ORTH-DION: ELIMINATING GEOMETRIC MISMATCH IN DISTRIBUTED LOW-RANK SPECTRAL OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

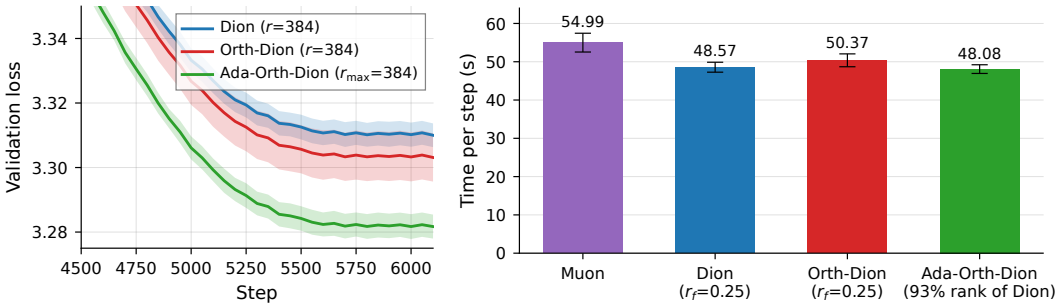
Low-rank gradient compression reduces communication in distributed training by representing updates with rank-r factors. Dion is a recent method that approximates Muon, a spectral optimizer that orthogonalizes momentum, using one step of power iteration followed by column normalization (rescaling each column of the right factor to unit length). This makes it compatible with fully sharded data parallel training, but it converges more slowly than full-rank spectral methods. We show that this gap is geometric: column normalization does not yield the rank-r polar factor that Muon implicitly targets, so the resulting direction violates the dual-norm constraint of the low-rank spectral geometry, and the rate picks up an extra factor of $\sqrt{r}$ even though the low-rank approximation of the gradient itself is accurate. The same mismatch enters the smoothness term and the error-feedback recursion in the analysis, which has a knock-on effect on empirical performance. We propose Orth-Dion, which replaces column normalization with QR orthogonalization of the right factor. Under non-Euclidean smoothness, with L_r the curvature constant along rank-r directions, Orth-Dion attains rate $O(\sqrt{L_r/T})$, matching exact spectral methods at the same per-step communication cost as Dion. The proof removes the bounded-drift assumption common in prior error-feedback analyses via a self-consistent fixed-point argument, and uses a time-averaged contraction that only requires the error sequence to contract on average rather than at every step. Experiments on large-scale language model pre-training validate the predicted $\sqrt{r}$ scaling and show that Orth-Dion closes the convergence gap to Muon at Dion's communication cost.

1 INTRODUCTION

Training large language models increasingly relies on distributed sharding schemes such as fully sharded data parallelism (Zhao et al., 2023), where optimizer design is constrained by both per-step computation and communication. Spectral optimizers such as Muon (Jordan et al., 2024) are attractive in this setting because they exploit the matrix structure of neural network parameters. Rather than applying an entrywise update, they orthogonalize momentum and take a step in a spectrally normalized direction. This geometry has been empirically effective, but it is difficult to scale under sharding, since forming and orthogonalizing full momentum matrices requires additional collective communication. Dion (Ahn et al., 2025b) addresses this bottleneck by replacing the full spectral update with a low-rank factorized one. Each matrix update is represented through rank-r factors, which are computed by one step of power iteration and communicated at cost proportional to $(m+n)r$ rather than mn.

At first glance, the remaining gap between Dion and full-rank spectral methods appears to be the unavoidable price of using rank-r updates (Ahn et al., 2025b; Carlson et al., 2015). Since Dion communicates only low-rank factors, one might expect its slower convergence to come from missing gradient directions outside the selected subspace. This interpretation is natural given the low-rank, power-iteration-based compression lineage of PowerSGD and Dion (Vogels et al., 2019; Ahn et al., 2025b). Under this view, the problem is where the update points, and Dion loses because its rank-r subspace is too small. We show that this explanation is incomplete. Dion's power iteration can identify a useful low-rank subspace, yet the subsequent update can still be shaped incorrectly inside that subspace. The source is Dion's final normalization step. After power iteration, Dion normalizes the columns of the right factor independently (Ahn et al., 2025b). This preserves the span of the factor,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) **Lower validation loss at matched rank.** LLaMA 320M, late-stage convergence (mean±2 std, full trajectory in Fig. 3).

(b) **Faster wallclock at scale.** Per-step time on a LLaMA 17.1B 50-step run (mean±2 std). Rank fraction r_f is the spectral update's rank divided by the parameter matrix's full rank.

Figure 1: **Proposed methods improve both convergence and wallclock. (a)** At matched rank, Orth-Dion and Ada-Orth-Dion reach Dion's plateau earlier and keep going. **(b)** Adaptive rank absorbs Orth-Dion's QR overhead and matches Dion's per-step time on the 17.1B model. (Ada-Orth-Dion's rank is pinned at $0.93\,r_f$, the steady-state rank reached on the 320M model; see App. L.2.)

but it does not produce the orthogonal right factor associated with a rank-r spectral update (Carlson et al., 2015; Bernstein & Newhouse, 2024b). As a result, Dion can move in essentially the same subspace as the desired low-rank polar direction while using a direction that is incorrectly scaled for spectral descent. We formalize this mismatch through a dual-norm factor ν_t, showing that column normalization can introduce a rank-dependent penalty even when the low-rank approximation itself is accurate.

Orth-Dion fixes this mismatch with the smallest possible change. It keeps Dion's power iteration, residual buffer, and low-rank communication pattern unchanged, but replaces column normalization with QR orthogonalization of the right factor. This makes the update geometrically aligned with the rank-r polar direction targeted by spectral methods. Under the same low-rank communication cost as Dion, Orth-Dion removes the rank-dependent normalization penalty and recovers the leading convergence rate of exact rank-r spectral updates. Because Dion-family methods use error feedback, the proof must control the coupled evolution of the residual buffer and the tracked subspace. We do this through a self-consistent residual-control argument and an amortized contraction condition, avoiding the bounded-buffer-drift assumption used in prior analyses.

We measure ν_t directly during Llama 3 320M pretraining. Dion's ν_t rises with rank across every layer and logged step; Orth-Dion's holds at $\nu_t \approx 1$. The mismatch is not a worst-case anomaly; it is what Dion does in the regime it is meant for. Holding r fixed and swapping column normalization for QR lowers validation loss at every rank we tested, which separates the QR fix from any rank-capacity effect. Because QR adds per-step compute, we pair it with adaptive rank to obtain Ada-Orth-Dion: the working rank shrinks toward each layer's intrinsic dimensionality, recovering Dion-like step time on a 17.1B model (Figure 1) while maintaining convergence gain. The combined method gives a low-rank distributed spectral optimizer with a better convergence–communication trade-off than Dion.

Contributions. We (i) identify a geometric source of suboptimality in Dion: column normalization inflates the Ky Fan dual norm by up to $\sqrt{r}$, producing a rank-dependent convergence penalty distinct from low-rank approximation error; (ii) propose **Orth-Dion**, a one-line ColNorm→QR replacement that forces ν_t=1 and recovers the exact low-rank spectral rate at Dion's communication cost; (iii) prove convergence under non-Euclidean smoothness via a self-consistent residual-control argument that eliminates bounded-buffer-drift and allows amortized rather than per-step contraction; and (iv) validate the mechanism and optimizer-level impact through direct ν_t measurements, matched-rank improvements on Llama 3 320M, and Ada-Orth-Dion matching Dion's wall-clock time at lower validation loss.

Algorithm 1 Orth-Dion / Stripped Dion (one step)

Require: Gradient $G_t \in \mathbb{R}^{m\times n}$, residual R_t, right factor V_{t-1}, step size η, EF coefficient β

1: $M_t \leftarrow G_t + R_t$ ▷ Buffer = gradient + error feedback
2: $U_t \leftarrow \mathrm{orth}(M_t V_{t-1})$ ▷ Left factor via QR, $O(mr^2)$
3: $W_t \leftarrow M_t^\top U_t$ ▷ Right factor (un-normalized), $O(mnr)$
4: $\bar{V}_t \leftarrow \mathrm{orth}(W_t)$ ▷ **Orth-Dion**: QR; *Dion*: $\mathrm{ColNorm}(W_t)$
5: $\hat{D}_t \leftarrow U_t \bar{V}_t^\top$ ▷ Low-rank update direction
6: $X_{t+1} \leftarrow X_t - \eta \hat{D}_t$ ▷ Parameter update
7: $R_{t+1} \leftarrow \beta\,(M_t - U_t(U_t^\top M_t))$ ▷ Error feedback
8: $V_t \leftarrow \bar{V}_t$ ▷ Warm-start for next step

2 Background: Geometry of Rank-Constrained Optimization

Steepest descent under norms. Given $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, steepest descent under norm $\|\cdot\|$ updates $X_{t+1} = X_t - \eta D_t^*$ with $D_t^* = \arg\max_{\|D\|\le 1}\langle \nabla f(X_t), D\rangle$. SGD uses ℓ_2; SignSGD (Bernstein et al., 2018) uses ℓ_∞; spectral methods use the operator norm. For distributed training, we constrain updates to rank $\le r$, yielding the *Ky Fan r-norm* geometry. The *Ky Fan r-norm* $\|A\|_{\mathrm{KF},r} = \sum_{i=1}^r \sigma_i(A)$ is the sum of the top-r singular values. Its dual is $\|D\|_{(r)} = \max\{\sigma_1(D),\ \|D\|_*/r\}$. Steepest descent in this geometry takes $D_t^* = \arg\max_{\|D\|_{(r)}\le 1}\langle M_t, D\rangle$. The solution is the rank-r polar factor $P_r(M_t) = U_r V_r^\top$ from the truncated SVD, achieving $\langle M_t, D_t^*\rangle = \|M_t\|_{\mathrm{KF},r}$.

The two algorithms. Both maintain buffer $M_t = G_t + R_t$ and use one power iteration step. The *sole difference* is the right-factor normalization:

	Stripped Dion[1]	**Orth-Dion (Proposed)**
Line 5:	$\bar{V}_t \leftarrow \mathrm{ColNorm}(W_t)$	$\bar{V}_t \leftarrow \mathrm{orth}(W_t)$ *(QR)*

where $W_t = M_t^\top U_t$ and $U_t = \mathrm{orth}(M_t V_{t-1})$. The update is $\hat{D}_t = U_t \bar{V}_t^\top$, and the error feedback is $R_{t+1} = \beta\,(I - U_t U_t^\top)\,M_t$. QR costs $O(nr^2)$ which is negligible compared to the $O(mnr)$ power step since $r \ll m$. The full procedure is given in Algorithm 1.

FSDP communication. Under FSDP, parameters are sharded across devices. Each training step involves: (1) all-gather to materialize full weights, (2) forward/backward pass, (3) reduce-scatter to re-shard gradients, (4) local optimizer step. Muon requires an *additional* all-gather/reduce-scatter pair for the full momentum matrix during orthogonalization. Dion and Orth-Dion avoid this by communicating only the low-rank factors ($O((m+n)r)$ vs. $O(mn)$), making them compatible with FSDP sharding at rank-proportional cost.

