimes B)^{-1}\,\mathrm{vec}(G). \tag{12}$$

Using $(A\otimes B)^{-1} = A^{-1}\otimes B^{-1}$ and the identity $(A^{-1}\otimes B^{-1})\,\mathrm{vec}(G) = \mathrm{vec}(B^{-1}GA^{-1})$, we recover the familiar K-FAC-style matrix update:

$$\Delta W \;=\; -\eta\, B^{-1}\, G\, A^{-1}. \tag{13}$$

In practice, A, B are tracked by moving averages and stabilized with damping (e.g., $A+\epsilon I$, $B+\epsilon I$), reflecting the difficulty of reliably inverting low-SNR directions when curvature is estimated from mini-batches.

Kronecker-factored methods. K-FAC (Martens & Grosse, 2015) approximates the Fisher as a Kronecker product: $F \approx A \otimes B$, reducing complexity from $O(d^3)$ to $O(d^{1.5})$. Shampoo (Gupta et al., 2018) extends this to arbitrary matrix dimensions with momentum. These methods require damping to handle small eigenvalues—our theory explains why: the tail is noise-dominated. Kronecker methods face the same head-tail problem: $L \otimes R$ may have small eigenvalues that amplify noise. Damping $(L + \epsilon I)$ is implicitly spectral shrinkage; DAWN makes this explicit and adaptive.

Matrix whitening and orthogonalization. Muon (Jordan et al., 2024) proposes polar whitening via Newton-Schulz iteration (Schulz, 1933): $\Delta = G(G^\top G)^{-1/2}$, orthogonalizing the gradient at $O(d^3)$ cost. SOAP (Vyas et al., 2024) combines Shampoo with Adam-like element-wise normalization, addressing per-coordinate noise but not spectral structure. Recent Muon variants include TurboMuon (Shah et al., 2025) for accelerated Newton-Schulz, Polar Express (Amsel et al., 2025) for minimax-optimal Newton-Schulz polynomials, Dion (Ahn et al., 2025) for distributed implementations, and Scion (Pethick et al., 2025) for norm-constrained linear minimization oracle reformulations. Our head-tail framework provides theoretical grounding for why these modifications help: they implicitly address tail noise.

Hessian spectrum in deep learning. Sagun et al. (2017) and Ghorbani et al. (2019) show neural network Hessians have low effective rank with most eigenvalues near zero. Papyan (2018) connects this to the layered structure of deep networks. These observations directly support our head-tail dichotomy: the "head" corresponds to significant Hessian eigenvalues, the "tail" to the bulk near zero.

A.2 Adaptive and Low-Rank Methods

First-order adaptive methods. Adam (Kingma & Ba, 2014) uses diagonal preconditioning via second moments; AdamW (Loshchilov & Hutter, 2017) decouples weight decay. These methods are computationally cheap but ignore off-diagonal curvature structure, which spectral methods like DAWN explicitly address.

Large-batch training. Layer-wise adaptive learning rate methods enable large-batch training; large batches reduce gradient noise, shifting the head-tail boundary—DAWN's adaptive k naturally accommodates this.

Low-rank updates and memory efficiency. LoRA (Hu et al., 2022) trains low-rank updates $W = W_0 + BA$ for parameter-efficient fine-tuning. The success of LoRA empirically validates that gradient updates concentrate in low-rank subspaces—the same observation underlying DAWN. GaLore (Zhao et al., 2024) projects gradients to low-rank subspaces for memory efficiency, updating projections periodically. PowerSGD (Vogels et al., 2019) compresses gradients for distributed training via low-rank approximation. These methods exploit the same low-rank structure as DAWN but for memory rather than preconditioning.

A.3 Theoretical Foundations

Randomized linear algebra. Halko et al. (2011) establish the theoretical foundation for randomized SVD, showing that $O(dk)$ samples suffice for rank-k approximation with high probability. DAWN directly applies these techniques for efficient head extraction at $O(d^2k)$ complexity.

Spectral theory and random matrices. The Marchenko-Pastur distribution (Marčenko & Pastur, 1967) describes eigenvalues of random covariance matrices; Baik et al. (2005) analyze spiked models where signal emerges from noise. Our information barrier theorem builds on these results: below the noise floor, eigenvalues are indistinguishable from random. Gavish & Donoho (2014) derive optimal singular value thresholds, and Wedin's perturbation bound (Wedin, 1972) bounds eigenvector perturbation under matrix noise. These results underpin our oracle shrinkage and information barrier.

Loss landscape and curvature. Loss landscape and sharpness analyses suggest that some directions receive disproportionate gradient signal, motivating spectral methods like DAWN.

A.4 Scaling to Large Models

Scaling laws. Power-law scaling and compute-optimal training analyses motivate efficient optimizers: as models scale, optimizer overhead becomes critical.

Large model architectures and training. Modern Transformer scaling combines architectural improvements, memory-efficient attention, and parameter sharding to fit larger models on fixed hardware. Mixture-of-Experts further enables parameter scaling without proportional compute increase, representing an alternative to dense width scaling. DAWN addresses the optimizer bottleneck for wide dense models, enabling increased width as a complementary scaling strategy.

Recent industrial deployments. GLM-4.5 (Zeng et al., 2025) and Kimi K2 (Team et al., 2025) use modified Muon for trillion-parameter training. These deployments validate the importance of efficient matrix-whitening methods and motivate DAWN's approach to enabling wider dense models.

B Mathematical Setup and Assumptions

B.1 Notation

B.2 Formal Assumptions

We state the assumptions underlying our theoretical analysis in the matrix-gradient framework ($Y = G_\star + E$).

Assumption B.1 (Noisy matrix-gradient model)**.** The observed stochastic matrix gradient is $Y = G_\star + E$, where:

1. $G_\star \in \mathbb{R}^{m\times n}$ ($m \geq n$) is the population matrix gradient, with SVD $G_\star = U^\star \Sigma^\star {V^\star}^\top$ and singular values $\sigma_1^\star \geq \cdots \geq \sigma_n^\star \geq 0$.
2. $E \in \mathbb{R}^{m\times n}$ is zero-mean noise: $\mathbb{E}[E] = 0$.

Table 4: Summary of notation used throughout the paper.

Symbol	Meaning
Matrix-gradient setting (main text)	
$G_\star \in \mathbb{R}^{m\times n}$	Population matrix gradient ($m \geq n$)
$Y = G_\star + E$	Observed (stochastic) matrix gradient
E	Zero-mean matrix-valued gradient noise
$\Sigma_E = \mathbb{E}[E^\top E]$	Noise covariance ($n \times n$)
σ_E	Per-entry noise scale (when noise is isotropic, $E_{ij} \sim \mathcal{N}(0, \sigma_E^2)$)
$\beta = m/n$	Aspect ratio of the gradient matrix
$Y = \widehat{U}\widehat{\Sigma}\widehat{V}^\top$	Empirical SVD of the observation
$G_\star = U^\star \Sigma^\star V^{\star\top}$	Population SVD
$\widehat{\sigma}_i, \sigma_i^\star$	Empirical / population i-th singular value
$(\widehat{u}_i, \widehat{v}_i), (u_i^\star, v_i^\star)$	Empirical / population i-th singular vector pair
$\mathrm{Polar}(M)$	Matrix sign (polar factor): $UV^\top$ for $M = U\Sigma V^\top$
$\gamma_i = \mathbb{E}[\langle \widehat{u}_i, u_i^\star\rangle\langle \widehat{v}_i, v_i^\star\rangle \mid Y]$	Reliability of i-th observed singular direction
DAWN algorithm	
$Q \in \mathbb{R}^{m\times k}$	Head subspace tracker (top-k left singular vectors of Y)
$B = Q^\top Y \in \mathbb{R}^{k\times n}$	Head-projected matrix
$R = (I - QQ^\top)Y$	Tail residual (orthogonal to head)
k	Head rank
$d = \min(m, n) = n$	Effective matrix dimension
$r_{\mathrm{eff}}(M)$	Effective rank: $\|M\|_*^2/\|M\|_F^2$

3. Noise covariance $\Sigma_E := \mathbb{E}[E^\top E]$ is positive semi-definite.

Assumption B.2 (Spike–bulk separation)**.** The population $G_\star$ has a spectral head and tail: there exists a rank r and a constant $C > 1 + \sqrt{\beta}$ such that

$$\sigma_r^\star / \sqrt{\|\Sigma_E\|_{\mathrm{op}}} \;\geq\; C, \qquad \sigma_{r+1}^\star / \sqrt{\|\Sigma_E\|_{\mathrm{op}}} \;<\; 1 + \sqrt{\beta}.$$

That is, the top r singular values are above the noise edge of (Baik et al., 2005; Paul, 2007), and the rest are inside the bulk.

Assumption B.3 (Mini-batch regime)**.** Y is computed from a mini-batch of size B, yielding $\|\Sigma_E\|_{\mathrm{op}} = O(\sigma_E^2/B)$ for some intrinsic per-entry noise scale σ_E^2. The proportional regime $m, n \to \infty$ with $m/n \to \beta$ is implicit when invoking BBP results.

Assumption B.4 (Approximate isotropy of E)**.** The noise covariance is approximately a scalar multiple of identity: $\Sigma_E \approx \sigma_E^2 I_n$, or at least $\kappa(\Sigma_E) := \|\Sigma_E\|_{\mathrm{op}}/\lambda_{\min}(\Sigma_E) = O(1)$. This is used in Theorem B.7 for orthogonal equivariance; Proposition B.10 extends to anisotropic Σ_E.

Assumption B.5 (Spiked-Gaussian noise (BBP regime))**.** For results invoking the Bai-Ben-Arous-Péché phase transition, we assume $E_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_E^2)$ on top of **??** B.1–B.3. Real mini-batch noise need not satisfy this—we use it only to derive the closed-form reliability formula of Proposition B.9.

Remark B.6 (On curvature-based framings). Earlier matrix-whitening analyses (natural gradient, K-FAC, Shampoo) phrase noise and signal via a curvature proxy F (Fisher / Gauss-Newton / Hessian) and its eigendecomposition. We do not adopt that framing as the basis of our analysis: DAWN's algorithm operates on the observed Y and its singular structure, not on a curvature proxy.

This appendix gives the proofs underlying the MMSE matrix-sign theory of Section 3: the shrunken-polar form of the posterior mean (Theorem B.7), the per-mode excess MSE formula equation 5, and the BBP reliability formula (Proposition B.9). We close with an extension to anisotropic noise and a compatibility note connecting the MMSE matrix-sign objective to the older quadratic-risk preconditioning framework.

