AdaSC-Muon: Adaptive Subspace Compression for Communication-Efficient Muon Training

Anonymous authors
Paper under double-blind review

Abstract

Muon has emerged as a promising optimizer for large-scale neural network training, but communication and orthogonalization costs can limit its efficiency in distributed settings. EF21-Muon by Gruntkowska et al. reduces communication by compressing momentum differences. Its best-performing compressor still requires workers to transmit local singular bases at every iteration, so the worker-to-server communication cost scales with the original matrix dimensions. We introduce adaptive subspace-compressed Muon (AdaSC-Muon), which uses shared subspaces for both momentum compression and Muon computation. The key insight is that aggregating projected momentum before orthogonalization enables a global update to be computed directly in a reduced matrix space. Workers reuse the shared bases and quantize projected momentum differences with error feedback. Periodic full-momentum corrections update the subspaces as momentum changes during training. We establish convergence guarantees and communication complexity bounds for AdaSC-Muon. Language model pretraining experiments with NanoGPT on FineWeb show that, compared with EF21-Muon, AdaSC-Muon and its diagonal variant reduce worker-to-server communication cost by up to 47.8% and reach the target validation loss in up to 53.8% less wall-clock time.

1 Introduction

Muon has demonstrated strong performance in language model pretraining by exploiting the matrix geometry of hidden-layer weights (Jordan et al., 2024; Liu et al., 2025; Shah et al., 2025). Its momentum orthogonalization corresponds to a linear minimization oracle (LMO) over the spectral-norm ball (Bernstein & Newhouse, 2024; Pethick et al., 2025; Riabinin et al., 2025). In distributed training, momentum aggregation and matrix orthogonalization introduce communication and computation costs that can limit these optimization gains. EF21-Muon reduces communication by compressing momentum differences with error feedback, then applies the LMO to the aggregated momentum approximations (Gruntkowska et al., 2026). Its largest reported communication reduction uses RankK and Natural compression (Gruntkowska et al., 2026; Horvóth et al., 2022). Each worker computes and transmits new singular factors at every iteration, so communication still scales with the original matrix dimensions. Accumulated corrections can also produce a full-rank momentum approximation, leaving the server to orthogonalize a matrix of the original size.

Subspace methods use low-rank projections to reduce communication, optimizer memory, and matrix computation. PowerSGD aggregates gradient factors obtained through shared projections (Vogels et al., 2019), while GaLore maintains optimizer states for projected gradients (Zhao et al., 2024). Dion constructs low-rank updates by aggregating products of momentum and projection matrices. Lazy-Dion keeps its right projection matrix fixed between periodic updates, avoiding intermediate QR factorizations (Ahn et al., 2025b). TSR-Adam projects gradients using shared left and right rank-r bases and maintains $r \times r$ first and second moments for AdamW updates (Dang et al., 2026). FedSubMuon also uses shared left and right bases, but performs local Muon steps before averaging the updated coefficient matrices (Chen et al., 2026b). Because orthogonalization is nonlinear, this local-update protocol generally differs from applying a spectral LMO to globally aggregated momentum.

These observations motivate compressing momentum in shared subspaces where aggregation and orthogonalization can both be performed on smaller matrices. The central challenge is to maintain an

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.

accurate approximation of projected global momentum while updating the shared bases to capture changes in its dominant singular directions.

Our approach. To address this challenge, we update the momentum approximation at every iteration while refreshing the shared bases periodically. Projected error feedback tracks momentum between basis updates, while full-momentum corrections provide the global approximation used to select new singular bases. We propose Adaptive Subspace-Compressed Muon (AdaSC-Muon), which uses shared left and right bases of rank r to track projected momentum in an $r \times r$ matrix. Workers quantize the differences between projected momentum and their stored approximations. The server updates its projected momentum approximation with the averaged corrections before applying the LMO. Basis reuse avoids repeated transmission of local singular factors, while the spectral update is computed within the reduced matrix space. Subspace adaptation requires information beyond the current projection. AdaSC-Muon therefore complements projected error feedback with periodic full-momentum corrections that determine new singular bases. Projecting the corrected full-momentum approximations onto these bases initializes the subsequent projected tracking. Our analysis accounts for the inherited compression error and quantifies how subspace rank and the basis update period affect convergence and communication efficiency. Our contributions are summarized as follows.