Assumptions. We use the following assumptions (details in Appendix A): **(A)** *Non-Euclidean smoothness*: $f(X+\Delta) \le f(X) + \langle \nabla f(X), \Delta\rangle + \frac{L_r}{2}\|\Delta\|_{(r)}^2$, where L_r captures curvature along low-rank directions; **(A′)** *Frobenius gradient-Lipschitz*: $\|\nabla f(Y) - \nabla f(X)\|_F \le L_F\|Y - X\|_F$ (independent of A in non-Euclidean geometry); **(B′)** *Gradient spectral gap*: $\sigma_r(G_t) - \sigma_{r+1}(G_t) \ge \Delta_{\mathrm{gap}} > 0$; **(B″)** *Small spectral tail*: $\sigma_{r+1}(G_t) \le \tau$ with $\tau = O(\eta)$ on the relevant horizon; **(C′)** *Gradient bounds*: $\|G_t\|_F \le G_F$, $\kappa_r(G_t) \le \kappa_G$.

One-step descent decomposition. Under Assumption A, any update $\hat{D}_t$ with $\|\hat{D}_t\|_{(r)} = \nu_t$ yields (proof in Appendix D):

$$f(X_{t+1}) \le f(X_t) - \eta\|G_t\|_{\mathrm{KF},r} + \eta\big(\delta_t + (1+\nu_t)\|R_t\|_{\mathrm{KF},r}\big) + \frac{L_r\nu_t^2}{2}\eta^2, \tag{1}$$

where $\delta_t = \|M_t\|_{\mathrm{KF},r} - \langle M_t, \hat{D}_t\rangle$ is the oracle defect. The dual norm factor ν_t appears in two critical places: the smoothness penalty $L_r\nu_t^2\eta^2/2$ and the residual coupling $(1+\nu_t)\|R_t\|_{\mathrm{KF},r}$.

[1] *Stripped Dion*: Dion's spectral update without the auxiliary scalar/Adam-side machinery (decoupled scalar LR, output-head LR scaling). Theorem 5.1 uses this form, but the QR correction is spectral-side only and transfers to full Dion, which all experiments use as the baseline (auxiliary machinery enabled for both); reported ν_t (Fig. 2) and matched-rank gains (Table 1) are in this setting.

3 THE $\sqrt{r}$ INEFFICIENCY

Column normalization produces $\bar{V}_t$ with unit-length columns, giving Gram matrix $\bar{V}_t^\top \bar{V}_t$ with diagonal entries 1 and potentially large off-diagonal entries. Since this is a correlation matrix:

Proposition 3.1 ($\sqrt{r}$ bound)**.** *For column-normalized $\bar{V}_t$: $\nu_t = \|\hat{D}_t\|_{(r)} = \|\bar{V}_t\|_{\mathrm{op}} \in [1, \sqrt{r}]$. Equality $\nu_t{=}1$ holds only when columns of W_t are already orthogonal.*

How ν_t enters the rate. Telescoping (1) and optimizing η:

$$\min_t \|G_t\|_{\mathrm{KF},r} \le \sqrt{\frac{2(f_0 - f_\infty) L_r \nu^2}{T}} + O(1/T), \quad \nu = \max_t \nu_t. \tag{2}$$

With $\nu{=}\sqrt{r}$ (ColNorm): rate $= O(\sqrt{L_r r/T})$. With $\nu{=}1$ (Orth-Dion): rate $= O(\sqrt{L_r/T})$. The $\sqrt{r}$ gap is purely geometric.

Geometric interpretation. The Ky Fan dual ball $\{D : \|D\|_{(r)} {\le} 1\}$ constrains both $\|D\|_{\mathrm{op}} {\le} 1$ and $\|D\|_F {\le} \sqrt{r}$. The polar factor $P_r(M_t)$ lies on its boundary ($\|\cdot\|_{\mathrm{op}}{=}1$, $\|\cdot\|_F{=}\sqrt{r}$). ColNorm preserves $\|\cdot\|_F{=}\sqrt{r}$ but allows $\|\cdot\|_{\mathrm{op}}$ up to $\sqrt{r}$—pushing the update *outside* the dual ball and incurring a factor-r smoothness penalty. The operator norm constraint $\|\hat{D}_t\|_{\mathrm{op}} {\le} 1$ is the binding constraint for any rank-r partial isometry. ColNorm violates it by concentrating singular values non-uniformly ($\sigma_{\max}(\hat{D}_t) \gg \sigma_{\min}(\hat{D}_t)$); QR orthogonalization forces all singular values to 1 by construction. The key insight: the inefficiency is not in *which subspace* is selected (Lemma 4.3 shows both algorithms select the same subspace) but in *how the update is scaled within that subspace*.

4 ORTH-DION: CORRECTING THE GEOMETRY

Lemma 4.1 (Partial isometry, $\nu_t{=}1$)**.** *If $U {\in} \mathbb{R}^{m\times r}$, $\bar{V} {\in} \mathbb{R}^{n\times r}$ both have orthonormal columns, then $\hat{D}{=}U\bar{V}^\top$ has $\|\hat{D}\|_{(r)}{=}1$.*

Lemma 4.2 (Non-negative defect)**.** *For Orth-Dion, $\delta_t \ge 0$ for all t.* (By Hölder: $\langle M_t, \hat{D}_t \rangle \le \|M_t\|_{\mathrm{KF},r} \cdot 1$.)

Lemma 4.3 (Same tracking)**.** *The projector error ε_t satisfies the same recursion for both algorithms, since $\mathrm{col}(\mathrm{orth}(W)) = \mathrm{col}(\mathrm{ColNorm}(W))$ and the error feedback depends only on U_t.*

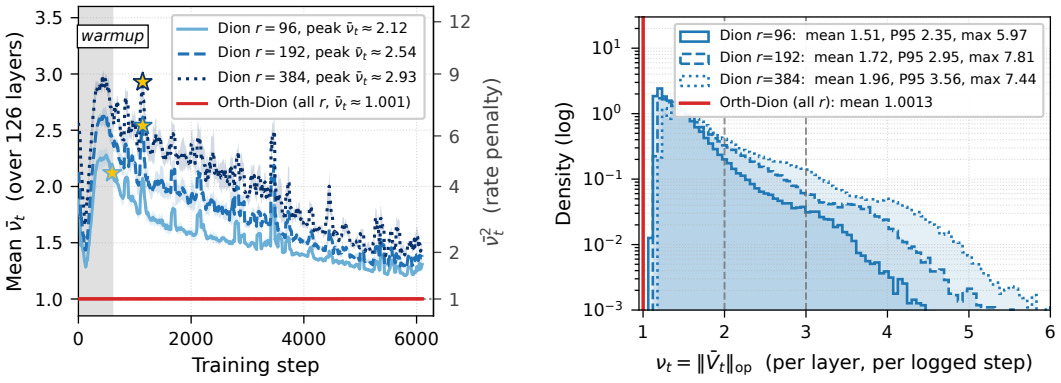
Consequence. The improvement is *entirely* in ν_t, not approximation quality. Lemma 4.3 ensures that both algorithms track the same gradient subspace, since the column spaces of $\mathrm{orth}(W)$ and $\mathrm{ColNorm}(W)$ are identical. At typical ranks, the quantitative impact is substantial. For $r{=}256$, the smoothness cost ratio is $L_r\nu^2/L_r = r = 256\times$, while the residual coupling ratio is $(1{+}\sqrt{r})/(1{+}1) = (1{+}16)/2 = 8.5\times$. These factors compound, inflating the leading convergence constant by $\sqrt{r}$ through both the smoothness and coupling terms.

The bound $\nu_t = \sqrt{r}$ is attained only when all columns of W_t are colinear. This pathological case does not arise in practice. But the inflation it predicts is real, rank-ordered, and persistent throughout training. Figure 2 measures ν_t during Llama 3 320M pretraining at $r \in \{96, 192, 384\}$. Figure 2a shows the time-course: Dion's curves separate cleanly by rank (peaks $\approx 2.1, 2.5, 2.9$, i.e. $\bar{v}_t^2 \approx 4.5{-}8.6$) and stay strictly above 1 throughout. Figure 2b shows the per-update dispersion: at $r{=}384$, $\sim 33\%$ of individual updates carry $\nu_t > 2$ and $\sim 10\%$ carry $\nu_t > 3$. In our runs ν_t never approaches the worst-case $\sqrt{r}$. It does climb with rank, though, and stays elevated through training. Orth-Dion's ν_t holds at 1 everywhere we measured. No hyperparameter setting fixes this; the problem is in the normalization step itself. Proofs of Lemmas 4.1–4.3 are in Appendix E.

5 CONVERGENCE THEORY

5.1 SELF-CONSISTENT DRIFT ELIMINATION

The original Dion analysis assumes bounded buffer drift ($\|M_t{-}M_{t-1}\|_{\mathrm{op}} \le \Delta_{\mathrm{drift}}$, Assumption D), but this is circular: $M_t{-}M_{t-1}$ contains the residual increment, which is itself $O(\|G_t\|)$ rather than small. We eliminate this assumption by exploiting an asymmetry: under Assumption A′,



(a) Layer-mean $\bar{\nu}_t$ vs. training step.

(b) Post-warmup pooled ν_t distribution.

Figure 2: **The dual-norm mismatch is real, rank-ordered, and persistent** (Llama 3 320M, $r \in \{96, 192, 384\}$, mean over 3 seeds; band $\pm$std). **(a)** Dion's $\bar{\nu}_t$ separates cleanly by rank and stays well above 1 throughout training; the right axis shows $\bar{\nu}_t^2$, the factor through which ν_t enters the smoothness term in (1). **(b)** Per-update ν_t pooled over layer, step, and seed: the inflation is dispersed, not transient. Orth-Dion collapses to $\nu_t \approx 1$ in both views (max ≈ 1.003).

the *gradient* subspace changes slowly ($\sin\Theta_{\max}(P_G^{t+1}, P_G^t) \le L_F\,\eta\sqrt{r}/\Delta_{\text{gap}} = O(\eta)$), while the buffer subspace may change abruptly. We track the former and *derive* the latter via a self-consistency loop on the coupled system $(\varepsilon_t, \mathcal{R}_t)$, where $\varepsilon_t = \sin\Theta_{\max}(P_M^t, P_G^t)$ is the tracking error and $\mathcal{R}_t = \|R_t\|_F/\|G_t\|_F$ the residual ratio. The recursions $\varepsilon_{t+1} \le \tilde{\gamma}_{\text{eff}}(1+\kappa_G)\varepsilon_t + O(\eta)$ and $\mathcal{R}_{t+1} \le h(\varepsilon_t, \mathcal{R}_t)$ form a contractive map Φ on $[0,1]^2$; for $\eta = O(\Delta_{\text{gap}}/(L_F\sqrt{r}\kappa_G))$, Banach fixed-point yields a unique fixed point with $\mathcal{R}_\infty, \varepsilon_\infty = O(\eta)$ *provided* the spectral tail is also small (Assumption B″); without B″ the conclusion is $\mathcal{R}_\infty = O(\tau + \eta)$. The residual recursion is controlled by the spectral tail $\sigma_{r+1}(M_t)$ plus a tracking-leakage term $\varepsilon_t\sigma_1(M_t)$ (Appendix F).