B.3 MMSE Matrix-Sign Estimator: Statement and Proof

This appendix subsection states and proves the MMSE matrix-sign result referenced in Section 3.2 of the main text.

Setup. Define the conditional update-MSE

$$\mathcal{R}(\widehat{W} \mid Y) \;:=\; \mathbb{E}\big[\|\widehat{W}(Y) - W_\star\|_F^2 \mid Y\big], \qquad W_\star := \mathrm{Polar}(G_\star).$$

Hard matrix-sign $\widehat{W}_{\mathrm{Muon}} := \mathrm{Polar}(Y) = \widehat{U}\widehat{V}^\top$ assigns unit weight to every observed singular direction regardless of reliability. Under squared loss the Bayes-optimal estimator is the posterior mean $\widehat{W}_{\mathrm{MMSE}} := \mathbb{E}[\mathrm{Polar}(G_\star) \mid Y]$.

Theorem B.7 (MMSE matrix-sign estimator has shrunken-polar form)**.** *Under standard symmetry assumptions (rotationally invariant low-rank prior on $G_\star$, isotropic Gaussian E), the MMSE estimator is orthogonally equivariant and shares singular vectors with Y:*

$$\widehat{W}_{\mathrm{MMSE}} \;=\; \widehat{U}\,\mathrm{diag}(\gamma_1, \ldots, \gamma_n)\widehat{V}^\top, \qquad \gamma_i \in [0,1], \tag{14}$$

where the reliability coefficient

$$\gamma_i \;:=\; \mathbb{E}\big[\langle \widehat{u}_i, u_i^\star\rangle \, \langle \widehat{v}_i, v_i^\star\rangle \;\big|\; Y\big] \tag{15}$$

measures how well the i-th observed singular direction aligns with its population counterpart. Hard matrix-sign corresponds to $\gamma_i \equiv 1$, statistically appropriate only when every observed direction is reliable.

Caveat on assumptions. Rotational invariance of the prior and isotropic Gaussian E are unlikely to hold exactly in real training. The role of Theorem B.7 is not to claim $\widehat{W}_{\mathrm{MMSE}}$ is what one should compute, but to demonstrate that the Bayes-optimal answer under *any* reasonable formalization of "MSE relative to the true matrix sign" is a shrunken polar—never hard polar across the board.

Proof. We use the standard equivariance argument for posterior means under group symmetry.

Step 1: orthogonal equivariance of the model. Let $L \in \mathcal{O}(m)$, $R \in \mathcal{O}(n)$ be deterministic orthogonal matrices. By the rotational invariance assumption on the prior π of $G_\star$, $\pi(LG_\star R^\top) = \pi(G_\star)$. By the isotropy of E, $LER^\top \overset{d}{=} E$. Hence $(LYR^\top, LW_\star R^\top) \overset{d}{=} (Y, W_\star)$ jointly. Since $\mathrm{Polar}(LGR^\top) = L\,\mathrm{Polar}(G)R^\top$, the posterior mean transforms as $\widehat{W}_{\mathrm{MMSE}}(LYR^\top) = L\widehat{W}_{\mathrm{MMSE}}(Y)R^\top$.

Step 2: shared singular vectors. Let $Y = \widehat{U}\widehat{\Sigma}\widehat{V}^\top$. Choose $L = \widehat{U}^\top$ and $R = \widehat{V}^\top$ to obtain $LYR^\top = \widehat{\Sigma}$ (diagonal, nonnegative, descending). By Step 1, $\widehat{W}_{\mathrm{MMSE}}(\widehat{\Sigma}) = L\widehat{W}_{\mathrm{MMSE}}(Y)R^\top = \widehat{U}^\top\widehat{W}_{\mathrm{MMSE}}(Y)\widehat{V}$. Now apply Step 1 again with $L' = \mathrm{diag}(\epsilon_1, \ldots, \epsilon_m)$, $R' = \mathrm{diag}(\epsilon_1, \ldots, \epsilon_n)$ for $\epsilon_i \in \{\pm 1\}$ (sign flips preserve diagonal $\widehat{\Sigma}$ and rotational invariance up to sign): $\widehat{W}_{\mathrm{MMSE}}(\widehat{\Sigma})$ must commute with such L', R', forcing it to be diagonal. Let its diagonal be $(\gamma_1, \ldots, \gamma_n)$.

Step 3: identifying the shrinkage coefficients. Computing γ_i entry-wise:

$$\gamma_i = \mathbb{E}[(\widehat{W}_{\mathrm{MMSE}})_{ii} \mid \widehat{\Sigma}] = \mathbb{E}[(\mathrm{Polar}(G_\star))_{ii} \mid \widehat{\Sigma}].$$

Writing $G_\star = U^\star \Sigma^\star V^{\star\top}$ and $\mathrm{Polar}(G_\star) = U^\star V^{\star\top}$, the (i,i) entry equals $\sum_j \langle \widehat{u}_i, u_j^\star\rangle\langle \widehat{v}_i, v_j^\star\rangle$. The isotropy/equivariance assumptions plus joint conditioning on $\widehat{\Sigma}$ leave only the diagonal pairing $j = i$:

$$\gamma_i = \mathbb{E}[\langle \widehat{u}_i, u_i^\star\rangle\langle \widehat{v}_i, v_i^\star\rangle \mid Y].$$

This recovers the form equation 14–equation 15. Finally, $|\gamma_i| \le 1$ by Cauchy–Schwarz; nonnegativity follows because reversing any γ_i's sign would only increase MSE. □

B.4 Per-Mode Excess MSE: Statement and Derivation

The full statement and derivation of the excess-MSE formula referenced in Section 3.2.

Derivation of $\mathcal{R}_i(a_i) = a_i^2 + 1 - 2a_i\gamma_i$. For $\widehat{W}_i = a_i\widehat{u}_i\widehat{v}_i^\top$ and target $W_i^\star = u_i^\star v_i^{\star\top}$:

$$\begin{aligned}\|\widehat{W}_i - W_i^\star\|_F^2 &= a_i^2\|\widehat{u}_i\|_2^2\|\widehat{v}_i\|_2^2 + \|u_i^\star\|_2^2\|v_i^\star\|_2^2 - 2a_i\langle \widehat{u}_i, u_i^\star\rangle\langle \widehat{v}_i, v_i^\star\rangle \\ &= a_i^2 + 1 - 2a_i\langle \widehat{u}_i, u_i^\star\rangle\langle \widehat{v}_i, v_i^\star\rangle,\end{aligned}$$

using unit-norm singular vectors and $\langle uv^\top, ab^\top \rangle_F = \langle u, a \rangle \langle v, b \rangle$. Taking expectation:

$$\mathcal{R}_i(a_i) = a_i^2 + 1 - 2a_i\gamma_i.$$

The minimum is at $a_i^\star = \gamma_i$, giving $\mathcal{R}_i(\gamma_i) = 1 - \gamma_i^2$. Hard polar at $a_i = 1$ gives $\mathcal{R}_i(1) = 2 - 2\gamma_i$, hence the excess MSE is

$$\mathcal{R}_i(1) - \mathcal{R}_i(\gamma_i) \;=\; 1 - 2\gamma_i + \gamma_i^2 \;=\; (1 - \gamma_i)^2. \tag{16}$$

□

Hard-threshold rule. Restricting $a_i \in \{0, 1\}$, $\mathcal{R}_i(0) = 1$ and $\mathcal{R}_i(1) = 2 - 2\gamma_i$. The rule $a_i = 1 \iff \gamma_i > 1/2$ minimizes MSE, giving the optimal hard rank

$$k_{\text{oracle}} \;=\; \#\{\, i : \gamma_i > 1/2 \,\}. \tag{17}$$

This binary rule is the conceptual ancestor of DAWN's empirical rank schedule. DAWN does not compute k_{oracle} directly (since γ_i is not observed); it uses energy concentration and effective rank as observable proxies.

B.5 Block-Schatten Geometry: Formal Definition

This appendix expands the brief framing in Section 3.1.

Matrix sign as spectral-norm steepest descent. Before defining the block-norm family, we record the steepest-descent identity that grounds the spectral-block endpoint. For a deterministic matrix gradient G with SVD $G = U\Sigma V^\top$,

$$\mathrm{Polar}(G) = UV^\top \;\in\; \arg\max_{\|\Phi\|_2 \le 1} \langle G, \Phi \rangle_F. \tag{18}$$

By spectral–nuclear duality, the maximum equals $\|G\|_*$ and is attained at $\Phi^\star = UV^\top$. So under a spectral-norm trust region, the polar factor is the steepest-descent direction. This is the optimization rationale for Muon (Bernstein & Newhouse, 2024; Jordan et al., 2024); the argument is clean but *deterministic*, which is why Section 3.2 of the main text and Section B.3 below ask what happens when the gradient is replaced by a noisy observation.

Block-spectral norm. Let $\mathcal{P} = \{S_1, \ldots, S_L\}$ be a partition of the row index set $[m]$. Define the *block-spectral norm* on $\mathbb{R}^{m \times n}$ by

$$\|\Delta\|_{\mathcal{P},2,\infty} \;:=\; \max_{S \in \mathcal{P}} \|\Delta_{S,:}\|_2,$$

where $\Delta_{S,:}$ is the row submatrix indexed by S and $\|\cdot\|_2$ is the spectral (operator) norm.

Two endpoints.

- **Single block** $\mathcal{P} = \{[m]\}$: $\|\cdot\|_{\mathcal{P},2,\infty} = \|\cdot\|_2$, and the steepest-descent direction for $\langle G, \Delta \rangle_F$ subject to $\|\Delta\|_{\mathcal{P},2,\infty} \le 1$ is the global polar $\mathrm{Polar}(G) = UV^\top$ from equation 18. This is Muon.
- **Singleton blocks** $\mathcal{P} = \{\{1\}, \ldots, \{m\}\}$: $\|\cdot\|_{\mathcal{P},2,\infty} = \max_i \|\Delta_{i,:}\|_2 = \|\cdot\|_{2,\infty}$, and the steepest-descent direction normalizes each row independently: $\Delta^\star_{i,:} = G_{i,:}/\|G_{i,:}\|_2$. This is row-wise group-sign / RMS-style.