- We propose AdaSC-Muon, which combines shared subspace compression with a spectral LMO computed from aggregated projected momentum. Projected error feedback supports communication between basis updates, while periodic full-momentum corrections determine new bases and initialize the projected approximations.
- For the full projected option with exact initialization and exact LMOs, we establish an $\mathcal{O}(K^{-1/4})$ stochastic convergence rate in the average expected nuclear norm of the gradient under the stated assumptions and parameter choices. The analysis controls compression errors, momentum variation, and the error inherited at basis updates.
- We derive communication complexity bounds that quantify the tradeoff between per-iteration communication and the convergence cost of projection. We also give a sufficient condition for a smaller asymptotic communication upper bound than normalized EF21-Muon at the same gradient accuracy.
- NanoGPT-124M pretraining on FineWeb demonstrates that AdaSC-Muon achieves comparable minimum validation loss at 5B tokens with 24.6% lower worker-to-server communication cost than EF21-Muon with RankK 15% and Natural compression. It reaches the target validation loss of 3.350 in 53.8% less wall-clock time.
- We evaluate diagonal compression of AdaSC-Muon to examine the tradeoff between communication reduction and optimization progress. Although it further reduces communication, it requires more training steps and wall-clock time than AdaSC-Muon to reach the same target validation loss.

2 Related Work and Preliminaries

2.1 Related Work

Matrix optimization and Muon. Adam and AdamW use elementwise gradient scaling (Kingma & Ba, 2014; Loshchilov & Hutter, 2017), while Shampoo and SOAP use matrix preconditioning (Gupta et al., 2018; Vyas et al., 2025). Muon orthogonalizes momentum (Jordan et al., 2024), with a spectral-norm interpretation connected to steepest descent and non-Euclidean trust-region methods (Bernstein & Newhouse, 2024; Kovalev, 2025). Scion and Gluon extend the LMO viewpoint to different layerwise norms (Pethick et al., 2025; Riabinin et al., 2025). Recent studies analyze convergence (Li & Hong, 2025), evaluate language model pretraining (Liu et al., 2025; Shah et al., 2025) and improve approximate orthogonalization (Amsel et al., 2026; Shulgin et al., 2025).

Compression with error feedback. Quantization reduces the precision of transmitted gradients (Seide et al., 2014; Alistarh et al., 2017; Horvóth et al., 2022), while sparsification sends selected components (Lin et al., 2017; Stich et al., 2018). Error feedback retains omitted information for later updates and supports convergence with biased compressors (Karimireddy et al., 2019; Beznosikov et al., 2023). EF21 compresses differences from stored gradient approximations (Richtárik et al., 2021), with extensions to momentum, bidirectional compression, and normalized updates (Fatkhullin et al., 2023; Gruntkowska et al., 2023; Khirirat et al., 2026). For Muon, the position of compression

relative to the LMO matters. EF21-Muon tracks momentum through compressed differences and aggregates the approximations before the global LMO (Gruntkowska et al., 2026). SignMuon signs locally orthogonalized momentum directions (Mishra et al., 2026), and error feedback applied to Muon's output can still fail to ensure convergence (Smirnova & Kravatskiy, 2026). AdaSC-Muon retains the aggregation order of EF21-Muon but compresses projected momentum differences between periodic full-momentum corrections.

Low-rank compression and subspace optimization. Shared projections support low-rank gradient aggregation in PowerSGD (Vogels et al., 2019) and random projection in SEPARATE (Zhao et al., 2025). GreedyLore combines error feedback with periodic factorization and adaptive column selection to maintain contractive compression (Chen et al., 2026a). For optimizer memory, GaLore stores states in gradient subspaces (Zhao et al., 2024), and LDAdam transfers states when the bases change (Robert et al., 2025). GoLore uses random subspaces to obtain convergence guarantees (He et al., 2024), while GUM corrects projection bias by randomly selecting layers for complementary updates (Pan et al., 2025). TSR-Adam aggregates gradients projected onto shared left and right bases and uses an AdamW update with moment states in the resulting $r \times r$ space. It refreshes these bases using random sketches (Dang et al., 2026). AdaSC-Muon uses a spectral LMO with error feedback on both projected and full momentum.

Distributed Muon. The distributed Muon implementation used for Moonlight partitions optimizer states with ZeRO-1 (Liu et al., 2025; Rajbhandari et al., 2020). MuonBP uses blockwise orthogonalization on matrix shards between periodic full-matrix orthogonalizations (Khaled et al., 2026). DiLoCo performs multiple local steps between synchronizations (Douillard et al., 2023), with MuLoCo using Muon as the local optimizer (Thérien et al., 2025). FedMuon corrects local LMO inputs with control variates inspired by SCAFFOLD (Takezawa et al., 2026; Karimireddy et al., 2020). In decentralized training, DeMuon uses gradient tracking (He et al., 2025), while SUDA-Muon unifies gradient tracking and other primal-dual communication schemes (Zhang et al., 2026).