5.2 Amortized Contraction

The per-step contraction $\rho_t = \tilde{\gamma}_t(1+\kappa_{G,t})$ may exceed 1 during warmup or landscape transitions, so a uniform per-step requirement is too strong. We instead impose *uniform amortized contraction*: there exist $\bar{\rho} \in (0,1)$, $C_\rho \ge 1$ such that

$$\prod_{i=s}^{t-1} \rho_i \;\le\; C_\rho\,\bar{\rho}^{\,t-s}, \qquad 0 \le s < t.$$

Every suffix product decays geometrically up to a constant, allowing transient $\rho_t > 1$ while ruling out persistent growth. Under this condition the tracking-error recursion $\varepsilon_t \le \rho_{t-1}\varepsilon_{t-1} + c_{t-1}$ unrolls to geometric convergence up to a bounded bias term (Theorem G.1, Appendix G)—the natural Lyapunov-style stability condition for subspace tracking, where stability is governed by contraction over windows rather than at every step.

5.3 Main Theorem

Theorem 5.1 (Convergence of Orth-Dion)**.** *Under Assumptions A, B', C' with β=1, ρ'<1, and $\eta = c/\sqrt{T}$:*

$$\min_{0 \le t < T} \|G_t\|_{\text{KF},r} \le \sqrt{\frac{2(f_0 - f_\infty)L_r}{T}} + O\left(\frac{1}{T}\right) \tag{3}$$

The leading constant matches exact-SVD Muon, where ν=1, and improves on Stripped Dion's $\sqrt{2(f_0-f_\infty)L_r r/T}$, where $\nu=\sqrt{r}$, by a factor of $\sqrt{r}$. The full per-step cost is $O(mnr) + O(nr^2)$, matching Stripped Dion. The improvement is geometric, since Orth-Dion achieves ν=1 by construction and eliminates the rank-dependent inflation in the smoothness term. The proof appears in Appendix H.

6 Error Feedback and the $\beta < 1$ Regimes

The one-step descent bound (1) contains two rank-dependent terms: the smoothness penalty $L_r\nu_t^2\eta^2/2$, addressed in Sections 3–4 via ν_t=1, and the residual coupling $(1+\nu_t)\|R_t\|_{\mathrm{KF},r}$, whose dynamics depend on β. With fixed ν_t, the residual recursion $R_{t+1} = \beta\,(I - U_tU_t^\top)(G_t+R_t)$ is governed by the cross-step correlation of the out-of-subspace gradient. Characterizing this correlation yields a 0.315-nat improvement on GPT-2 45M at zero additional compute (Table 2) and shows that EF is mechanistically distinct from heavy-ball momentum. Define the *gradient subspace persistence* $\phi_t := \cos\angle\big((I-P_t)G_t,\ (I-P_{t-1})G_{t-1}\big)$ and the *R-norm ratio* $\mathcal{R}_t := \|R_t\|_F/\|G_t\|_F$. Details in Appendix K.

Regime 1 (Coherent, $\phi_t>0$): Out-of-subspace components persist across steps, so $R_T \approx \sum_t(I-P_t)G_t$ accumulates coherently. The subspace eventually rotates to capture this direction; β=1 is optimal.

Regime 2 (Stochastic, $\phi_t\approx0$): Typical of LLM training with stochastic mini-batches. The residual performs a random walk, acting as *implicit heavy-ball momentum* with effective coefficient $\hat\varepsilon^2$, where $\hat\varepsilon_t = \|(I-P_t)M_t\|_F/\|M_t\|_F$ is the tracking error. Reducing β below 1 boosts this effect at zero extra compute. The optimal β satisfies $\beta^* \approx 1-2\hat\varepsilon$; for GPT-2 with baseline tracking error $\hat\varepsilon\approx0.35$ measured at β=1, this predicts $\beta^* \ll 1$. Empirically, val loss decreases monotonically from β=1 down to $\beta\approx0.1$ (Table 2), and $\mathcal{R}_t^*$ grows from 0.3 to 3.2 over the same range, directly tracking the implicit-momentum accumulation the theory predicts. Explicit Polyak momentum is not a substitute: PolyakDion (μ=0.95, $R\equiv0$) reaches val. loss 3.162 vs. 3.056 for Dion at β=0.3, since EF reshapes the buffer *before* compression while explicit momentum acts in the full parameter space.

Regime 3 (Anti-correlated, $\phi_t<0$): Residual components cancel across steps; the buffer diverges and EF is harmful. $\beta\to0$ is necessary.

This framework extends EF21 (Richtárik et al., 2021), which assumes independent lost signal (Regime 2 only) and misses the coherent and anti-correlated cases. Full derivations are in Appendix K.

7 Ada-Orth-Dion: Adaptive Rank as Intrinsic Dimensionality

Although Orth-Dion achieves convergence improvement with a minimal fix, the extra QR orthogonalization adds compute overhead per step. To mitigate this and match Dion's wall-clock speed, we couple Orth-Dion with a per-layer adaptive rank mechanism to reduce the rank and the overhead, yielding *Ada-Orth-Dion*.

From contraction to intrinsic dimensionality. The contraction condition $\tilde\gamma_{i,r}(1+\kappa_{G,i,r}) < 1$ governing Theorem 5.1 decomposes into two competing terms per layer i: $\tilde\gamma_{i,r} = \sigma_{r+1}(M_{i,t})/\sigma_r(M_{i,t})$ *decreases* with r (more signal captured), while $\kappa_{G,i,r} = \sigma_1(G_{i,t})/\sigma_r(G_{i,t})$ *increases* with r (the tail is harder to track). The critical rank $r_i^* = \max\{r : \tilde\gamma_{i,r}(1+\kappa_{G,i,r}) < 1\}$ marks the signal-to-noise transition and is the layer's *intrinsic dimensionality*: above it, power iteration diverges; below it, the spectral update is provably contracting.

Lightweight estimator. We approximate r_i^* from Dion's existing factors with no extra SVD: $\widehat{\mathrm{erank}}(M_t) = \exp(-\sum_i \hat p_i \log \hat p_i)$ where $\hat p_i = \hat\sigma_i/\sum_j\hat\sigma_j$ and $\hat\sigma_i = \|[W_t]_i\|_2$. The rank is updated via EMA smoothing (α), a buffer multiplier (γ), clipping, and rounding to multiples of 8 for GPU efficiency (full policy in Algorithm 4, Appendix B.2).

QR remains essential at adaptive rank. Adaptive rank and QR address different problems. Adaptive rank lowers the working rank, which reduces compute and communication, but it does not change how the update is normalized: by Proposition 3.1, a ColNorm update still suffers the inflation at whatever rank is selected. QR removes this inflation (Lemma 4.1), so the two are complementary, and Ada-Orth-Dion uses both. Ada-Orth-Dion adjusts each layer's rank during training while staying below the fixed-rank budget, reaches the lowest validation loss among the low-rank methods (3.282), and reaches Dion's best loss in 18.3% fewer steps.

Table 1: Llama 3 320M pre-training on C4 (3.2B tokens, 8×GH200). Mean ± 2 std over 3 seeds. The steps reduction represents the percent reduction in steps to reach the Dion's minimum validation loss at equivalent rank. *Paired* Δ: for each seed, we subtract the Dion run's validation loss from the Orth-Dion run's (both runs use the same seed); the table reports the mean difference, the p-value of a paired t-test, and how many of the 3 seeds favour Orth-Dion. Negative values mean Orth-Dion is better. The ± ranges of the two runs overlap because the loss varies more across seeds than between methods, but Orth-Dion is better on every seed.

Method	Val. Ppl	Val. Loss	Final Step $\bar{\nu}_t$	Avg. Rank	Steps Reduction	Paired Δ vs. Dion
AdamW	28.46 ± 1.59	3.348 ± 0.055	—	—	—	—
Muon (Full, r=768)	25.92 ± 0.25	3.255 ± 0.010	1.00	full	—	—
Dion (r=96)	29.58 ± 0.25	3.387 ± 0.009	1.20 ± 0.01	96	—	—
Orth-Dion (r=96)	29.37 ± 0.09	3.380 ± 0.003	1.00	96	$13.1\% \pm 1.6\%$	-0.0072 (p=0.043, 3/3)
Dion (r=192)	28.37 ± 0.06	3.345 ± 0.002	1.28 ± 0.10	192	—	—
Orth-Dion (r=192)	28.16 ± 0.13	3.338 ± 0.005	1.00	192	$12.6\% \pm 1.9\%$	-0.0075 (p=0.010, 3/3)
Dion (r=384)	27.39 ± 0.10	3.310 ± 0.004	1.34 ± 0.01	384	—	—
Orth-Dion (r=384)	27.20 ± 0.20	3.303 ± 0.007	1.00	384	$12.3\% \pm 4.3\%$	-0.0069 (p=0.031, 3/3)
Ada-Orth-Dion ($r_{max} = 384$)	26.62 ± 0.10	3.282 ± 0.004	1.00	356.9 ± 12.2	$18.3\% \pm 1.0\%$	—

8 Experiments

Our experiments answer: (1) Does replacing ColNorm with QR improve convergence at matched rank, as predicted by Lemma 4.1? (2) Does Ada-Orth-Dion recover Dion's per-step wall-clock on a larger model? (3) How does the error-feedback coefficient β interact with the corrected geometry?

8.1 LLM Pre-training: Llama 3 320M

Setup. 320M-param Llama 3 (Grattafiori et al., 2024) on C4 (Raffel et al., 2020), 3.2B tokens (Chinchilla-optimal (Hoffmann et al., 2022)), seq len 2048, 8×GH200 with FSDP. All methods use cosine LR schedule with warmup, β=1.0 (unless stated), and identical hyperparameters except the normalization. Baselines: AdamW (β_1=0.95, β_2=0.95), Muon (full-rank with Newton–Schulz), Dion ($r \in \{96, 192, 384\}$ with ColNorm), Orth-Dion (same ranks with QR), and Ada-Orth-Dion (adaptive rank, maximum rank r_{max} at 384, rank initialized to r_{max}).

Findings (Table 1). *(a) Matched-rank improvement.* Orth-Dion lowers val. loss over Dion at every rank and reaches Dion's best loss in 12–13% fewer steps, which equates to $\approx$ 10% reduction in wall-clock time (Figure 3a). This confirms that replacing ColNorm with QR improves convergence. *(b) Direct theory validation.* Dion's measured $\bar{\nu}_t$ inflates monotonically with r (1.20→1.28→1.34) while Orth-Dion has $\bar{\nu}_t = 1.00$ across all ranks, confirming Proposition 3.1 and Lemma 4.1. *(c) Adaptive rank.* Ada-Orth-Dion beats all low-rank baselines at an average rank of 356.9. The rank reduction can lead to wall-clock time improvement in large scale models, as discussed in Section 8.3.