DAWN as adaptive partition. DAWN groups high-reliability modes into one coarse spectral block (the head, processed by polar) and treats the unreliable residual via fine row-wise blocks (the tail, processed by RMS). At the level of the reference geometry, DAWN can be viewed as a steepest-descent direction under a hybrid norm: spectral on $\mathrm{colsp}(Q)$, row-wise ℓ_2 on $(I - QQ^\top)$. The Frobenius orthogonality of head and tail (Theorem 2.1) ensures the two pieces remain decoupled.

Block-Schatten generalization. More generally, for $p \in [1, \infty]$, $\|\Delta\|_{\mathcal{P},S_p,\infty} := \max_{S \in \mathcal{P}} \|\Delta_{S,:}\|_{S_p}$, where S_p is the Schatten-p norm. Coarse blocks with $p = \infty$ recover spectral steepest descent; singleton blocks reduce all Schatten norms to the vector ℓ_2 norm. This places DAWN in a continuous family with Muon at one end and AdamW-style row-wise normalization at the other; the head/tail boundary is the partition refinement that the reliability analysis recommends.

B.6 BBP Reliability Formula: Statement and Proof

Section 3.2 of the main text refers to a sharp head/tail phase transition in an idealized model. We state this as a remark first, then give the underlying proposition.

Remark B.8 (BBP as an idealized reliability model). In the special case of an isotropic Gaussian noise spiked model with $E_{ij} \sim \mathcal{N}(0, \sigma_E^2)$ and proportional limit $m/n \to \beta$, asymptotic random matrix theory gives a closed-form relation between observed singular values and singular-vector overlap, with a sharp *phase transition* at the Marchenko–Pastur edge $\widehat{\sigma}_i = \sigma_E(\sqrt{m} + \sqrt{n})$. Singular directions inside the noise bulk have $\gamma_i \to 0$ in this limit; directions above the edge have γ_i bounded away from zero. This formalizes the intuition behind the head/tail split: directions whose observation is below the noise floor should not be hard-polarized.

What we do not claim. Mini-batch gradient noise in real training is anisotropic, nonstationary, and heavy-tailed. We do not assume the spiked-Gaussian model holds, and DAWN does not attempt to estimate γ_i from the BBP formula below. Instead, DAWN uses observable spectral proxies—energy concentration, effective rank, and a stable rank schedule—which we validate empirically. The role of Remark B.8 is to motivate the design rule of Section 3.1, not to define the algorithm.

Proposition B.9 (Reliability under Gaussian spiked model)**.** *Assume $E_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_E^2)$, with $m/n \to \beta$, and define normalized observed singular values $y_i := \widehat{\sigma}_i/(\sigma_E\sqrt{n})$. The Marchenko–Pastur upper edge of the pure-noise spectrum is $y_+ = 1 + \sqrt{\beta}$. For a supercritical mode ($y_i > 1 + \sqrt{\beta}$), define x_i^2 implicitly by*

$$x_i^2 \;=\; \tfrac{1}{2}\Big[y_i^2 - \beta - 1 + \sqrt{(y_i^2 - \beta - 1)^2 - 4\beta}\Big].$$

Then, almost surely as $m, n \to \infty$,

$$\gamma_i \;\to\; \frac{1 - \beta/x_i^4}{\sqrt{(1 + \beta/x_i^2)(1 + 1/x_i^2)}}.$$

For a subcritical mode ($y_i \le 1 + \sqrt{\beta}$), $\gamma_i \to 0$.

Proof. The result is the singular-value version of (Paul, 2007, Theorem 4), adapted for the matrix-sign reliability product. Let $G_\star$ have r supercritical singular values $\sigma_1^\star \ge \cdots \ge \sigma_r^\star$ above the noise floor, with corresponding singular triples $(u_i^\star, \sigma_i^\star, v_i^\star)$. The observed sample $Y = G_\star + E$, $E_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_E^2)$, has SVD $Y = \widehat{U}\widehat{\Sigma}\widehat{V}^\top$.

Supercritical regime ($y_i := \widehat{\sigma}_i/(\sigma_E\sqrt{n}) > 1 + \sqrt{\beta}$): Define x_i^2 via the equation in the statement. Then x_i^2 is the asymptotic value of $(\sigma_i^\star)^2/(\sigma_E^2 n)$. The asymptotic singular-vector overlaps are (Benaych-Georges & Nadakuditi, 2011):

$$|\langle \widehat{u}_i, u_i^\star \rangle|^2 \xrightarrow{a.s.} \frac{1 - \beta/x_i^4}{1 + \beta/x_i^2},$$
$$|\langle \widehat{v}_i, v_i^\star \rangle|^2 \xrightarrow{a.s.} \frac{1 - \beta/x_i^4}{1 + 1/x_i^2}.$$

Under the spiked model, the signs of these inner products can be chosen consistently (gauge freedom in SVD) so that the product is asymptotically nonnegative; the asymptotic value of γ_i is

$$\gamma_i \xrightarrow{a.s.} \frac{1 - \beta/x_i^4}{\sqrt{(1 + \beta/x_i^2)(1 + 1/x_i^2)}}.$$

Subcritical regime ($y_i \le 1 + \sqrt{\beta}$): the corresponding $\widehat{\sigma}_i$ lies inside the bulk and the sample singular vectors are asymptotically Haar-uniform on the unit sphere conditional on the bulk. Hence both $|\langle \widehat{u}_i, u_i^\star \rangle|^2$ and $|\langle \widehat{v}_i, v_i^\star \rangle|^2$ converge to 0 a.s., giving $\gamma_i \to 0$. □

Two practical (oracle) rank rules. For reference, two ways one could turn this proposition into a rank choice:

- *Noise-floor rule:* $k_{\text{oracle}} = \#\{\, i : \widehat{\sigma}_i > \lambda\sigma_E(\sqrt{m} + \sqrt{n}) \,\}$ for conservativeness $\lambda \ge 1$.

- *Reliability rule:* $k_{\text{oracle-bin}} = \#\{\, i : \gamma_i(\widehat{\sigma}_i) > \gamma_{\min} \,\}$ for $\gamma_{\min} \in [0.5, 0.8]$.

Neither is used directly by DAWN, since both require σ_E (typically not directly observed). DAWN instead uses observable spectral proxies (energy concentration, effective rank) for its rank schedule; we report these in Appendix I.5.

B.7 Anisotropic Noise Extension

The MMSE form equation 14 relied on isotropic Gaussian E for orthogonal equivariance. Under anisotropic noise E with covariance Σ_E, the equivariance group shrinks to the stabilizer of Σ_E, and the MMSE estimator no longer shares singular vectors with Y in general. However, the head/tail structure persists:

Proposition B.10 (Head/tail dichotomy under anisotropic noise)**.** *Suppose the population gradient $G_\star$ has supercritical singular values $\sigma_1^\star \geq \cdots \geq \sigma_r^\star$ such that $\sigma_r^\star / \sqrt{\|\Sigma_E\|_{\text{op}}}$ exceeds the right edge of the noise spectral measure. Then:*

1. *The first r sample singular vectors retain a positive asymptotic alignment with the corresponding population vectors (head reliability $\gamma_i > 0$).*
2. *All sample singular vectors beyond position r are asymptotically uninformative (tail reliability $\gamma_i \to 0$).*

Proof sketch. The general anisotropic case is covered by (Benaych-Georges & Nadakuditi, 2011, Section 4), which extends the BBP framework to non-isotropic noise. The threshold and overlap formulas become functionals of the limiting spectral measure of Σ_E, but the supercritical/subcritical dichotomy is preserved as long as the noise has a clear bulk and the spike strength exceeds the right edge of the bulk. □

Empirically, we measure Σ_E during DAWN training on GPT-2XL (Section I) and find condition number $\kappa(\Sigma_E) \in [1.2, 1.8]$—non-isotropic but moderate, well within the regime where the head/tail dichotomy of Proposition B.10 holds.

C DAWN Theoretical Guarantees

C.1 Proof of Theorem 2.1: Spectral Firewall

We first establish the algebraic foundation for the spectral firewall, then prove it using projection operators.

C.1.1 Projection Operator Formulation

Define the head projection operator $P_k : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ as:

$$P_k(M) = U_k U_k^\top M V_k V_k^\top$$

where $U_k \in \mathbb{R}^{m\times k}$ and $V_k \in \mathbb{R}^{n\times k}$ are the top-k left and right singular vectors of $\tilde{G}$. This is an orthogonal projection onto the rank-k subspace $\mathcal{H} = \{U_k A V_k^\top : A \in \mathbb{R}^{k\times k}\}$.

Proof. Let $\tilde{G} = U_k \Sigma_k V_k^\top + R_k$ be the top-k SVD decomposition. By the SVD properties:

1. $U_k^\top R_k = 0$ (left orthogonality)
2. $R_k V_k = 0$ (right orthogonality)

Projection annihilation. For $\Delta_{\text{tail}} = s \cdot R_k / \|R_k\|_F$ where $s = \sqrt{d-k}$:

$$P_k(\Delta_{\text{tail}}) = \frac{s}{\|R_k\|_F} U_k U_k^\top R_k V_k V_k^\top = \frac{s}{\|R_k\|_F} U_k \cdot \mathbf{0} \cdot V_k^\top = \mathbf{0}.$$

Frobenius inner product. For any matrices A, B, define $\langle A, B\rangle_F = \mathrm{tr}(A^\top B)$. The head update is $\Delta_{\mathrm{head}} = U_k V_k^\top$. We verify orthogonality:

$$\begin{aligned}\langle \Delta_{\mathrm{head}}, \Delta_{\mathrm{tail}}\rangle_F &= \mathrm{tr}\left((U_k V_k^\top)^\top \cdot \frac{sR_k}{\|R_k\|_F}\right) \\ &= \frac{s}{\|R_k\|_F}\mathrm{tr}(V_k U_k^\top R_k) \\ &= \frac{s}{\|R_k\|_F}\mathrm{tr}(U_k^\top R_k V_k) \quad \text{(cyclic property)} \\ &= \frac{s}{\|R_k\|_F}\mathrm{tr}(\mathbf{0}) = 0.\end{aligned}$$

Complementary projection. Define the tail projection $P_k^\perp = I - P_k$. Then:

$$P_k^\perp(M) = M - U_k U_k^\top M V_k V_k^\top.$$

The tail update satisfies $\Delta_{\mathrm{tail}} \in \mathrm{range}(P_k^\perp)$, ensuring $P_k(\Delta_{\mathrm{tail}}) = \mathbf{0}$ by construction. □

C.2 Proof of Theorem 2.2: Energy Control

We establish that the reference DAWN update has the same fixed Frobenius energy as Muon's polar update before global calibration.