Low-rank Muon updates. Dion computes low-rank updates from aggregated momentum products, and Lazy-Dion reuses a right basis between periodic updates (Ahn et al., 2025b). Trion selects directions from a fixed discrete cosine transform and orthogonalizes a lower-dimensional momentum representation (Modoranu et al., 2025). Dion2 selects momentum rows or columns (Ahn et al., 2025a), while Orth-Dion replaces Dion's right-factor column normalization with QR orthogonalization (Nakamori et al., 2026). FedSubMuon also uses shared left and right bases and small coefficient matrices. It performs local Muon steps on these coefficients before averaging them, and its gradient-tracking extension refreshes the bases using aggregated projected probe gradients (Chen et al., 2026b). AdaSC-Muon instead averages compressed momentum corrections to update a projected approximation before applying the server LMO. Its bases are selected from a periodically corrected global momentum matrix, and its projected approximations are reinitialized from the corrected full-momentum states. The analysis accounts for the error inherited at these basis changes, subsequent projected compression errors, and momentum changes between basis updates.

2.2 Preliminaries

Problem setup. We study synchronous optimization with N workers and a common parameter matrix $X \in \mathbb{R}^{m\times n}$, where $1 \le n \le m$. The objective is $\min_X f(X) = \frac{1}{N}\sum_{j=1}^{N} f_j(X)$, with $f_j(X) = \mathbb{E}_{\xi\sim\mathcal{D}_j} F_j(X;\xi)$. Here j indexes worker, F_j is the sample loss, and $\mathcal{D}_j$ is the local data distribution. These distributions may differ. Let k denote the K model updates, with $k = 0, \dots, K-1$. At iteration $k \ge 1$, worker j draws a fresh sample ξ_j^k, independently of past randomness and the other workers, and evaluates $\nabla f_j(X^k;\xi_j^k) = \nabla_X F_j(X^k;\xi_j^k)$.

Matrix notation. For matrices of the same dimensions, $\langle A, B\rangle = \mathrm{tr}(A^\top B)$. Write the singular values of A in decreasing order as $\sigma_i(A)$. Its spectral, nuclear, and Frobenius norms are $\|A\|_{\mathrm{op}} = \sigma_1(A)$, $\|A\|_* = \sum_i \sigma_i(A)$, and $\|A\|_{\mathrm{F}} = (\sum_i \sigma_i(A)^2)^{1/2}$. For $A \in \mathbb{R}^{m\times n}$ and an integer rank $1 \le r \le n$, $\mathrm{SVD}_r(A)$ returns paired leading singular bases $Q_L \in \mathbb{R}^{m\times r}$ and $Q_R \in \mathbb{R}^{n\times r}$ with $Q_L^\top Q_L = Q_R^\top Q_R = I_r$, where I_r is the identity matrix. If the rank of A is below r, the remaining columns are chosen from the corresponding nullspaces. For the compact SVD $A = U\Sigma V^\top$ over positive singular values, define $\mathrm{polar}(A) = UV^\top$ and $\mathrm{polar}(0) = 0$. The operators $\mathrm{diag}(A)$ and $\mathrm{Diag}(v)$ respectively extract a diagonal vector and form a diagonal matrix from a vector v.

Muon. Given a stochastic gradient g^k at X^k, Muon forms the momentum $M^k = (1-\beta)M^{k-1} + \beta g^k$ and takes $X^{k+1} = X^k + \eta\,\mathrm{LMO}(M^k)$, where $\beta \in (0,1]$ and $\eta > 0$ is the step size. We use the minimizing spectral LMO,

$$\mathrm{LMO}(A) = -\operatorname{polar}(A) \in \operatorname*{arg\,min}_{\|D\|_{\mathrm{op}} \le 1} \langle A, D\rangle, \qquad \langle A, \mathrm{LMO}(A)\rangle = -\|A\|_*. \tag{1}$$

The constraint defines the update direction D, while the parameter matrix remains unconstrained. Newton–Schulz (NS) iteration approximates the polar factor in practice (Jordan et al., 2024).