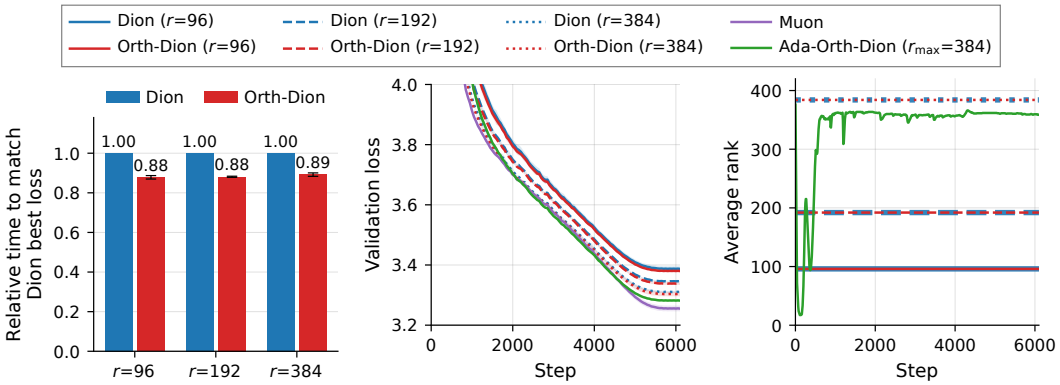
8.2 Measuring ν_t at LLM Scale

The central theoretical prediction is that Dion's convergence degrades as $\sqrt{r}$ while Orth-Dion is rank-independent. We validate this with two experiments.

Mechanism check. Figure 2 (§3) measures $\nu_t = \|\bar{V}_t\|_{\text{op}}$ for every matrix layer and logged step during Llama 3 320M pretraining at $r \in \{96, 192, 384\}$. Both the time-course (Figure 2a) and per-update distribution (Figure 2b) confirm Proposition 3.1: Dion's ν_t inflates monotonically with r throughout training, while Orth-Dion saturates the partial-isometry bound $\nu_t = 1$ across every (rank, layer, step) sample (Lemma 4.1).

Matched-rank, hyperparameter-controlled comparison. The one-step descent bound (1) couples three rank-dependent terms: the captured-gradient mass $\|G_t\|_{\text{KF},r}$ and the residual $\|R_t\|_{\text{KF},r}$ both improve with r, while the smoothness penalty $L_r \nu_t^2 \eta^2 / 2$ inflates with ν_t^2. Their relative weight is regime-dependent, and the matched-r Orth-Dion vs. Dion comparison, where the only difference is ColNorm vs. QR on Line 4 of Algorithm 1, is the clean isolation of the geometric correction.

On Llama 3 320M, the residual and descent gains from a larger r outweigh the smoothness penalty, so Dion's val. loss falls as r grows (Figure 3b). Dion's $\bar{\nu}_t$ does rise with rank, from 1.20 at r=96 to 1.34



(a) Wall-clock time to match Dion, normalized to mean Dion loss=1. Orth-Dion: mean±2 std.

(b) Validation loss on c4_validation, logged every 50 steps.

(c) Mean per-block effective rank during training.

Figure 3: **Training dynamics for Llama 3 320M (6,100 steps, 8×GH200 FSDP). (a)** Orth-Dion reaches Dion's best loss at matched rank in $\sim 0.89\times$ wall-clock time across $r \in \{96, 192, 384\}$. **(b)** At every rank, Orth-Dion lands at lower C4 validation loss than Dion, consistent with Lemma 4.1. **(c)** Ada-Orth-Dion (init r=384) has $r \approx 357$ on average; the fixed-rank curves are flat by construction.

at r=384, but the residual benefit at higher r wins. With r fixed, the only difference between Dion and Orth-Dion is ColNorm vs. QR, which translates directly to the dual-norm inflation. Orth-Dion beats Dion at every rank we tested (Table 1).

8.3 Wall-Clock on Llama 3 17.1B

Figure 1b measures per-step time on Llama 3 with 17.1B parameters on 8×A40 with FSDP full-shard. The largest per-block matrix is 16384×43776, roughly 450× the size of the largest 320M matrix (768×2048). Combined with A40s' lower per-link bandwidth than NVLink-class accelerators, this pushes NCCL collectives into the bandwidth-bound regime, where communication volume dominates step time and the gap between low-rank and full-rank updates is most visible. Two findings: *(i)* fixed-rank Orth-Dion is slightly slower than Dion at matched r due to the QR step, and *(ii)* Ada-Orth-Dion closes that gap and matches Dion's per-step time. Combined with the convergence advantage of Section 8.1, Ada-Orth-Dion gives the same per-step time as Dion at lower validation loss. Full architecture, hardware, and rank-policy details are in Appendix L.2.

8.4 Error Feedback: β-Sweep

Table 2 validates the $\beta<1$ regime on GPT-2 45M (wikitext-103, r=64, 30k steps), where $\phi_t \approx 0$ (Regime 2). Full 9-point sweep in App. K. Val loss decreases monotonically as β is reduced from 1.0 down to $\beta\approx0.1$, with the empirical optimum at $\beta\approx0.1$ (val. loss 2.978 vs. 3.293 for β=1), a 0.315-nat improvement obtained at zero additional compute, in line with the theoretical prediction $\beta^* \ll 1$. The R-norm ratio $\mathcal{R}_t^*$ grows from 0.3 at β=1 to 3.2 at β=0.1, directly tracking the predicted implicit-momentum accumulation. Disabling EF entirely (β=0) hurts by 0.261 nats relative to β=1, confirming that EF's contribution is essential even when heavily damped.

Table 2: β-sweep on GPT-2 45M/WikiText-103. Smaller β improves loss until $\beta\approx0.05$–0.1 while $\mathcal{R}_t^*$ grows; disabling EF is worse.

β	**Val loss**	Δ **vs.** β=1	$\mathcal{R}_t^*$
0.05	2.982	−0.311	4.6
0.1	**2.978**	−0.315	3.2
0.3	3.056	−0.237	1.7
0.5	3.111	−0.183	1.2
1.0 (baseline)	3.293	—	0.3
0 (no EF)	3.554	+0.261	—

9 Analysis and Discussion

The $\sqrt{r}$ factor is rank-dependent, and rank is the parameter practitioners scale up. The realized ν_t in our Llama 3 runs sits well below the worst-case $\sqrt{r}$ envelope (Fig. 2), so the absolute gap at any

fixed rank is gentler than the worst case; nonetheless the inflation is monotone in r and persists across layers and training steps. This gap is geometric rather than statistical. Lemma 4.3 shows both algorithms track the same gradient subspace with identical error, so ColNorm does not lose information. It produces a direction that overshoots the dual-norm unit ball. The QR fix removes the inflation from the norm geometry of the update, independent of how tight the worst-case bound is.

10 Related Work

Spectral-norm optimizers. Shampoo (Gupta et al., 2018) and SOAP (Vyas et al., 2024) precondition with Kronecker factors of second-moment statistics. Muon (Jordan et al., 2024) instead takes the orthogonal polar factor of the gradient. That is the closed-form steepest-descent direction under the spectral norm in the modular-norm framework of Bernstein & Newhouse (2024a) computed via Newton–Schulz iteration. All three require $O(mn)$ communication per matrix.

Error feedback for biased compressors. Error feedback originates with 1-bit SGD (Seide et al., 2014) and was developed through QSGD (Alistarh et al., 2017), sparsified SGD (Stich et al., 2018), and the EF21 framework (Richtárik et al., 2021). Karimireddy et al. (2019) showed EF restores convergence for sign-based methods. All assume Euclidean smoothness, under which our setting's bounds are loose by $\min(m, n)/r$. Our ϕ_t-framework adapts the EF analysis to operator-norm geometry and resolves the cross-step correlation between residual and signal that standard analyses implicitly absorb into a worst-case constant (Regimes 1 and 3).

Low-rank and adaptive distributed training. PowerSGD (Vogels et al., 2019) performs power iteration under Frobenius geometry, where column normalization is harmless. The Ky Fan mismatch is specific to the operator-norm setting. Dion (Ahn et al., 2025b) inherits ColNorm from this lineage and (as we show) pays a $\sqrt{r}$ penalty for it. Dion2 (Ahn et al., 2025a) selects row/column subsets without touching the normalization step. GaLore (Zhao et al., 2024) and ReLoRA (Lialin et al., 2024) operate in Frobenius geometry and so neither encounter nor benefit from the QR fix. MuonBP (Khaled et al., 2025) keeps full-rank updates but synchronizes only periodically, addressing communication along an orthogonal axis. For adaptive rank, AdaLoRA (Zhang et al., 2023) prunes singular values by importance during fine-tuning, and adaptive GaLore variants set rank from gradient-projection statistics. Our contraction condition instead ties per-layer rank to the gradient spectrum's signal-to-noise transition in the from-scratch optimization regime, where the relevant geometry is operator-norm rather than Frobenius.

11 Conclusion

We identified a $\sqrt{r}$ geometric inefficiency in low-rank spectral optimization. Column normalization inflates the dual norm ν_t by up to $\sqrt{r}$, producing a convergence penalty that grows with rank. Replacing ColNorm with QR orthogonalization forces $\nu_t{=}1$ and recovers the exact polar rate targeted by Muon. Direct ν_t measurement on Llama 3 320M confirms the predicted rank-dependent inflation in Dion and that QR removes it identically.

Ada-Orth-Dion couples the geometric fix with per-layer adaptive rank that targets each layer's intrinsic dimensionality. The two corrections are complementary: QR removes the rank-dependent inflation, and adaptive rank reduces the working rank to offset QR's per-step compute overhead. It achieves the lowest validation loss (3.282) of any low-rank method evaluated, reaches within 0.027 nats of full-rank Muon (3.255) at substantially lower per-step communication, and attains Dion's best loss in 18.3% fewer steps. On the 17.1B model (Figure 1b, Appendix L.2), this convergence advantage holds at Dion-comparable per-step time. Together, the fixed-rank Orth-Dion result establishes the geometric correction, and Ada-Orth-Dion turns it into a practical low-rank spectral optimizer with a strictly better convergence–communication trade-off than Dion.

Limitations. Our analysis assumes deterministic gradients. Extending to the stochastic setting requires bounding the interaction between mini-batch noise and subspace tracking. The adaptive rank policy uses a heuristic estimator. Tighter connections to the contraction condition would enable provable rank guarantees. Full 17.1B convergence is untested due to the compute limit.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing code for the experiments, to help derive proofs of the theoretical results, and to edit the paper for readability. We have not used generative AI tools to generate research ideas or produce experimental results. We have reviewed all AI-assisted work and take responsibility for the final content of this work.

REFERENCES

Kwangjun Ahn, Noah Amsel, and John Langford. Dion2: A simple method to shrink matrix in muon. *arXiv preprint arXiv:2512.16928*, 2025a.

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*, 2025b.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, 2017.

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

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *OPT 2024: Optimization for Machine Learning*, 2024b. doi: 10.48550/arXiv.2409.20325. URL https://openreview.net/forum?id=ux18f5nOpD. NeurIPS 2024 Workshop; also available as arXiv:2409.20325.

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

David E. Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned spectral descent for deep learning. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2971–2979, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/f50a6c02a3fc5a3a5d4d9391f05f3efc-Abstract.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Angela Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 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. Accessed: 2025-7-3.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *International Conference on Machine Learning*, pp. 3252–3261. PMLR, 2019.