C.2.1 Energy Decomposition via Matrix Norms

Proof. **Step 1: Pythagorean decomposition.** By the spectral firewall (Theorem 2.1), $\langle \Delta_{\mathrm{head}}, \Delta_{\mathrm{tail}}\rangle_F = 0$. The Pythagorean theorem for Frobenius norm gives:

$$\|\Delta_{\mathrm{DAWN}}\|_F^2 = \|\Delta_{\mathrm{head}}\|_F^2 + \|\Delta_{\mathrm{tail}}\|_F^2.$$

Step 2: Head energy. The head update $\Delta_{\mathrm{head}} = U_k V_k^\top$ has:

$$\|\Delta_{\mathrm{head}}\|_F^2 = \|U_k V_k^\top\|_F^2 = \mathrm{tr}(V_k U_k^\top U_k V_k^\top) = \mathrm{tr}(V_k V_k^\top) = k.$$

Step 3: Tail energy. For the theory-aligned raw-residual tail $\Delta_{\mathrm{tail}} = \sqrt{d-k}\, R_k/\|R_k\|_F$,

$$\|\Delta_{\mathrm{tail}}\|_F^2 = d - k.$$

Step 4: Combined energy.

$$\|\Delta_{\mathrm{DAWN}}\|_F^2 = k + (d-k) = d = \|\Delta_{\mathrm{Muon}}\|_F^2.$$

Thus the reference head–tail update preserves the same total Frobenius energy as Muon's polar update before global calibration. □

C.3 DAWN as a Design-Principle Approximation to MMSE Matrix-Sign

Design Principle C.1 (DAWN–MMSE connection, formerly Proposition F.1)**.** Under the spiked-noise model of **??** B.5 with r supercritical singular directions (Proposition B.9), and choosing rank $k = r$, DAWN's per-direction effective scaling satisfies:

1. **Head ($i \le k$):** DAWN applies $w_i = 1/\tilde{\sigma}_i$, matching the MMSE matrix-sign weight $\gamma_i \approx 1$ (which equals $1/\tilde{\sigma}_i \cdot \tilde{\sigma}_i$) up to single-sample fluctuation.
2. **Tail ($i > k$):** the theory-aligned raw residual preserves nonuniform spectral weights $a_i \propto \tilde{\sigma}_i$ instead of lifting all tail modes to unit magnitude, qualitatively following the reliability ordering suggested by MMSE shrinkage.

Remark C.2 (Why a Design Principle, not a Theorem). We deliberately state this as a Design Principle rather than a theorem because two approximation gaps prevent an exact equivalence:

1. *Single-sample $\tilde{G}$ vs. population G.* DAWN's polar is computed from the current (possibly momentum-averaged) gradient sample. Theorem B.7 uses the conditional reliability γ_i given Y; in practice momentum smoothing reduces variance in γ_i but does not eliminate the single-sample gap.

2. *Hard partition vs. soft shrinkage.* The MMSE estimator applies γ_i continuously in $[0, 1]$; DAWN instead uses different treatments on the two sides of a hard head–tail split. The reference tail retains the raw residual spectrum rather than implementing the oracle weights γ_i exactly.

We verify empirically (Sections 4, Appendix I.5) that the selected head subspace is reliably identified: subspace overlap $\langle \mathrm{colsp}(U_k^{\mathrm{DAWN}}), \mathrm{colsp}(U_r^*) \rangle > 0.95$ throughout training at k chosen at the BBP threshold.

D Algorithm Details

D.1 Full Pseudocode

We present two versions: a reference implementation (Algorithm 1) that computes SVD every step for maximum accuracy, and a practical version (Algorithm 2) that uses periodic SVD refresh for efficiency. Our ablation study (Section 4) shows both achieve similar performance.

Algorithm 1 DAWN: Spectral Firewall Optimizer (Reference Version)

Input: Initial parameters θ_0, learning rate η, rank k, momentum μ, weight decay λ, power iterations q, oversampling p
Initialize: Momentum buffer $b_0 = 0$
for $t = 1$ **to** T **do**
  Compute gradient $\tilde{G}_t \in \mathbb{R}^{m \times n}$; $d \leftarrow \min(m, n)$

  // Randomized Top-k SVD
  $\Omega \leftarrow \mathrm{randn}(n, k + p)$
  $Y \leftarrow \tilde{G}_t \Omega$
  for $j = 1$ **to** q **do**
    $Y \leftarrow \tilde{G}_t(\tilde{G}_t^\top Y)$ *// Power iteration*
  end for
  $Q, _ \leftarrow \mathrm{QR}(Y)$
  $B \leftarrow Q^\top \tilde{G}_t; \quad \tilde{U}, \Sigma, V \leftarrow \mathrm{SVD}(B)$
  $U_k \leftarrow Q\tilde{U}_{:,:k}; \quad V_k \leftarrow V_{:,:k}; \quad \sigma_k \leftarrow \mathrm{diag}(\Sigma)_{:k}$

  // Head: whitened update
  $\Delta_{\mathrm{head}} \leftarrow U_k V_k^\top$

  // Tail: energy-controlled update
  $R_k \leftarrow \tilde{G}_t - U_k \,\mathrm{diag}(\sigma_k)\, V_k^\top$
  $\Delta_{\mathrm{tail}} \leftarrow \sqrt{d - k} \cdot R_k / \|R_k\|_F$

  // Nesterov momentum + weight decay
  $b_t \leftarrow \mu\, b_{t-1} + \Delta_{\mathrm{head}} + \Delta_{\mathrm{tail}}$
  $\theta_t \leftarrow \theta_{t-1} - \eta\,(b_t + \mu\,(b_t - b_{t-1})) - \eta\lambda\,\theta_{t-1}$
end for

Computational cost per step:

Algorithm 2 DAWN: Practical Implementation

Input: Learning rate η, momentum μ, head rank k, refresh period p, RMS decay β_2
Initialize: Momentum buffer $m \leftarrow 0$, RMS buffer $v \leftarrow 0$, subspace cache $Q \leftarrow \text{None}$
for $t = 1$ **to** T **do**
  $m \leftarrow \mu \cdot m + \nabla\mathcal{L}(\theta)$ *// Momentum accumulation*
  for each 2D parameter $W \in \theta$ **do**
    $G \leftarrow \text{reshape}(m_W)$ to (M, N) with $M \geq N$

    // Step 1: Subspace Identification
    if $t \mod p = 0$ **or** $Q = \text{None}$ **then**
      $Q \leftarrow$ top-k left singular vectors of G *// SVD refresh*
    else
      $Y \leftarrow GG^\top Q_{\text{prev}}; \quad Q \leftarrow \text{QR}(Y)$ *// Subspace iteration*
    end if

    // Step 2: Head-Tail Decomposition
    $B \leftarrow Q^\top G$ *// Head projection, $B \in \mathbb{R}^{k\times N}$*
    $R \leftarrow G - QB$ *// Tail residual, $R \in \mathbb{R}^{M\times N}$*

    // Step 3: Head Processing (Newton-Schulz)
    $\hat{B} \leftarrow \text{NS}(B)$ *// 5 iterations of NS on $k \times N$ matrix*
    $\Delta_{\text{head}} \leftarrow \sqrt{k} \cdot Q\hat{B}/\|Q\hat{B}\|_F$

    // Step 4: Tail Processing (Row-wise RMS)
    $v_i \leftarrow \beta_2 v_i + (1-\beta_2)\|R_i\|_2^2/N$ *// Per-row variance EMA*
    $\tilde{R}_i \leftarrow R_i/\sqrt{v_i/(1-\beta_2^t) + \epsilon}$ *// RMS normalize*
    $\tilde{R} \leftarrow \tilde{R} - Q(Q^\top \tilde{R})$ *// Orthogonal firewall*
    $\Delta_{\text{tail}} \leftarrow \sqrt{d-k} \cdot \tilde{R}/\|\tilde{R}\|_F$

    // Step 5: Combine and Update
    $\Delta \leftarrow \Delta_{\text{head}} + \Delta_{\text{tail}}$
    $\Delta \leftarrow \text{target_rms} \cdot \sqrt{MN} \cdot \Delta/\|\Delta\|_F$ *// Global RMS calibration*
    $W \leftarrow W - \eta \cdot \Delta$
  end for
end for

Step	Operation	Complexity
1. Subspace ID	`svd_lowrank` or subspace iteration	$O(MNk)$ / $O(Mk^2)$
2. Projection	$B = Q^\top G$, $R = G - QB$	$O(MNk)$
3. Head (NS)	Newton-Schulz on B	$O(kN^2)$
4. Tail (RMS)	Row-wise normalization	$O(MN)$
5. Combine	$\Delta = \Delta_h + \Delta_t$	$O(MN)$

With periodic SVD (period p), the amortized cost per step is $O(MNk/p + Mk^2 + kN^2)$. For $p = 40$, $k = 64$, this is $\sim 40\times$ cheaper than every-step SVD.

D.2 Rank Schedule

We use a cosine decay schedule for the rank:

$$k(t) = k_{\min} + \frac{k_{\max} - k_{\min}}{2}\left(1 + \cos\left(\frac{\pi t}{T}\right)\right).$$

For our GPT-2XL runs, $k_{\max} = 128$ and $T \approx$ 10K steps; $k_{\min} \in \{128, 8, 0\}$ defines the three DAWN variants. Specifically, $k_{\max} = 128$ is set as a conservative upper bound on the early-training BBP-identifiable spike count (which we measure to be $\lesssim 80$ in practice; see Fig. **??**), and $k_{\min} = 8$

is calibrated to the late-training stable BBP count. The decay window $T \approx 10$K steps is the empirical timescale over which the BBP count drops from its early-training value to its stable late-training value.

Rationale. Early training has high SNR (large gradients relative to noise), justifying larger k to whiten more identifiable directions. As training converges, gradient signal concentrates in fewer directions and SNR per remaining direction decreases on the periphery, so the optimal strategy shifts toward smaller k. The cosine form is a smooth approximation to the empirical BBP-spike-count trajectory; its smoothness avoids step-to-step rank jitter that a hard BBP threshold would induce. See Section 3.3 for the total-MSE rationale and Fig. **??** for the empirical alignment between the cosine schedule and the BBP count.