Ahmed Khaled, Kaan Ozkara, Tao Yu, Mingyi Hong, and Youngsuk Park. Muonbp: Faster muon via block-periodic orthogonalization. *arXiv preprint arXiv:2510.16981*, 2025.

Vladislav Lialin, Sherin Muckatira, Namrata Shivagunde, and Anna Rumshisky. ReLoRA: High-rank training through low-rank updates. In *International Conference on Learning Representations*, 2024. arXiv:2307.05695.

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

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34, 2021.

Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Convergence bound and critical batch size of muon optimizer. *arXiv preprint arXiv:2507.01598*, 2025.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems*, 2018.

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

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023.

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

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Proceedings of the VLDB Endowment*, 16(12):3848–3860, 2023.

APPENDIX

A ASSUMPTIONS (FULL STATEMENTS)

Assumption A.1 (Non-Euclidean smoothness)**.** There exists $L_r > 0$ such that for all X, Δ with $\mathrm{rank}(\Delta) \le r$:

$$f(X+\Delta) \le f(X) + \langle \nabla f(X), \Delta \rangle + \frac{L_r}{2}\|\Delta\|_{(r)}^2.$$

This uses the Ky Fan dual norm $\|\cdot\|_{(r)}$ rather than the Frobenius norm. For rank-r updates $\Delta = U\bar{V}^\top$ with orthonormal $U, \bar{V}$: $\|\Delta\|_{(r)} = 1$, so the smoothness cost reduces to $L_r/2$ independent of r.

Assumption A is a one-sided function-value inequality. In Frobenius/Euclidean geometry, a matching descent lemma and a gradient-Lipschitz bound are equivalent; this equivalence *fails* for the KF dual norm $\|\cdot\|_{(r)}$. Assumption A therefore does not by itself imply any bound of the form $\|\nabla f(Y) - \nabla f(X)\|_\bullet \le L_r \|Y - X\|_\bullet$. Steps that need such a bound (the gradient-subspace drift in Appendix F and the resulting fixed-point argument) require Assumption A′ below.

Assumption A.2 (Gradient Lipschitz in Frobenius – A′)**.** There exists $L_F > 0$ such that for all $X, Y \in \mathbb{R}^{m\times n}$:

$$\|\nabla f(Y) - \nabla f(X)\|_F \le L_F \|Y - X\|_F.$$

This is the standard Euclidean gradient-Lipschitz condition. It is independent of Assumption A in general, and is the constant that drives the Wedin-type gradient-subspace drift bound used in Appendix F. (Where the framework manuscript refers to a drift constant of $L_r\sqrt{r}$, the rigorous constant is $L_F\sqrt{r}$.)

Assumption A.3 (Gradient spectral gap – B′)**.** There exist $\Delta_{\mathrm{gap}} > 0$ and $\sigma_{\min} > 0$ such that for all $t \ge 0$: $\sigma_r(G_t) - \sigma_{r+1}(G_t) \ge \Delta_{\mathrm{gap}}$ and $\sigma_r(G_t) \ge \sigma_{\min}$. This is a property of the *gradient* (the optimization landscape), not the algorithm's internal state. It replaces the original Dion framework's Assumption B, which imposed a gap on the *buffer* M_t—a quantity that depends on the algorithm's own error feedback and is therefore circular.

Local validity. Assumption B′ is intended to hold over the regime in which the gradient still has its meaningful rank-r structure (early/mid training). Near convergence, where $\|G_t\|_{\mathrm{KF},r} \to 0$, the lower bound $\sigma_r(G_t) \ge \sigma_{\min}$ may fail; the convergence results should be read as applying over any horizon on which B′ is in force.

Assumption A.4 (Small spectral tail – B″)**.** There exists $\tau \ge 0$ such that for all $t \ge 0$: $\sigma_{r+1}(G_t) \le \tau$, with $\tau = O(\eta)$ in the regime where the convergence theorem is asserted (cf. Appendix F).

Why this is needed. The self-consistent residual recursion of Appendix F yields the operator-norm bound $\|R_{t+1}\|_{\mathrm{op}} \le \sigma_{r+1}(M_t) + \varepsilon_t \sigma_1(M_t)$. Assumption B′ controls the *gap* but not the *tail*; without B″, the floor $\sigma_{r+1}(M_t)$ is not $O(\eta)$, and the advertised $R_\infty = O(\eta)$ conclusion does not follow. With B″, the boxed rate of Theorem 5.1 acquires an additive $O(\tau)$ term, which is absorbed when $\tau = O(\eta)$. When B″ fails (large tail), the rate becomes $\sqrt{2(f_0 - f_\infty)L_r/T} + O(\tau) + O(1/T)$, i.e. a floor of order τ.

Assumption A.5 (Gradient bounds – C′)**.** There exist $G_F, G_K, \kappa_G > 0$ such that for all $t \ge 0$: $\|G_t\|_F \le G_F$, $\|G_t\|_{\mathrm{KF},r} \le G_K$, $\kappa_r(G_t) \le \kappa_G$.

Notational map for stability constants. Three closely related contraction symbols appear across the paper. To keep them straight:

- $\rho_t := \tilde{\gamma}_t(1 + \kappa_{G,t})$ is the *instantaneous* per-step contraction.
- $\rho' < 1$ denotes a *uniform per-step* bound $\rho_t \le \rho'$ for all t (used in Theorem 5.1 for the cleanest statement).
- $\bar{\rho} \in (0,1)$ together with $C_\rho \ge 1$ are the *amortized / suffix-product* constants used when $\rho_t > 1$ is allowed transiently: $\prod_{i=s}^{t-1} \rho_i \le C_\rho\, \bar{\rho}^{t-s}$ for all $s < t$ (Section 5.2).

The uniform condition is the special case $\bar{\rho} = \rho'$, $C_\rho = 1$. Theorem statements that say "$\rho' < 1$" are tight conditions; the amortized version (Theorem G.1) yields the same conclusion under the suffix-product condition above.

B Algorithm Details

B.1 Full Algorithms

Algorithm 2 Stripped Dion (Baseline)

Input: learning rate η, error-feedback coefficient β, rank r
for $t = 0, 1, \ldots, T-1$ **do**
  $M_t \leftarrow G_t + R_t$ *(buffer = gradient + residual)*
  $U_t \leftarrow \mathrm{orth}(M_t V_{t-1})$ *(one step of subspace iteration)*
  $W_t \leftarrow M_t^\top U_t, \quad \bar{V}_t \leftarrow \mathrm{ColNorm}(W_t)$ ***(column normalization)***
  $\hat{D}_t \leftarrow U_t \bar{V}_t^\top$ *(rank-r update direction)*
  $X_{t+1} \leftarrow X_t - \eta_t \hat{D}_t$
  $B_t \leftarrow U_t U_t^\top M_t, \quad R_{t+1} \leftarrow \beta\,(M_t - B_t)$ *(error feedback)*
end for

Algorithm 3 Orth-Dion (Proposed)

Input: learning rate η, error-feedback coefficient β, rank r
for $t = 0, 1, \ldots, T-1$ **do**
  $M_t \leftarrow G_t + R_t$ *(buffer = gradient + residual)*
  $U_t \leftarrow \mathrm{orth}(M_t V_{t-1})$ *(one step of subspace iteration)*
  $W_t \leftarrow M_t^\top U_t, \quad \bar{V}_t \leftarrow \mathrm{orth}(W_t)$ ***(QR orthogonalization)***
  $\hat{D}_t \leftarrow U_t \bar{V}_t^\top$ *(rank-r update direction)*
  $X_{t+1} \leftarrow X_t - \eta_t \hat{D}_t$
  $B_t \leftarrow U_t U_t^\top M_t, \quad R_{t+1} \leftarrow \beta\,(M_t - B_t)$ *(error feedback)*
end for

B.2 Ada-Orth-Dion Rank Selection

Definition B.1 (Estimated effective rank)**.** Given power-iteration factors $U_t \in \mathbb{R}^{m\times r}$ and $W_t = M_t^\top U_t \in \mathbb{R}^{n\times r}$, define $\hat{\sigma}_i = \|[W_t]_i\|_2$ and:

$$\widehat{\mathrm{erank}}(M_t) = \exp\Big(-\sum_{i=1}^{r} \hat{p}_i \log \hat{p}_i\Big), \qquad \hat{p}_i = \frac{\hat{\sigma}_i}{\sum_j \hat{\sigma}_j}.$$

Algorithm 4 Ada-Orth-Dion Rank Selection

Require: estimate e_t, current rank r_t, EMA $\bar{e}_{t-1}$, bounds $r_{\min}, r_{\max}$, smoothing α, buffer γ
$\bar{e}_t \leftarrow \alpha\, e_t + (1-\alpha)\,\bar{e}_{t-1}$
$r_{\mathrm{target}} \leftarrow \lceil \gamma\, \bar{e}_t \rceil$
$r_{t+1} \leftarrow \mathrm{clip}(r_{\mathrm{target}},\ r_{\min},\ r_{\max})$, rounded to multiple of 8
if $r_{t+1} \neq r_t$ **then**
  Resize U, V to rank r_{t+1} (truncate or pad)
end if

Connection to contraction condition. The estimated effective rank is a spectral entropy measure. When the gradient has a sharp spectral gap at position k, the effective rank concentrates near k regardless of the nominal r. By tracking $\widehat{\mathrm{erank}}$ and adjusting r_t, Ada-Orth-Dion approximates $r_i^* = \max\{r : \tilde{\gamma}_{i,r}(1+\kappa_{G,i,r}) < 1\}$ without explicit SVD.

C Norms and Key Quantities

For $A \in \mathbb{R}^{m\times n}$ with singular values $\sigma_1(A) \geq \sigma_2(A) \geq \ldots$:

Notation	Name	Definition
$\|A\|_F$	Frobenius	$\sqrt{\sum_{i,j} A_{ij}^2} = \sqrt{\sum_i \sigma_i^2}$
$\|A\|_{\text{op}}$	Operator (spectral)	$\sigma_1(A)$
$\|A\|_*$	Nuclear	$\sum_i \sigma_i(A)$
$\|A\|_{\text{KF},r}$	Ky Fan r-norm	$\sum_{i=1}^r \sigma_i(A)$
$\|A\|_{(r)}$	Dual of KF r-norm	$\max\{\sigma_1(A),\ \|A\|_*/r\}$
$\kappa_r(A)$	Condition at rank r	$\sigma_1(A)/\sigma_r(A)$

Key Hölder identity. $\langle A, B\rangle \le \|A\|_{\text{KF},r} \cdot \|B\|_{(r)}$, with equality *attained* at $B = P_r(A)$ (the rank-r polar factor): then $\|P_r(A)\|_{(r)} = 1$ and $\langle A, P_r(A)\rangle = \|A\|_{\text{KF},r}$. Equality is not in general unique — e.g. for r=1 and $A = \text{diag}(1,1)$, both $B_1 = \text{diag}(1,0)$ and $B_2 = \text{diag}(0,1)$ saturate the bound.

Key quantities per iteration. $\varepsilon_t = \|U_t U_t^\top - U_{r,t} U_{r,t}^\top\|_{\text{op}}$ (projector error), $\delta_t = \|M_t\|_{\text{KF},r} - \langle M_t, \hat{D}_t\rangle$ (oracle defect), $\nu_t = \|\hat{D}_t\|_{(r)}$ (dual norm factor), $\tilde{\gamma}_t = \sigma_{r+1}(M_t)/\sigma_r(M_t)$ (buffer spectral ratio).

D Proof of One-Step Descent (Equation 1)

Proof. **Step 1 (Smoothness):** $f(X_{t+1}) \le f(X_t) + \langle G_t, -\eta \hat{D}_t\rangle + \frac{L_r}{2}\|\eta \hat{D}_t\|_{(r)}^2 = f(X_t) - \eta\langle G_t, \hat{D}_t\rangle + \frac{L_r \nu_t^2}{2}\eta^2$.

Step 2 (Gradient-buffer decomposition): $\langle G_t, \hat{D}_t\rangle = \langle M_t, \hat{D}_t\rangle - \langle R_t, \hat{D}_t\rangle$.

Step 3 (LMO defect): $\langle M_t, \hat{D}_t\rangle = \|M_t\|_{\text{KF},r} - \delta_t$.

Step 4 (Triangle inequality): $\|M_t\|_{\text{KF},r} \ge \|G_t\|_{\text{KF},r} - \|R_t\|_{\text{KF},r}$.

Step 5 (Hölder for R_t term): $|\langle R_t, \hat{D}_t\rangle| \le \|R_t\|_{\text{KF},r} \cdot \nu_t$.

Assembly: $-\langle G_t, \hat{D}_t\rangle \le -\|G_t\|_{\text{KF},r} + \delta_t + (1+\nu_t)\|R_t\|_{\text{KF},r}$. Substituting into Step 1 completes the proof. □

E Orth-Dion Proofs

Proof of Lemma 4.1. $\hat{D}^\top \hat{D} = \bar{V} U^\top U \bar{V}^\top = \bar{V}\bar{V}^\top$ is an orthogonal projector with eigenvalues $\{0,1\}$. Therefore $\hat{D}$ has r singular values equal to 1 and the rest 0. Hence $\sigma_1(\hat{D}) = 1$ and $\|\hat{D}\|_* = r$, so $\|\hat{D}\|_{(r)} = \max\{1,\ \|\hat{D}\|_*/r\} = \max\{1,1\} = 1$. □

Proof of Lemma 4.2. By Hölder: $\langle M_t, \hat{D}_t\rangle \le \|M_t\|_{\text{KF},r} \cdot \|\hat{D}_t\|_{(r)} = \|M_t\|_{\text{KF},r} \cdot 1$. So $\delta_t = \|M_t\|_{\text{KF},r} - \langle M_t, \hat{D}_t\rangle \ge 0$.

Significance: For ColNorm, $\nu_t > 1$ means $\hat{D}_t$ can violate the Hölder bound, making δ_t potentially negative and complicating the analysis. □

Proof of Lemma 4.3. The tracking recursion depends on three quantities: (1) Column space of V_{t-1}: $\text{col}(\text{orth}(W)) = \text{col}(W) = \text{col}(\text{ColNorm}(W))$ for full-rank W. (2) Spectral gap of M_t: unchanged between algorithms. (3) Buffer drift $\|M_t - M_{t-1}\|_{\text{op}}$: depends on η, β, unchanged. Therefore ε_t satisfies the same recursion and bounds. □

Defect bound for Orth-Dion. Under the projector-level bound with Lemma 4.3: $0 \le \delta_t \le C'_\delta \|M_t\|_F \cdot \varepsilon_t$ where $C'_\delta = (1+\kappa_G)\sqrt{2r}$. The constant is the same order as for Stripped Dion; the improvement comes entirely from ν_t.

F Self-Consistent Drift Elimination (Full Details)

F.1 Why the Original Assumption D Was Circular

Assumption D required $\|M_t - M_{t-1}\|_{\text{op}} \le \Delta_{\text{drift}}$. But $M_t - M_{t-1} = (G_t - G_{t-1}) + (R_t - R_{t-1})$. Under Assumption A′ (Frobenius gradient-Lipschitz, Appendix A.2), the gradient difference is small: $\|G_{t+1} - G_t\|_F \le L_F \|X_{t+1} - X_t\|_F = L_F \eta \|\hat{D}_t\|_F \le L_F \eta \sqrt{r}$. The residual difference is *not* small: $R_{t+1} - R_t = \beta\,(I - U_t U_t^\top) M_t - R_t$, so $\|R_{t+1} - R_t\|_{\text{op}} = O(\|G_t\|)$. Assumption D is inconsistent with the algorithm's actual dynamics.

F.2 Gradient Subspace Stability

Proposition F.1 (Gradient subspace drift)**.** *Under Assumptions A′ and B′:* $\sin\Theta_{\max}(P_G^{t+1}, P_G^t) \le L_F\,\eta\,\sqrt{r}/\Delta_{\text{gap}}$.

Proof. By A′: $\|G_{t+1} - G_t\|_F \le L_F\,\|X_{t+1} - X_t\|_F = L_F \eta \|\hat{D}_t\|_F \le L_F \eta \sqrt{r}$, where the last step uses $\|\hat{D}_t\|_F \le \sqrt{r}$ (every $\hat{D}_t$ has at most r singular values, each at most 1). Wedin's $\sin\Theta$ theorem with eigengap Δ_{gap} then yields the stated bound.

Remark on the constant. An earlier version of this proof attempted to derive a Frobenius gradient-Lipschitz bound from Assumption A. That step is not valid: A is a one-sided function-value smoothness inequality in the KF dual norm and does not imply a Frobenius gradient-Lipschitz estimate. We therefore introduce A′ as a separate assumption and state the drift bound with constant $L_F\sqrt{r}$, not $L_r\sqrt{r}$. This substitution propagates through all downstream uses of the drift bound (in particular, the fixed-point map below). □

F.3 Buffer Spectral Gap Transfer

Proposition F.2. *Under Assumptions B', C', if* $\|R_t\|_{\text{op}} \le \Delta_{\text{gap}}/2$*: (1)* $\sigma_r(M_t) \ge \sigma_{\min} - \Delta_{\text{gap}}/2 > 0$*; (2)* $\tilde{\gamma}_{\text{eff}} \le (\sigma_{r+1}(G_t) + \|R_t\|_{\text{op}})/(\sigma_r(G_t) - \|R_t\|_{\text{op}})$*; (3) Buffer gap:* $\sigma_r(M_t) - \sigma_{r+1}(M_t) \ge \Delta_{\text{gap}} - 2\|R_t\|_{\text{op}}$*. All follow from Weyl's inequality.*

F.4 Self-Consistent Residual Bound

With $\beta = 1$ and projector error ε_t, the residual update yields the operator-norm bound

$$\|R_{t+1}\|_{\text{op}} \;\le\; \sigma_{r+1}(M_t) + \varepsilon_t \cdot \sigma_1(M_t).$$

This decomposition has two components: the spectral *tail* $\sigma_{r+1}(M_t)$ (a property of the buffer's low-rank approximation error, controlled by Assumption B″ together with Weyl-type buffer-vs-gradient comparison), and the tracking-leakage term $\varepsilon_t\sigma_1(M_t)$ (controlled by the contraction $\rho' < 1$).

Assumption B′ controls only the gradient gap, not the tail; even with perfect tracking ($\varepsilon_t = 0$) the floor $\sigma_{r+1}(M_t)$ remains, and Weyl gives only $\sigma_{r+1}(M_t) \le \sigma_{r+1}(G_t) + \|R_t\|_{\text{op}}$. The advertised conclusion $R_\infty = O(\eta)$ therefore requires the additional small-tail Assumption B″ (Appendix A.4); with $\sigma_{r+1}(G_t) \le \tau$, the recursion has a fixed point $R_\infty = O(\tau + \eta)$, which collapses to $O(\eta)$ when $\tau = O(\eta)$. When B″ fails (e.g. on heavy-tailed gradients), the conclusion weakens to a tail floor of order τ in Theorem 5.1.

F.5 The Fixed Point

Proposition F.3 (Self-consistent fixed point)**.** *Assume A, A′, B′, B″ (with tail* τ*), C′, and let* $\rho' = \tilde{\gamma}_{\text{eff}}(1 + \kappa_G) < 1$*. For* $\eta = O(\Delta_{\text{gap}}/(L_F\sqrt{r}\kappa_G))$*, the coupled system*

$$\varepsilon'_\infty = \frac{\tilde{\gamma}_{\text{eff}} \cdot D_{\text{eff}}}{1 - \rho'}, \quad R_\infty \le \underbrace{\sigma_{r+1}(M)}_{\le \tau + \|R_\infty\|_{\text{op}}\ (B''+\text{Weyl})} + \varepsilon'_\infty \sigma_1,$$

$$D_{\text{eff}} = \frac{L_F\eta\sqrt{r} + R_\infty}{\Delta_{\text{gap}} - 2R_\infty}, \quad \tilde{\gamma}_{\text{eff}} = \frac{\sigma_{r+1}(G) + R_\infty}{\sigma_r(G) - R_\infty}$$

has a unique small fixed point with $R_\infty = O(\tau + \eta)$ and $\varepsilon'_\infty = O(\eta)$. In particular, when $\tau = O(\eta)$, both quantities are $O(\eta)$.

Proof. Banach iteration from $(0, 0)$. First iterate: $D_{\text{eff}}^{(0)} = L_F \eta \sqrt{r}/\Delta_{\text{gap}} = O(\eta)$ (uses A′), $\tilde{\gamma}_{\text{eff}}^{(0)} = \tilde{\gamma}_G$, $\varepsilon'^{(0)} = O(\eta)$, $R^{(1)} \le \tau + O(\eta)$ (using B″ to bound the tail floor). Each subsequent iterate perturbs by $O(\eta^2)$. The map is Lipschitz with constant $O(\eta)$ near the fixed point, hence contractive.

What this proves precisely. Under A, A′, B′, B″, C′, the coupled system has a fixed point with $R_\infty, \varepsilon'_\infty$ as stated, on any finite horizon over which the assumptions hold. Assumption D is *derived* in this regime, not posited. Outside this regime — in particular when B″ fails — the drift is not eliminated; it is re-expressed in terms of the tail τ, which can be large. □

G Amortized Contraction (Full Proof)

Theorem G.1 (Amortized contraction)**.** *Let $\rho_t = \tilde{\gamma}_t(1 + \kappa_{G,t})$. Suppose there exist constants $\bar{\rho} \in (0, 1)$ and $C_\rho \ge 1$ such that the* suffix-product condition *holds: for every $0 \le s < t$,*

$$\prod_{i=s}^{t-1} \rho_i \;\le\; C_\rho\, \bar{\rho}^{\,t-s}.$$

Then the conclusion of Theorem 5.1 holds with ρ' replaced by $\bar{\rho}$ and an additional multiplicative constant C_ρ in the tracking-error and residual-coupling contributions.