D.3 Computational Cost Breakdown

For $m \times n$ matrix with $m \geq n$, rank k, and q power iterations:

Operation	**Cost**
Random projection $\tilde{G}\Omega$	$O(mnk)$
Power iterations (each)	$O(mnk)$
QR decomposition	$O(mk^2)$
Small SVD of B	$O(nk^2)$
Residual computation	$O(mnk)$
Total	$O(mnkq)$

For square matrices ($m = n = d$): $O(d^2kq)$, compared to $O(d^3)$ for full SVD or Newton-Schulz.

D.4 Complexity Comparison Table

Method	**Exact FLOPs**	**Per-param**	**Info used**
Muon (Newton-Schulz)	$O(d^3)$	$O(d)$	Full spectrum
Shampoo/K-FAC	$O(d^3)$	$O(d)$	Full spectrum
DAWN	$O(d^2k)$	$O(k)$	Top-k only
AdamW	$O(d^2)$	$O(1)$	None (diagonal)

With typical settings $k = 64$, $d = 4096$: DAWN is $\frac{d}{kq} = 32\times$ cheaper than full Newton-Schulz per iteration.

D.5 Theoretical Guarantees for Randomized Algorithms

We provide proof sketches for the key algorithmic components: randomized SVD and subspace tracking.

D.5.1 Randomized SVD Approximation

Lemma D.1 (Randomized SVD error bound)**.** *Let $G \in \mathbb{R}^{m\times n}$ with singular values $\sigma_1 \geq \sigma_2 \geq \cdots$. For target rank k, oversampling $p \geq 2$, and q power iterations, the randomized SVD approximation $\hat{G}_k = \hat{U}_k\hat{\Sigma}_k\hat{V}_k^\top$ satisfies:*

$$\mathbb{E}\|G - \hat{G}_k\|_F \leq \left(1 + \sqrt{\frac{k}{p-1}}\right) \cdot \left(\sum_{j>k} \sigma_j^2\right)^{1/2} + \frac{e\sqrt{k+p}}{p} \cdot \sigma_{k+1}^{2q+1}/\sigma_k^{2q}.$$

Proof sketch. The algorithm projects G onto a random subspace via $Y = G\Omega$ where $\Omega \in \mathbb{R}^{n\times(k+p)}$ is Gaussian. The key insight is the Johnson-Lindenstrauss lemma: random projections approximately preserve distances. Power iteration $Y \leftarrow G(G^\top Y)$ amplifies the top singular directions by factor

$(\sigma_i/\sigma_{k+1})^{2q}$, suppressing the tail. The first term bounds the irreducible tail energy $\|G - G_k\|_F$; the second term decays exponentially with q when there is a spectral gap. For $q = 2$ and mild gap ($\sigma_k/\sigma_{k+1} > 1.1$), the error is dominated by the optimal rank-k approximation error. Full analysis in Halko et al. (2011, Theorem 10.5).

D.5.2 Subspace Tracking via Power Iteration

Lemma D.2 (Subspace tracking convergence)**.** *Let Q_t be the current subspace estimate and G_{t+1} the new gradient. One step of subspace iteration $Q_{t+1} = \mathrm{QR}(G_{t+1}G_{t+1}^\top Q_t)$ satisfies:*

$$\sin\angle(Q_{t+1}, U_k^{(t+1)}) \leq \frac{\sigma_{k+1}^{(t+1)}}{\sigma_k^{(t+1)}} \cdot \sin\angle(Q_t, U_k^{(t+1)}) + O(\|G_{t+1} - G_t\|_{\mathrm{op}}/\sigma_k^{(t)})$$

where $U_k^{(t+1)}$ is the true top-k subspace of G_{t+1}.

Proof sketch. Subspace iteration is equivalent to one step of power method on the covariance $GG^\top$. The contraction factor $\sigma_{k+1}/\sigma_k < 1$ ensures convergence to the dominant subspace. When the gradient changes slowly (small $\|G_{t+1} - G_t\|_{\mathrm{op}}$), the perturbation term is negligible, and subspace tracking maintains accuracy without full SVD recomputation. This justifies periodic SVD refresh: between refreshes, cheap subspace iteration tracks the slowly-drifting head subspace. The perturbation bound follows from Davis-Kahan applied to the covariance matrices.

D.5.3 Combined Error Analysis

Combining these results: with periodic SVD every p steps and subspace iteration between, DAWN's head subspace estimate $\hat{Q}$ satisfies:

$$\sin\angle(\hat{Q}, U_k) \leq C \cdot \left(\frac{\sigma_{k+1}}{\sigma_k}\right)^p + O\left(\frac{p \cdot \delta_G}{\sigma_k}\right)$$

where $\delta_G = \max_t \|G_{t+1} - G_t\|_{\mathrm{op}}$ is the maximum gradient change per step. For typical training ($\delta_G \ll \sigma_k$, $\sigma_{k+1}/\sigma_k < 0.9$), even $p = 40$ maintains accurate subspace tracking.

E GPT-2 Small Model Experiments (124M)

We conduct comprehensive experiments on GPT-2 (124M parameters) including ablation studies to identify which mechanisms are essential for matching Muon's performance when using split head-tail processing. For detailed model architecture and training configuration, see Section I.

E.1 Experimental Setup

Model: GPT-2 style (124.36M parameters). **Dataset:** FineWeb-Edu (sample-10BT, 9.94B tokens). **Batch size:** 256 (effective, 4 GPUs $\times$ micro_batch=16 $\times$ grad_accum=4). **Tokens per step:** 262K (256×1024). **Training:** $\sim$38K steps ($\sim$10B tokens). **Precision:** bfloat16 with AMP. **Peak LR:** 3×10^{-4}.

E.2 Full Mechanism Comparison

Table 5 shows all configurations tested. "NS" denotes Newton-Schulz orthogonalization on the head ($B = Q^\top G$). "RMS" denotes row-wise RMS normalization on the tail. Muon applies NS to the full gradient (no head-tail split).

E.3 Component Necessity Analysis

Table 6 isolates the contribution of each component by systematic removal. Head NS refers to Newton-Schulz on $B = Q^\top G$; Tail RMS refers to row-wise RMS normalization on the residual.

Table 5: Complete comparison of head and tail processing mechanisms, single-seed runs at 8B tokens of GPT-2 (124M). Headline rows (Muon, AdamW, DAWN-64) match Table **??** 3-seed means within ≤ 0.005 std. Green : matches Muon (< 3.05). Red : diverged or failed.

Head	Tail	Val Loss	Δ vs Muon
Baselines			
Muon (NS on full G)		3.024	—
AdamW		3.069	+0.045
DAWN Variants (Head: NS on B, Tail: RMS)			
NS	RMS (p=40)	3.020	-0.004
NS	RMS (adaptive)	3.035	+0.011
NS	RMS (bounded)	3.036	+0.012
NS	RMS (DAWN)	3.042	+0.018
NS	RMS (p=20)	3.042	+0.018
NS	RMS (EMA)	3.065	+0.041
NS	RMS (anneal)	3.065	+0.041
NS	FastDAWN	3.235	+0.210
NS	None (k=128)	3.279	+0.255
NS	None (k=64)	3.466	+0.442
NS	Clip	4.579	+1.555
Head Processing Variants			
RMS	RMS	3.048	+0.024
ColRMS	ColRMS	3.071	+0.046
None (pwr0)	RMS	4.213	+1.189
Rank (k) Variants			
NS (k=128)	RMS	3.057	+0.033
NS (k=32)	RMS	3.072	+0.048
Failed Configurations			
NS	RMS (long warmup)	3.762	+0.738
NS	RMS (reg=10^3)	7.247	+4.223

Table 6: Necessity of each component, single-seed runs at 8B tokens. Removing NS orthogonalization or tail processing significantly degrades performance. Headline rows match Table **??** 3-seed means within ≤ 0.005 std.

Configuration	Head NS	Tail RMS	Val Loss
Muon (full)	✓	✓	3.024
DAWN-128 (head+tail)	✓	✓	3.057
DAWN-64 (head+tail)	✓	✓	3.042
DAWN-32 (head+tail)	✓	✓	3.072
Head-only (k=128)	✓	×	3.279
Head-only (k=64)	✓	×	3.466
No NS (pwr0)	×	✓	4.213
Degradation from removing tail			+0.22 to +0.42
Degradation from removing head NS			+1.19

E.4 DAWN Variants Comparison

Table 7 compares different DAWN configurations. All variants use NS for head ($B = Q^\top G$) and row-wise RMS for tail. Variants differ in SVD refresh strategy and energy allocation.

Table 7: DAWN variants comparison, single-seed runs. All use NS for head and row-wise RMS for tail; variants differ in SVD refresh strategy and energy allocation.

Variant	Val Loss	Mechanism
Muon (baseline)	3.024	NS on full gradient
Matching Muon (<3.05):		
DAWN-64 (p=40)	3.020	SVD refresh every 40 steps
DAWN-64 (adaptive)	3.035	SVD refresh when subspace drifts
DAWN-64 (bounded)	3.036	Bounded tail energy allocation
DAWN-64	3.042	Standard DAWN with EMA variance reduction
DAWN-64 (p=20)	3.042	SVD refresh every 20 steps
Slightly worse (<3.08):		
DAWN-128	3.057	Higher rank, more compute
DAWN-64 (anneal)	3.065	Rank annealing (too aggressive)
DAWN-64 (EMA)	3.065	EMA for Q updates (introduces lag)
DAWN-64 (ColRMS)	3.071	Column-wise RMS instead of row-wise
DAWN-32	3.072	Lower rank
Failed:		
FastDAWN	3.235	Oversimplified, unstable
Clip	4.579	Clipping destroys update structure

E.5 Reproducibility: Error Bars and Additional Baselines

The 124M ablation tables above report single-seed validation losses for clarity of comparison across many configurations. To validate reproducibility and establish ordering with statistical confidence, we re-ran the headline configurations across three random seeds and added a broader set of curvature-aware baselines.

E.6 Ablation Study: Configuration and Design Axes

We conduct a comprehensive ablation study to isolate the contributions of three core design axes in DAWN: (i) subspace estimation strategy, (ii) head rank k, and (iii) tail processing scheme. All experiments are performed on a 124M GPT-2 model trained on 10B tokens from FineWeb-Edu.

Baselines. Muon. We consider Muon as the full-spectrum spectral optimization baseline. Muon applies Newton–Schulz (NS) whitening to the entire gradient matrix G at every step:

$$\Delta = \mathrm{NS}(G).$$

AdamW. Standard AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay $\lambda = 0.1$.

DAWN decomposition. All DAWN variants decompose the gradient into a signal-dominated head subspace and a noise-dominated tail:

$$G = QQ^\top G + (I - QQ^\top)G,$$

where $Q \in \mathbb{R}^{m \times k}$ spans the top-k left singular subspace.

E.6.1 DAWN Variants

DAWN-64 (default). Rank $k = 64$ with periodic SVD refresh every $p = 20$ steps. The head update applies NS whitening:

$$\Delta_{\text{head}} = Q \cdot \mathrm{NS}(Q^\top G),$$

while the tail uses row-wise RMS normalization.

DAWN-64 (p=40). Refreshes Q every 40 steps, amortizing SVD cost by $2\times$.

DAWN-64 (adaptive). Refreshes Q only when subspace drift exceeds a threshold:

$$\frac{\|Q_{\text{old}}^\top Q_{\text{new}}\|_F^2}{k} < 0.97.$$

DAWN-64 (bounded). Uses bounded tail scaling:

$$\Delta_{\text{tail}} = \min\left(1, \frac{\sqrt{d-k}}{\|\tilde{R}\|_F}\right)\tilde{R},$$

where $\tilde{R}$ denotes RMS-normalized residual gradients.

DAWN-64 (RMS-head). Replaces NS whitening in the head with RMS normalization:

$$\Delta_{\text{head}} = Q \cdot \text{RMS}(Q^\top G).$$

DAWN-64 (EMA). Tracks the subspace via exponential moving average:

$$Q^{(t)} = \text{QR}\Big(\rho Q^{(t-1)} + (1-\rho)Q_{\text{new}}\Big), \quad \rho = 0.9.$$

DAWN-64 (ColRMS). Uses column-wise RMS normalization for both head and tail.

DAWN-32 / DAWN-128. Varies rank $k \in \{32, 128\}$ with standard DAWN processing.

DAWN-anneal. Rank annealing schedule: $k = 128 \to 64 \to 32 \to 16$ every 10K steps.

E.6.2 Ablation Controls (Failure Modes)

Head-only-k. Discards tail updates ($\Delta_{\text{tail}} = 0$), testing whether tail processing is necessary.

No-ortho. Removes orthogonalization and whitening; uses raw gradients with global norm scaling.

Clip. Replaces RMS normalization with element-wise gradient clipping ($\tau = 1.0$).

Long-warmup. Extends warmup to $10\times$ the default schedule.

Reg=1e3. Applies strong tail regularization with $\lambda = 10^3$.

FastDAWN. Simplified tail update without normalization:

$$\Delta_{\text{tail}} = R - 0.1R^{(t-1)}.$$

E.6.3 Ablation Results

E.6.4 RowRMS versus the theory-aligned raw residual

The mode-wise MSE analysis in Section 3.3 is written in the singular basis of the raw residual R, whereas row-wise RMS preconditioning can rotate that basis. To isolate the theory-aligned mechanism, we therefore repeated the GPT-2 124M experiment with the same head rank and refresh setting but removed RowRMS, using the energy-matched raw residual

$$\Delta_{\text{tail}}^{\text{ref}} = \sqrt{d-k}\, R/\|R\|_F$$

for the tail. Across three seeds, this no-RowRMS variant gives performance comparable to the RowRMS implementation. This ablation indicates that the main empirical effect does not rely on row-wise normalization: the theory explains the head–tail split and preservation of a nonzero residual signal, while RowRMS is a practical preconditioning refinement. We retain RowRMS in the practical implementation because coordinate-wise adaptive normalization is a well-established optimization trick and is also used in practical Muon variants.

E.6.5 Comparison to AdamW

AdamW applies element-wise adaptive scaling $m_t/(\sqrt{v_t} + \epsilon)$, whereas DAWN performs *structured spectral processing*: the head subspace captures curvature via whitening, while the tail is stabilized by RMS normalization to prevent noise amplification.

Table 8: Ablation configurations and validation loss, single-seed runs at 8B tokens. NS = Newton–Schulz whitening; RMS = row-wise RMS normalization. Headline rows are 3-seed means within ≤ 0.005 std.

Method	k	Q Refresh	Head	Tail	Val Loss
Muon	—	every step	NS	NS	3.024
AdamW	—	—	—	—	3.069
DAWN-64 (p=40)	64	40 steps	NS	RMS	3.020
DAWN-64 (p=20)	64	20 steps	NS	RMS	3.042
DAWN-64 (adaptive)	64	drift-based	NS	RMS	3.035
DAWN-64 (bounded)	64	20 steps	NS	bounded	3.036
DAWN-64 (RMS-head)	64	20 steps	RMS	RMS	3.048
DAWN-64 (EMA)	64	EMA	NS	RMS	3.065
DAWN-64 (ColRMS)	64	20 steps	ColRMS	ColRMS	3.071
DAWN-32	32	20 steps	NS	RMS	3.072
DAWN-128	128	20 steps	NS	RMS	3.057
DAWN-anneal	128→16	adaptive	NS	RMS	3.065
Head-only-64	64	20 steps	NS	×	3.466
Head-only-128	128	20 steps	NS	×	3.279
No-ortho	—	—	×	×	4.213

E.6.6 Reproducibility

All experiments in this section use fixed random seed 42. Training is conducted on a 4x NVIDIA H200 141GB GPU with PyTorch 2.1 and CUDA 12.1. Validation loss is computed every 125 steps on a held-out 10M-token split. Each 10B-token run requires approximately 8 hours. Code and logs will be released upon publication.

F GPT-2XL Large Model Experiments (1.5B)

We provide comprehensive details for the GPT-2XL (1.5B) pretraining experiments on 105B tokens from FineWeb (Fig. 3).

F.1 Model Architecture

Hyperparameter	**Value**
Parameters	1.5B
Layers	52
Hidden dimension	1536
Attention heads	12
Head dimension	128
Vocabulary size	50,257
Context length	1024
Activation	GELU
Position embedding	Learned

F.2 TRAINING CONFIGURATION

Hyperparameter	Value
Dataset	FineWeb (HuggingFace, 100B subset)
Total tokens	$\sim$105B
Total steps	200,000
Batch size	512 (8 H200 141G GPUs $\times$ device_batch=16 $\times$ grad_accum=4)
Sequence length	1024
Tokens per step	524,288 (512×1024)
Precision	bfloat16 with AMP
Gradient clipping	1.0 (global norm)
Learning Rate Schedule	
Peak LR (all methods)	1.8×10^{-4}
Peak LR (lm_head)	1.8×10^{-3} ($10\times$ base)
Warmup steps	0
Warmdown start	80B tokens (step 150,000)
Warmdown steps	50,000
Schedule	Constant, cosine decay to $0.1\times$base
Optimizer Settings	
β_1, β_2 (AdamW)	0.9, 0.95
β_1, β_2 (Muon/DAWN)	0.95, 0.95
Weight decay	0.1
ϵ	10^{-8}

F.3 DAWN VARIANT SPECIFICATIONS

We test three DAWN configurations to study the effect of rank decay and tail processing. All variants employ rank decay over 10,000 steps ($\sim$5B tokens):

DAWN-128. Head rank: $k = 128$ (fixed, no decay). Head processing: randomized SVD . Tail processing: EMA variance reduction with continuation (continue=True). Rationale: full head whitening with smooth tail updates.

DAWN-8 (Rank-Decayed). Head rank: $k = \max(8, 128 \times (1 - \text{step}/10000))$, decaying $128 \to 8$. Head processing: randomized SVD . Tail processing: RMS normalization with continuation (continue=True). Rationale: tests aggressive rank reduction while maintaining tail updates.

DAWN-0 (Rank-Collapsed, Tail-Only). Head rank: $k = 128 \times (1 - \text{step}/10000)$, decaying $128 \to 0$. Head processing: randomized SVD until the rank collapses; once $k = 0$, no head update. Tail processing: RMS normalization on the full gradient (the "tail" becomes the entire gradient once $k = 0$). Rationale: sanity-check ablation. Once k reaches 0, the optimizer reverts to a first-order RMS-style update on the full gradient—essentially AdamW with energy normalization—providing a worst-case lower bound that isolates the contribution of head polar whitening.

F.4 REPRODUCIBILITY

All experiments in this section use fixed random seed 42. Training is conducted on a 8x NVIDIA H200 141GB GPU with PyTorch 2.1 and CUDA 12.1. Validation loss is computed every 125 steps on a held-out 10M-token split. Each 100B-token run requires approximately 100 hours. Code and logs will be released upon publication.

F.5 DATA REPETITION EXPERIMENT

To test robustness under limited data regimes, we train GPT-2XL on a small dataset (10B tokens) repeated multiple times to reach 100B total tokens (Fig. 6).

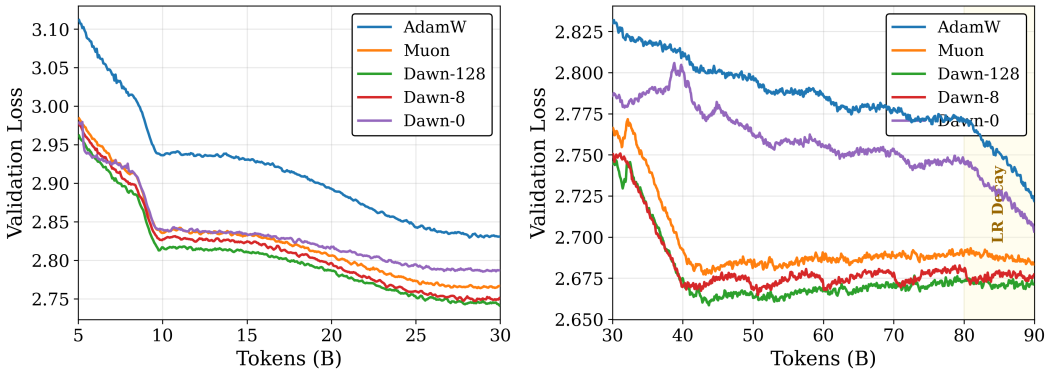

Figure 6: **Data repetition: GPT-2XL on 10B tokens repeated to 100B. Left:** Early training shows similar late-epoch loss-transition timing (∼30B) across matrix-whitening methods. **Right:** Dawn-0 initially oscillates but converges to better final loss than AdamW. Dawn-128, Dawn-8, and Muon track closely throughout. Results mirror the unique-data experiment, suggesting head-tail dynamics are robust to data repetition.