Proof. The tracking recursion with variable ρ_t is $\varepsilon_t \le \rho_{t-1}\varepsilon_{t-1} + c_{t-1}$. Unrolling,

$$\varepsilon_T \;\le\; \Big(\prod_{i=0}^{T-1} \rho_i\Big)\varepsilon_0 \;+\; \sum_{s=0}^{T-1} c_s \Big(\prod_{i=s+1}^{T-1} \rho_i\Big).$$

Apply the suffix-product condition twice. *Initial term:* $\prod_{i=0}^{T-1} \rho_i \le C_\rho\, \bar{\rho}^T \to 0$. *Sum term:* each partial product $\prod_{i=s+1}^{T-1} \rho_i \le C_\rho\, \bar{\rho}^{T-1-s}$, so $\sum_s c_s \prod_{i=s+1}^{T-1} \rho_i \le C_\rho\, \frac{\max_s c_s}{1-\bar{\rho}}$. Both terms are uniformly controlled, and the rest of the proof of Theorem 5.1 (Appendix H) goes through with $1/(1-\rho')$ replaced by $C_\rho/(1-\bar{\rho})$.

Why the suffix-product condition (not the time-averaged $\bar{\rho} = \exp((1/T)\sum_t \log \rho_t)$) is required. A condition on the *full-horizon* geometric mean does not control *suffix* products: a sequence with $\rho_t < 1$ for $t < T/2$ and $\rho_t > 1$ for $t \ge T/2$ can have full-horizon mean < 1 while every suffix product over $[T/2, T)$ grows. The earlier statement "$\bar{\rho} = \exp((1/T)\sum_t \log \rho_t) < 1$ implies each partial product is bounded by a power of $\bar{\rho}$ by AM-GM" is therefore not valid as written. The strong suffix-product condition above is exactly the body's Section 5.2 statement and is the natural Lyapunov-style stability condition for subspace tracking. □

H Complete Proof of Theorem 5.1

Proof. **Step 1 (Telescoping).** Sum the Orth-Dion descent bound over $t = 0, \dots, T-1$:

$$\sum_t \eta \|G_t\|_{\text{KF},r} \le (f_0 - f_\infty) + \sum_t \eta(\delta_t + 2\|R_t\|_{\text{KF},r}) + \tfrac{L_r}{2} T\eta^2.$$

Step 2 (Error terms, β=1). By the self-consistent analysis: $\varepsilon'_\infty = O(\eta)$. Defect: $\delta_t \le C'_\delta G_F \varepsilon_t$. Residual: $\|R_{t+1}\|_{\text{KF},r} \le C_R G_F \varepsilon_t$. Error sum splits: $\sum_t \varepsilon_t \le \varepsilon_0/(1-\rho') + T\varepsilon'_\infty = O(1) + O(T\eta)$. Therefore $\sum_t \eta(\delta_t + 2\|R_t\|_{\text{KF},r}) = O(T\eta^2)$.

Step 3 (Assembly). $\eta T \min_t \|G_t\|_{\text{KF},r} \le (f_0 - f_\infty) + (\frac{L_r}{2} + C)T\eta^2$. Dividing: $\min_t \|G_t\|_{\text{KF},r} \le \frac{f_0 - f_\infty}{\eta T} + \frac{L_r \eta}{2} + O(1/T)$.

Step 4 (Optimize $\eta = c/\sqrt{T}$). $\min_c [\frac{f_0 - f_\infty}{c} + \frac{L_r c}{2}]$ at $c^* = \sqrt{2(f_0 - f_\infty)/L_r}$. Substituting: $\min_t \|G_t\|_{\text{KF},r} \le \sqrt{2(f_0 - f_\infty)L_r/T} + O(1/T)$. □

I Convergence Rate Comparison

The exact-Muon column in Table 3 should be read as a geometry-level comparison in the rank-r Ky Fan setting. It is complementary to recent full-rank Muon analyses, which prove convergence for practical Muon variants with Nesterov momentum and weight decay and derive critical batch sizes (Sato et al., 2025). Our focus is instead the low-rank spectral update geometry and the rank-dependent dual-norm inflation caused by column normalization.

Table 3: Complete rate comparison across algorithm variants.

Property	**Stripped Dion**	**Orth-Dion**	**Exact Muon**
ν_t	$\leq \sqrt{r}$	$= 1$	$= 1$
δ_t	Signed	≥ 0	$= 0$
Smoothness/step	$\frac{L_r r}{2}\eta^2$	$\frac{L_r}{2}\eta^2$	$\frac{L_r}{2}\eta^2$
Residual coupling	$(1+\sqrt{r})\lvert R\rvert$	$2\lvert R\rvert$	0
Rate	$O(\sqrt{L_r r/T})$	$O(\sqrt{L_r/T})$	$O(\sqrt{L_r/T})$
Iterations to ε	$O(L_r r/\varepsilon^2)$	$O(L_r/\varepsilon^2)$	$O(L_r/\varepsilon^2)$
Per-step cost	$O(mnr)$	$O(mnr)+O(nr^2)$	$O(mn\min\{m,n\})$

J Contraction Condition Comparison

The convergence proof requires the per-step buffer-tracking contraction rate $\tilde{\gamma}$ to lie strictly below an admissibility ceiling that depends on how tightly the residual amplification can be controlled. The original Dion ceiling scales as $1/(1+2\sqrt{2r}\kappa)$, so it shrinks as the rank grows. Our analysis replaces this with the rank-independent ceiling $1/(1+\kappa)$, which depends only on the conditioning of the gradient covariance. Table 4 reports both ceilings at the representative regime r=64, κ=10 used throughout our LLM experiments: the original bound forces $\tilde{\gamma} < 0.0044$ while ours permits $\tilde{\gamma} < 0.091$, a 22.6× relaxation.

Table 4: Contraction condition at r=64, κ=10. Our Ky Fan dual-norm analysis (Section 5) relaxes the original Dion bound by 22.6×.

Framework	**Condition**	r=64	**Relaxation**
Original Dion	$\tilde{\gamma} < 1/(1+2\sqrt{2r}\kappa)$	$\tilde{\gamma} < 0.0044$	—
Ours	$\tilde{\gamma} < 1/(1+\kappa)$	$\tilde{\gamma} < 0.091$	22.6×

K $\beta < 1$ Regimes: Additional Details

This appendix supplements Section 6 with the full three-regime taxonomy of error-feedback dynamics, the complete β-sweep used to validate Regime 2 on GPT-2 45M, and the implicit-momentum derivation behind the prediction $\beta^* \approx 1 - 2\hat{\varepsilon}$. With ν_t fixed by Orth-Dion, the residual recursion $R_{t+1} = \beta(I - U_t U_t^\top)(G_t + R_t)$ is driven entirely by the cross-step correlation of the out-of-subspace gradient, summarised through the persistence $\phi_t := \cos\angle((I-P_t)G_t, (I-P_{t-1})G_{t-1})$. The sign of ϕ_t determines which of three qualitatively distinct dynamical regimes the optimizer is operating in, and therefore which value of β is optimal.

Table 5 summarises the three regimes. In Regime 1 ($\phi_t > 0$), gradients are coherent across steps, residuals accumulate constructively, and the tracked subspace eventually rotates to absorb them; full retention ($\beta = 1$) is optimal because it preserves the coherent signal. In Regime 2 ($\phi_t \approx 0$, the typical setting for stochastic LLM pretraining), residual increments are nearly independent, so R_T executes a $\sqrt{T}$ random walk that, when projected back onto the tracked subspace, behaves like Polyak momentum with effective coefficient $\hat{\varepsilon}^2$; reducing β amplifies this implicit-momentum signal at zero additional compute. In Regime 3 ($\phi_t < 0$), residuals anti-correlate across steps and the EF accumulator chases a moving target, so vanishing β is required to suppress harmful interference.

Table 5: Three regimes of error feedback dynamics, governed by the gradient persistence ϕ_t.

Regime	ϕ_t	**R-norm**	**EF effect**	**Optimal** β
1. Coherent	> 0	accumulates	strongly beneficial	$\beta = 1$
2. Stochastic	≈ 0	$\sqrt{T}$ random walk	implicit momentum	$\beta \approx 0.3$
3. Anti-correlated	< 0	diverges	harmful	$\beta \to 0$

Table 6 reports the complete nine-point β-sweep on GPT-2 45M (WikiText-103, r=64, 30k steps), of which the body Table 2 is a six-point subset. All runs share identical optimizer hyperparameters apart from β, and the empirical persistence $\phi_t \approx -0.01$ places the configuration squarely in Regime 2. Validation loss decreases monotonically as β shrinks from 1.0, with the empirical optimum at β≈0.1 (val. loss 2.978 vs. 3.293 at β=1); the R-norm ratio $\mathcal{R}_t^*$ grows monotonically from 0.3 at β=1 to 4.6 at β=0.05, directly tracking the predicted implicit-momentum accumulation. Disabling EF entirely (β=0) is far worse than any β>0 in the sweep, indicating that the gain from β<1 comes from tuning the implicit-momentum coefficient rather than from removing the residual altogether.

Table 6: Full β-sweep on GPT-2 45M/WikiText-103 (r=64, 30k steps; Regime 2, ϕ_t≈0). Validation loss is monotonically non-increasing in β from β=1 down to β≈0.05–0.1; below this the implicit-momentum mechanism saturates (Section 6). Disabling EF entirely (β=0) is substantially worse than any β>0, confirming EF is essential even when heavily damped. Body Table 2 reports the subset {0, 0.05, 0.1, 0.3, 0.5, 1.0}.

β	**Val loss**	Δ **vs.** β=1	$\mathcal{R}_t^*$
0.05	2.982	−0.311	4.6
0.1	**2.978**	−0.315	3.2
0.2	3.019	−0.274	2.2
0.3	3.056	−0.237	1.7
0.47	3.076	−0.217	1.3
0.5	3.111	−0.183	1.2
0.7	3.177	−0.116	0.9
1.0 (baseline)	3.293	—	0.3
0 (no EF)	3.554	+0.261	—

K.1 Key Quantities

Tracking error: $\hat{\varepsilon}_t := \|(I - U_t U_t^\top) M_t\|_F / \|M_t\|_F$. For ColNorm: $\hat{\varepsilon}_t \approx 0.35$ in LLM training. For Orth-Dion: $\hat{\varepsilon}_t \approx 0.007$.

Gradient subspace persistence: $\phi_t := \cos \angle((I-P_t)G_t,\ (I-P_{t-1})G_{t-1})$ (distinct from the per-step contraction ρ_t of Section 5.2). $\phi_t > 0$: directions coherent. $\phi_t \approx 0$: approximately independent (LLM typical: $\phi_t \approx -0.01$). $\phi_t < 0$: anti-aligned.

R-norm ratio: $\mathcal{R}_t := \|R_t\|_F / \|G_t\|_F$. Empirically $\mathcal{R}_t \approx 0$ for β=1; grows monotonically as β decreases. Optimal sweet spot: $\mathcal{R}_t^* \in [1.5, 2.0]$.

K.2 Implicit Momentum Mechanism (Regime 2)

When $\phi_t \approx 0$, the residual R_T is a sum of approximately independent terms: $R_T \approx \sum_{t=0}^{T-1} (I-P_t)G_t$. By random-walk scaling: $\|R_T\|_F \sim \sqrt{T} \cdot \hat{\varepsilon} \cdot \|G\|_F$.