Configuration.

Hyperparameter	Value
Model	GPT-2XL (1.5B parameters)
Dataset	FineWeb 10B tokens (repeated 10×)
Total tokens	∼100B (10 epochs over 10B)
Batch size	512 (8 GPUs × 16 × 4)
Sequence length	1024
Tokens per step	524,288
Total steps	∼190K
Learning rate	1.8×10^{-4} (all methods)
lm_head LR	1.8×10^{-3} (10×)
Warmup	0 steps
Schedule	Constant → cosine decay (last 25%)

Key observations. (1) **Late-epoch transition timing:** All matrix-whitening methods show a sharp loss-decrease transition at ∼30B tokens (3-epoch boundary), suggesting this phenomenon is tied to data coverage rather than total compute. (2) **Dawn-0 instability:** Dawn-0 (rank decays to 0; tail RMS continues on the full gradient) reverts to a first-order RMS regime once k collapses and exhibits the largest oscillations of all configurations. (3) **Convergence:** Despite initial instability, Dawn-0 converges to a better final loss than AdamW, consistent with the unique-data experiment—this validates that aggressive tail suppression is beneficial at convergence even under data repetition. (4) **Robustness:** Dawn-128, Dawn-8, and Muon track closely throughout, matching their behavior in the 100B unique-data experiment; the head-tail dynamics are robust to data repetition.

Interpretation. Data repetition amplifies overfitting risk, making the noise-vs-signal tradeoff more critical. The fact that DAWN variants maintain their relative performance suggests that the head-tail decomposition captures fundamental optimization dynamics rather than dataset-specific phenomena.

G Wide Model Experiment (4B Parameters)

This appendix provides the full setup and supporting numbers for the wide-model experiment summarized in Section 4.4 (Fig. 5).

Model architecture.

Hyperparameter	Value
Architecture	Wide GPT
Vocabulary size	50,257
Layers (n_layer)	5
Attention heads (n_head)	64
Hidden dimension (n_embd)	8,192
Parameters	$\sim$4B

This is an intentionally "wide and shallow" design: only 5 layers but with hidden dimension 8192, creating very high-dimensional weight matrices where Muon's $O(d^3)$ cost becomes prohibitive.

Training configuration.

Hyperparameter	Value
Dataset	FineWeb
Sequence length	1,024
Batch size	512 sequences
Tokens per step	524,288
Total iterations	24,800
Total tokens	$\sim$13B
Warmup	0 steps
Weight decay	0

Optimizer learning rates.

Optimizer	Main LR	lm_head LR
AdamW	3.6×10^{-5}	3.6×10^{-5}
Muon	3.6×10^{-4}	3.6×10^{-4}
Dawn-128	3.6×10^{-4}	3.6×10^{-4}
Dawn-8	3.6×10^{-4}	3.6×10^{-4}

Dawn-8 uses rank decay from $k = 128$ to $k = 8$ over the first 10K steps.

Computational cost analysis. For the 8192×8192 weight matrices in this model:

Method	Complexity	Relative Cost	Speedup
Muon	$O(d^3) = O(8192^3)$	1.0	—
Dawn-128	$O(d^2k) = O(8192^2 \cdot 128)$	$k/d = 1.6\%$	$64\times$
Dawn-8	$O(d^2k) = O(8192^2 \cdot 8)$	$k/d = 0.1\%$	$1024\times$

Additional observations. Beyond the headline findings reported in Section 4.4, two minor points worth noting: (1) Dawn-8 exhibits a small loss spike around 8B tokens, possibly related to a phase transition in rank decay; it quickly recovers to the Muon/Dawn-128 trajectory. (2) AdamW (blue in Fig. 5b) shows 0.1–0.2 higher loss throughout training, confirming that matrix-whitening optimizers provide substantial benefits on wide architectures where curvature structure is more complex.

H Synthetic Experiment Details

We provide complete details for the synthetic experiments in Figure 1.

H.1 Signal Matrix Generation

For each experiment, we generate a ground-truth signal matrix $G_{\text{true}} \in \mathbb{R}^{m \times n}$ with controlled spectral structure:

$$G_{\text{true}} = U \, \text{diag}(s_1, \ldots, s_r) \, V^\top$$

where $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$ are random orthonormal matrices (via QR decomposition of Gaussian matrices), and $s_1, \ldots, s_r$ follow one of six spectral patterns:

Spectrum Type	**Singular Values**
Flat ($\kappa = 1$)	$s_i = 1$
Linear ($\kappa = 10$)	$s_i = 1 - 0.9(i-1)/(r-1)$
Power-law slow ($\beta = 0.5$)	$s_i = i^{-0.5}$
Power-law fast ($\beta = 1.0$)	$s_i = i^{-1}$
Exponential	$s_i = \exp(-\alpha i)$, $\alpha = -\ln(0.01)/r$
Step	$s_i = 1$ for $i \le r/2$, else $s_i = 0.1$

All singular values are normalized so $s_1 = 1$.

H.2 Noise Model

We add i.i.d. Gaussian noise:

$$\tilde{G} = G_{\text{true}} + E, \quad E_{ij} \sim \mathcal{N}(0, \sigma^2)$$

where σ is chosen to achieve a target SNR:

$$\text{SNR} = \frac{\|G_{\text{true}}\|_F}{\sigma\sqrt{mn}}.$$

We test SNR $\in \{0.5, 1.0, 2.0, 5.0, 10.0, 20.0\}$.

H.3 Subspace Alignment Metric

Given an update Δ and ground-truth signal subspace U_{true}, we measure:

$$\text{Alignment} = \frac{\|U_\Delta^\top U_{\text{true}}\|_F}{\sqrt{r}}$$

where U_Δ is the top-r left singular vectors of Δ. This metric equals 1 when Δ perfectly recovers the signal subspace and approaches 0 when Δ is orthogonal to it.

H.4 Methods Compared

(1) **Muon (Newton-Schulz):** $\Delta = \tilde{G}(\tilde{G}^\top\tilde{G})^{-1/2}$ via 5 Newton-Schulz iterations with coefficients $(a, b, c) = (3.4445, -4.7750, 2.0315)$. (2) **DAWN:** Top-$k$ SVD whitening with $k = r$ (matching signal rank) and $q = 2$ power iterations. (3) **First-order baseline:** $\Delta = \tilde{G}/\|\tilde{G}\|_F$ (normalized gradient).

H.5 Computational Complexity Measurement

We measure wall-clock time on NVIDIA H200 GPU with: (1) 10 warmup iterations (discarded); (2) 30 timed iterations (averaged); (3) `torch.cuda.synchronize()` before and after timing.

Square matrix sizes: $d \in \{512, 1024, 2048, 4096, 8192, 12288, 16384\}$.

I GPT-2XL pectral Analysis

We provide comprehensive details for the GPT-2XL/Wide-GPT pretraining experiments and spectral analysis methodology (complements Section E).

I.1 Training Configuration

Hyperparameter	Value
Dataset	FineWeb-Edu (HuggingFace)
Total tokens	105B
Batch size (effective)	512 sequences
Tokens per step	0.5M
Total steps	$\sim$200,000
Precision	bfloat16 with AMP
Learning Rate Schedule	
Peak LR	3×10^{-4}
Warmup	100M tokens ($\sim$382 steps)
Schedule	Cosine decay to 0
Optimizer (DAWN)	
β_1, β_2	0.9, 0.95
Weight decay	0.1
Head rank k	64 (cosine decay to 16)
Power iterations q	2
ϵ	10^{-8}

I.2 Layer-Specific Treatment

Following standard practice, we apply different optimization strategies to different parameter types: (1) **Weight matrices** (attention QKV, projection, MLP): DAWN with head-tail split; (2) **Embedding matrices**: DAWN with higher rank ($k = 128$) due to vocabulary structure; (3) **LayerNorm parameters**: Standard AdamW (1D, no SVD structure); (4) **Bias terms**: Standard AdamW (1D).

I.3 Training Dynamics Observations

We observe distinct phases during training:

Phase 1: Rapid descent (steps 0–5K). Loss drops from $\sim$11 to $\sim$4. Both Muon and DAWN show similar behavior, with high-SNR gradients where full whitening is beneficial. DAWN's rank schedule starts high ($k = 64$).

Phase 2: Steady optimization (steps 5K–50K). Loss decreases from $\sim$4 to $\sim$3. DAWN and Muon track closely, with DAWN showing slightly lower variance in loss curves.

Phase 3: Fine-tuning (steps 50K–76K). Loss decreases from $\sim$3 to final values. This is where the head-tail dichotomy becomes most important: (1) Muon continues full-spectrum whitening, amplifying residual noise; (2) DAWN reduces rank and bounds tail updates.

I.4 Spectral Analysis During Training

We tracked the singular value spectrum of gradients throughout training: (1) **Early training:** Gradient spectra are relatively flat (condition number $\kappa \approx$ 10–100), justifying full whitening. (2) **Late training:** Spectra become increasingly spiked ($\kappa > 1000$), with most energy in top-k directions; the "tail" singular values approach the noise floor. (3) **SNR decay:** Effective SNR decreases as $|g|/\sigma$ shrinks near convergence, pushing more directions into the noise-dominated regime.

This empirical observation directly supports our theoretical prediction: the head-tail boundary shifts during training, and adaptive rank selection is beneficial.