The projected residual $P_T R_T$ acts as a momentum term: $\mathbb{E}[P_T R_T] \approx \hat{\varepsilon}^2 \cdot \sum_{t<T} \gamma^{T-t} \nabla f(w_t)$ for effective decay $\gamma \in (0, 1)$ depending on subspace rotation rate. This is equivalent to heavy-ball momentum with coefficient $\hat{\varepsilon}^2$.

For ColNorm ($\hat{\varepsilon} \approx 0.35$): implicit momentum ≈ 0.12—substantial. For Orth-Dion ($\hat{\varepsilon} \approx 0.007$): implicit momentum $\approx 5 \times 10^{-5}$—negligible. *This is why ColNorm+EF empirically outperforms Orth-Dion+EF even when* $\phi_t \approx 0$.

K.3 Experimental Evidence: EF $\neq$ Explicit Momentum

PolyakDion (explicit Polyak momentum μ=0.95, $R\equiv 0$ by design) achieves val loss 3.162. Dion with β=0.3 achieves 3.056—a gap of 0.106 nats. The mechanism: EF accumulation reshapes $M_t = G_t + R_t$ *before* the rank-r compression step, directly influencing which subspace is tracked. Explicit momentum is applied before compression and does not interact with subspace selection in the same way.

L LLM Pre-training Details

Our code will be available at https://anonymous.4open.science/r/orth_dion-4DD4/README.md.

L.1 320M Pre-training

Table 7: Llama 3 320M architecture.

Hyperparameter	Value
Hidden dimension	768
Intermediate dimension (FFN)	2048
Number of layers	18
Number of attention heads	12
Number of KV heads	12
Sequence length	2048
Vocabulary size	128,256
Total parameters	$\sim$320M

Table 8: Training hyperparameters (320M).

Hyperparameter	Value
Token budget	3.2B
Warmup steps	2000
LR schedule	cosine decay
Gradient clipping	1.0
Batch size	256
GPUs	8 $\times$ GH200 (120GB)
FSDP	Full Shard, bf16

In this paper, we used torchtitan stack to train a Llama 3 architecture model, using C4 data, on the Miyabi-G supercomputer at JCAHPC, with each run using 8 GH200 (120GB) GPUs distributed across 8 nodes interconnected via InfiniBand NDR200. The number of tokens (3.2B) for the training was calculated based on the Chinchilla optimal heuristic. For Dion, we used the microsoft implementation Ahn et al. (2025b), and our proposed Dion variants extended that implementation. For AdamW, we swept the joint grid of learning rate $\{0.004, 0.008, 0.012, 0.016\}$ and β_1 $\{0.9, 0.925, 0.95, 0.975\}$. For Muon and Dion, we swept the joint grid of learning rate $\{0.004, 0.008, 0.012, 0.016\}$ and output head LR scaling $\{\text{True}, \text{False}\}$, where the latter controls whether a shape-dependent rescaling is applied to the learning rate of the final output projection (whose fan-out, the vocabulary size, is much larger than that of the hidden matrices). We reused Dion's best-performing configuration for our proposed methods without further tuning. For Ada-Orth-Dion, we set the maximum rank at $r_{max} = 384$ and initialized the rank at r_{max} to conduct comparison with Dion at $r = 384$. Table 7, Table 8, and Table 9 list the model architecture, training hyperparameters, and optimizer hyperparameters used for the experiments, respectively. A single 320M run (6,100 steps, 3.2B tokens) takes 58–76 minutes on 8 $\times$ GH200 depending on the optimizer, with peak VRAM of $\sim$40 GB/GPU under FSDP full-shard.

Table 9: Optimizer hyperparameters (320M). Values marked with † were tuned via sweep. For Muon, Dion, and Ada-Orth-Dion, the scalar-parameter group (e.g., embeddings, biases, layer norms) is optimized with AdamW-style updates whose hyperparameters are reused from the best AdamW configuration.

Hyperparameter	**AdamW**	**Muon**	**Dion**	**Ada-Orth-Dion**
Learning rate†	0.012	0.016	0.012	0.012
Output head LR scaling†	—	False	False	False
β_1	0.95†	0.9	0.9	0.9
β_2	0.95	0.95	0.95	0.95
ϵ	1e−8	1e−8	1e−8	1e−8
Weight decay	0.1	0.1	0.1	0.1
Momentum factor	—	0.95	—	—
γ	—	—	—	1.425
Scalar LR	—	0.016	0.012	0.012
Scalar β_1	—	0.95	0.95	0.95
Scalar β_2	—	0.95	0.95	0.95
Scalar ϵ	—	1e−8	1e−8	1e−8
Scalar weight decay	—	0.1	0.1	0.1

Table 10: `w16k_l4` architecture.

Hyperparameter	**Value**
Hidden dimension	16,384
Intermediate dimension (FFN)	43,776
Number of layers	4
Number of attention heads	64
Number of KV heads	64
Head dimension	256
Sequence length	2048
Vocabulary size	128,256
Total parameters	∼17.1B

L.2 Wide-and-Shallow Large-Scale Wall-Clock Study

To examine wall-clock behavior in the bandwidth-bound regime, we additionally measured per-step time on a wide-and-shallow Llama 3 architecture (referred to as `w16k_l4`) with approximately 17.1B parameters (Figure 1b). Table 10 and Table 11 list the model architecture and training used for the wall-clock study. Each 50-step measurement run takes approximately 40 minutes on 8 × A40 (48 GB GDDR6 per GPU, ∼44.5 GB usable after framework overhead), with peak VRAM of ∼43 GB/GPU under FSDP full-shard. Aside from the rank fractions set to different values, we reused the best-performing configurations from the 320M sweep for optimizer hyperparameters. The model architecture is intentionally wide and shallow so that the per-block matrices (e.g., FFN matrices of shape $43{,}776 \times 16{,}384$) are large enough for NCCL collectives to transition from the latency-bound regime to the bandwidth-bound regime. The 8 × A40 hardware was chosen for the same reason: A40s have substantially lower inter-GPU bandwidth than current-generation training accelerators (e.g., H100s with NVLink), so the bandwidth-bound regime dominates at smaller collective payload sizes than it would on a higher-end system, making the wall-clock differences between communication-light and communication-heavy optimizers more visible. In this regime, the reduced communication volume of low-rank Dion-family updates relative to Muon's full-rank Newton–Schulz updates translates directly into wall-clock savings.

For Ada-Orth-Dion, we executed the full adaptive rank algorithm, including the per-block effective rank computation and bookkeeping, but overrode the resulting rank assignment so that the actual working rank was always held fixed at $\rho = 0.93 \times 0.25 = 0.2325$, i.e., 93% of Dion's fixed rank fraction, regardless of the computed effective rank or the schedule's shrink/grow decision. The 93%

Table 11: Training hyperparameters for the `w16k_l4` wall-clock study.

Hyperparameter	Value
Batch size	8
Sequence length	2048
Measurement steps	50 (mean of steps 2–50)
GPUs	8 × A40 (44.45GB)
FSDP	Full Shard, bf16

factor was chosen to match the average effective rank that Ada-Orth-Dion's adaptive schedule settled at in the 320M experiment, so the wide-model setting reflects its empirically observed steady-state. Since the run still goes through the full Ada-Orth-Dion code path (sort-and-pack batching, per-block adaptive bookkeeping, Cholesky-QR), any algorithmic and implementation overhead inherent to Ada-Orth-Dion is reflected in the measurement, while holding the rank constant isolates whether the rank reduction enabled by the adaptive mechanism is sufficient to recover this overhead and yield a net wall-clock improvement over fixed-rank Orth-Dion or Dion.

L.3 Per-Step Time Across Communication Regimes

To examine how the QR overhead depends on the communication regime, we additionally measured per-step time and ran training on 8×GH200 interconnected via InfiniBand NDR200. Together with the A40 setting of Appendix L.2, the measured settings span the communication axis, from very large matrices over slow links down to small matrices over fast links. QR's per-step overhead grows as communication stops hiding it (Table 12, all at matched rank).

Table 12: Per-step time of Dion and Orth-Dion at matched rank, from most to least communication-bound (mean ± 2 std over per-step times within each run). *QR overhead*: Orth-Dion's per-step time relative to Dion's. *Time to Dion's best loss*: Orth-Dion's wall-clock to reach Dion's best validation loss, relative to Dion's.

Setting	Dion (s/step)	Orth-Dion (s/step)	QR overhead	Time to Dion's best loss
17.1B, 8×A40 (Figure 1b)	48.57 ± 1.30	50.37 ± 1.68	+3.7%	—
17.1B, 8×GH200	5.46 ± 0.01	5.85 ± 0.01	+7.0%	—
1.5B, 8×GH200 (r_f=0.5)	1.66 ± 0.06	1.82 ± 0.06	+9.6%	1.02× (1 seed)
320M, 8×GH200 (Table 1 runs)	0.728 ± 0.006	0.748 ± 0.007	+2.8%	0.88–0.89×

Two factors move against each other down the table. Communication's share of step time shrinks, which exposes QR's compute, and QR's $O(nr^2)$ cost shrinks relative to the $O(mnr)$ power step as the matrices and rank get small. The first effect drives the rise across the large-model rows; the second is why the 320M row falls back.

L.4 Convergence at Larger Scale: Llama 3 1.5B

Table 13: Llama 3 1.5B architecture.

Hyperparameter	Value
Hidden dimension	1,792
Intermediate dimension (FFN)	4,864
Number of layers	26
Number of attention heads	16
Number of KV heads	16
Sequence length	2,048
Vocabulary size	128,256
Total parameters	∼1.5B

Table 14: Training hyperparameters (1.5B).

Hyperparameter	Value
Token budget	15B
Rank r (rank fraction)	896 (0.5)
Warmup steps	2,861
LR schedule	cosine decay
Gradient clipping	1.0
Batch size	256 seqs (∼0.5M tokens)
GPUs	8 × GH200 (120GB)
FSDP	Full Shard, bf16
Seeds	1

To test whether Orth-Dion's convergence advantage carries over to larger models, we train a 1.5B model to convergence, $4.7\times$ the size of the 320M model.

Setup. Tables 13 and 14 list the model architecture and training hyperparameters. Dion and Orth-Dion use the same hyperparameters, taken from the best 320M configuration. Due to compute cost, we train each optimizer once (a single seed), rather than three times as at 320M.

Results. Orth-Dion reaches Dion's best validation loss (2.8225) in 5.3% fewer steps. At the end of training, Orth-Dion's loss is lower (2.817 vs. 2.823): Dion's loss stops improving over its last $\sim$1,500 steps, while Orth-Dion's keeps decreasing. Dion's measured $\bar{\nu}_t$ is 2.01 at this rank, higher than at 320M, consistent with the inflation growing with rank. The result suggests that convergence advantage carries over to larger models.

In wall-clock time, Orth-Dion reaches Dion's best loss at $1.02\times$ Dion's time, because the 9.6% per-step QR overhead in this setting (Appendix L.3) cancels the 5.3% step saving. So at this scale, fixed-rank Orth-Dion reaches Dion's loss in about the same time, and its benefit is the lower final loss.