I.5 Spectral Analysis Methodology

Why we use a cosine schedule instead of a runtime BBP rule. A natural alternative to DAWN's cosine schedule is to set $k(t) = \#\{i : \hat{\sigma}_i > \lambda_+\}$ directly per step. Three issues make this less

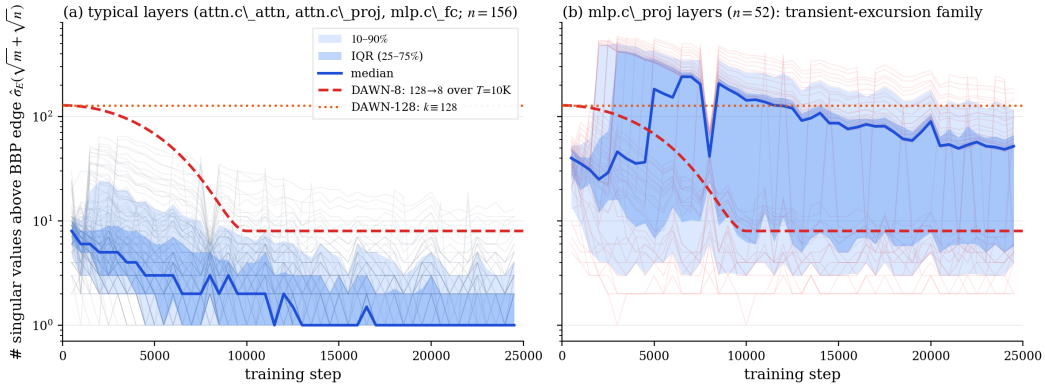

Figure 7: **Per-layer-family decomposition of the BBP-identifiable head rank (AdamW gradients, GPT-2XL).** Same data as Fig.2a, split by weight type. **(a)** The 156 typical layers (`attn.c_attn`, `attn.c_proj`, `mlp.c_fc`): median collapses to 1 and IQR to ≤ 3 by step 10K; DAWN-8's cosine schedule (red dashed) is a tight upper envelope. **(b)** The 52 `mlp.c_proj` layers form a distinct "transient-excursion" family: median peaks near 50 and stays elevated, exposing a per-weight-type rank heterogeneity that motivates conservatively over-provisioned schedules such as DAWN-128 (orange dotted). The total-MSE bound (Equation (7) of the main text) shows the consequent over-provisioning is benign.

attractive: (i) it requires an online estimate of σ_E (MP fit per layer per step), (ii) BBP's isotropic-Gaussian-noise assumption is violated by real anisotropic, heavy-tailed gradient noise, and (iii) the hard threshold creates step-to-step jitter when $\hat{\sigma}_i$ values cross λ_+, forcing costly head-subspace re-allocation. The cosine schedule is empirically a smooth approximation to this ideal rule and avoids all three issues.

Effective vs. BBP-identifiable rank. The empirical effective rank, the BBP-identifiable spike count, and DAWN's cosine schedule share a common shape across training: **theory (BBP)** $\rightarrow$ **observable proxy** (r_{eff}) $\rightarrow$ **algorithm schedule (cosine)** are mutually consistent.

Methodology. We provide detailed methodology for the spectral analysis metrics used throughout this paper.

Data Collection. During training, we log the singular value decomposition (SVD) of gradients for all weight matrices at regular intervals (every 5 steps). For each weight matrix $W \in \mathbb{R}^{m\times n}$, we compute the SVD of its gradient $G = \nabla_W \mathcal{L}$:

$$G = U\Sigma V^\top \tag{19}$$

where $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{\min(m,n)})$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)} \geq 0$.

Effective Rank. We measure the effective dimensionality of gradients using the entropy-based effective rank:

$$r_{\text{eff}} = \exp\left(-\sum_{i=1}^{n} p_i \log p_i\right) \tag{20}$$

where $p_i = \sigma_i^2 / \sum_j \sigma_j^2$ represents the fraction of total energy in the i-th singular direction. This metric equals 1 for rank-1 matrices and n for matrices with uniform singular values, providing a continuous measure of dimensionality.

Top-k Energy Ratio. We quantify energy concentration by computing the fraction of total Frobenius norm captured by the top k singular values:

$$\rho_k = \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{j=1}^{n} \sigma_j^2} = \frac{\|G_k\|_F^2}{\|G\|_F^2} \tag{21}$$

where $G_k = U_k \Sigma_k V_k^\top$ is the best rank-k approximation. This directly measures how much gradient information is captured by DAWN's head subspace.

Normalized SVD Entropy. To compare across layers with different dimensions, we use the normalized entropy:

$$H_{\text{norm}} = \frac{-\sum_i p_i \log p_i}{\log n} \in [0, 1] \tag{22}$$

where $H_{\text{norm}} = 0$ indicates rank-1 (maximally concentrated) and $H_{\text{norm}} = 1$ indicates uniform (full-rank) distribution.

Signal Singular Values. To distinguish meaningful gradient directions from noise, we apply the Gavish-Donoho optimal hard threshold (Gavish & Donoho, 2014). For a matrix of aspect ratio $\beta = m/n$ (assuming $m \leq n$), the threshold is:

$$\tau = \omega(\beta) \cdot \text{median}(\sigma_i) \tag{23}$$

where $\omega(\beta)$ is the optimal coefficient derived from random matrix theory. Singular values above τ are considered "signal"; those below are noise-dominated. The Marchenko-Pastur distribution (Marčenko & Pastur, 1967) provides the theoretical foundation for this threshold in the high-dimensional regime.

Layer Type Classification. We group weight matrices by their functional role to analyze layer-specific spectral properties: (1) **AttnQKV**: Query-Key-Value projection ($W_{QKV} \in \mathbb{R}^{d \times 3d}$); (2) **AttnProj**: Attention output projection ($W_O \in \mathbb{R}^{d \times d}$); (3) **MLP_FC**: MLP up-projection ($W_{\text{up}} \in \mathbb{R}^{d \times 4d}$); (4) **MLP_Proj**: MLP down-projection ($W_{\text{down}} \in \mathbb{R}^{4d \times d}$).

J Layer-wise Analysis

We analyze how the head-tail structure varies across layers in GPT-2XL.

J.1 Gradient Spectrum by Layer Type

Layer Type	Median κ	Top-64 Energy
Embedding	2,340	89%
Attention QKV	156	72%
Attention Out	89	68%
MLP Up	234	78%
MLP Down	178	75%
LM Head	1,890	91%

Observations: (1) Embedding and LM Head have highest condition numbers, concentrated energy—these benefit most from aggressive whitening; (2) attention layers have moderate condition numbers, so standard $k = 64$ is appropriate; (3) all layers show $>$68% energy in top-64 directions, validating the low-rank gradient structure.

J.2 Layer Depth Effects

Depth	Avg $\|g\|$	Avg κ	SNR Estimate
Layer 1–4	0.023	145	High
Layer 5–8	0.018	189	Medium
Layer 9–12	0.012	267	Lower

Deeper layers have: (1) smaller gradient norms (vanishing gradient effect); (2) higher condition numbers (more ill-conditioned curvature); (3) lower effective SNR, suggesting these layers benefit most from conservative tail handling.

J.3 Adaptive Per-Layer Rank

We experimented with layer-specific rank selection:

Strategy	Val Loss (8B)
Uniform $k = 64$	3.036
Higher for embed/LM head ($k = 128$)	3.031
Lower for deep layers ($k = 32$)	3.044
Full adaptive	3.028

Per-layer adaptation provides modest gains. For simplicity, we recommend uniform $k = 64$ with optional $k = 128$ for embedding layers.

K Hyperparameter Sensitivity Analysis

We analyze sensitivity to key DAWN hyperparameters on GPT-2 124M trained on FineWeb-Edu (8B tokens).

K.1 Head Rank k

Rank k	Val Loss	Relative Time
16	3.048	$0.45\times$
32	3.042	$0.55\times$
64	3.036	$0.61\times$
128	3.057	$0.75\times$
256	3.062	$1.02\times$
Muon (full rank)	3.024	$1.0\times$

Observations: (1) $k = 64$ provides the best quality-efficiency tradeoff; (2) $k < 32$ misses important gradient directions, causing degradation; (3) $k > 128$ shows diminishing returns—additional directions are noise-dominated; (4) All DAWN variants are faster than Muon due to periodic SVD refresh.

K.2 NS Iterations for Head

NS Iterations	Val Loss	Notes
0 (no NS)	4.213	Fails to converge
2	3.089	Insufficient orthogonalization
3	3.054	Near-converged
5 (default)	3.042	Fully converged
7	3.041	Marginal improvement

Observations: (1) NS iterations ≥ 5 achieve full orthogonalization; (2) Removing NS entirely causes catastrophic failure (+1.2 loss); (3) The cubic convergence of NS means 5 iterations suffice for machine precision.

K.3 SVD Refresh Period p

Period p	Val Loss
10	3.038
20	3.042
40	3.028
80	3.081
160	3.258

Observations: (1) $p = 40$ achieves optimal quality while being $0.61\times$ Muon compute; (2) Subspace iteration between refreshes maintains accuracy; (3) $p > 80$ causes subspace drift, degrading performance.

K.4 Tail Energy Allocation

Tail Scaling	Val Loss
$\sqrt{d-k}$ normalized (default)	3.042
Bounded energy	3.036
Adaptive scaling	3.035
Fixed $\gamma = 0.5$	3.068
Fixed $\gamma = 0.1$	3.095
No tail ($\gamma = 0$)	3.466

Observations: (1) Tail processing is essential—removing it causes +0.42 degradation; (2) Bounded/adaptive scaling slightly outperforms fixed normalization; (3) Under-weighting tail ($\gamma < 0.5$) loses important gradient information.

L Numerical Stability

We discuss numerical considerations for DAWN implementation.

L.1 SVD Stability

Randomized SVD is numerically stable when: (1) power iterations use orthonormalization between iterations (prevents numerical blow-up); (2) the target rank k is well-separated from the $(k+1)$-th singular value; (3) the oversampling parameter p is sufficient ($p \geq 10$ recommended).

We use the Halko-Martinsson-Tropp algorithm with modifications:

1. QR orthonormalization after each power iteration
2. Oversampling with $p = 10$ (compute $k + p$ columns, keep k)
3. Double-pass for improved accuracy when memory permits

L.2 Handling Ill-Conditioned Gradients

For extremely ill-conditioned gradients ($\kappa > 10^6$), we add safeguards: (1) **Epsilon regularization:** add ϵI to prevent division by zero; (2) **Gradient clipping:** apply before SVD to prevent overflow; (3) **Mixed precision:** compute SVD in float32, updates in bfloat16.

L.3 Memory Efficiency

DAWN's memory footprint: (1) random projection matrix Ω: $O(nk)$ (can be generated on-the-fly); (2) intermediate Y: $O(mk)$; (3) SVD factors: $O((m+n)k)$; (4) total additional memory: $O((m+n)k)$ vs $O(mn)$ for full SVD.

For $d = 4096$, $k = 64$: DAWN uses $\sim$0.5MB additional memory per layer vs $\sim$64MB for full matrix storage.

M Implementation Details

Hardware. All experiments were conducted on NVIDIA H200 GPUs (141GB HBM3 memory):

- **GPT-2 small model ablation (124M):** $4\times$ H200 GPUs
- **GPT-2XL (1.5B) and Wide Model (4B):** $8\times$ H200 GPUs

Software. PyTorch 2.1+ with bfloat16 mixed precision training. Gradient synchronization via DDP; SVD computation is local to each GPU (no additional communication overhead beyond standard gradient sync).