EF21-Muon and compression notation. Let $M_j^k \in \mathbb{R}^{m\times n}$ be worker j's momentum, with average $\overline{M}^k = \frac{1}{N}\sum_j M_j^k$. EF21-Muon (Gruntkowska et al., 2026) stores local approximations G_j^k and their average G^k. With uncompressed server broadcasts, it updates

$$G_j^k = G_j^{k-1} + \mathcal{C}(M_j^k - G_j^{k-1}), \qquad G^k = \frac{1}{N}\sum_j G_j^k, \qquad X^{k+1} = X^k + \eta\,\mathrm{LMO}(G^k). \tag{2}$$

The compressor $\mathcal{C}$ may be randomized, and $\mathcal{C}(A)$ denotes its decoded output for an input matrix A. RankK and Natural compression transmit quantized singular factors and values of a rank-s approximation, where $1 \le s \le n$. Although each correction has rank at most s, the accumulated matrix G^k can have full rank. AdaSC-Muon instead uses $\mathcal{C}_{\mathrm{a}}\colon \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ for full-momentum differences compression and $\mathcal{C}_{\mathrm{b}}\colon \mathbb{R}^{r\times r} \to \mathbb{R}^{r\times r}$ for projected differences compression. The integer $T \ge 1$ denotes the number of iterations between shared-basis updates. The matrices C_j^k and C^k denote the local and averaged full-momentum approximations. The projected momentum is B_j^k, its local approximation is $\widehat{B}_j^k$, and the averaged projected approximation is $\widehat{B}^k$. The compressed full and projected corrections are ΔC_j^k and R_j^k, respectively. The server computes a projected LMO direction $W^k \in \mathbb{R}^{r\times r}$ and a model update direction $O^k \in \mathbb{R}^{m\times n}$. Section 3 specifies their update rules.

3 Adaptive Subspace-Compressed Muon

3.1 Algorithm Description

AdaSC-Muon uses shared subspaces for momentum compression and the global Muon update. The server selects bases $Q_L \in \mathbb{R}^{m\times r}$ and $Q_R \in \mathbb{R}^{n\times r}$ from the top-r singular vectors of its global momentum approximation. Workers reuse these bases for T iterations. This avoids transmitting new local singular factors at every iteration, as required by the RankK compressor in EF21-Muon. Between basis updates, workers apply error feedback to their projected momentum matrices. Worker j forms $B_j = Q_L^\top M_j Q_R$ and compresses $B_j - \widehat{B}_j$ using $\mathcal{C}_{\mathrm{b}}$. The worker sends the correction and adds it to its stored approximation $\widehat{B}_j$. The server averages the corrections from all workers and adds the result to its stored matrix $\widehat{B}$ as $\widehat{B} = \frac{1}{N}\sum_j \widehat{B}_j$. Because all workers use the same bases, the average of their projected momentum matrices equals the projection of the global momentum. The server applies the LMO to the aggregated projected momentum approximation $\widehat{B}^k \in \mathbb{R}^{r\times r}$ and updates the model X^{k+1}.

The dominant directions of global momentum can change during training. To select new bases, the server uses the global momentum approximation C. Every T iterations, worker j sends the compressed difference $\mathcal{C}_{\mathrm{a}}(M_j - C_j)$ and adds it to C_j, its stored full-momentum approximation. The server adds the average of these corrections to C, computes the top-r singular bases of the updated C, and broadcasts them. Workers and server project the updated C_j and C onto the new bases to initialize their projected approximations.

The full update rule is presented in Algorithm 1. We provide two compression options for AdaSC-Muon. Between basis updates, Option I compresses the full projected momentum difference $B_j^k - \widehat{B}_j^{k-1} \in \mathbb{R}^{r\times r}$ and sends R_j^k. Option II compresses only the r diagonal values of the same difference. We establish convergence for Option I. Details of Option II are provided in Appendix E.3.

Algorithm 1: AdaSC-Muon

Input: Initial matrix X^0, step sizes η, momentum parameter $\beta \in (0,1]$, rank $1 \le r \le n$, period $T \ge 1$, full-momentum compressor $\mathcal{C}_{\mathrm{a}}$, projected compressor $\mathcal{C}_{\mathrm{b}}$, iterations K

Workers $M_j^0 \leftarrow \nabla f_j(X^0)$, $C_j^0 \leftarrow M_j^0$ and send C_j^0

Server $C^0 \leftarrow \frac{1}{N}\sum_j C_j^0$, $(Q_L^0, Q_R^0) \leftarrow \mathrm{SVD}_r(C^0)$, $\widehat{B}^0 \leftarrow (Q_L^0)^\top C^0 Q_R^0$, $W^0 \leftarrow \mathrm{LMO}(\widehat{B}^0)$, $O^0 \leftarrow Q_L^0 W^0 (Q_R^0)^\top$, $X^1 \leftarrow X^0 + \eta^0 O^0$, broadcast X^1, Q_L^0, Q_R^0, and set $\widehat{B}_j^0 \leftarrow (Q_L^0)^\top C_j^0 Q_R^0$

for $k = 1, \dots, K-1$ **do**

  Workers $M_j^k \leftarrow (1-\beta) M_j^{k-1} + \beta \nabla f_j(X^k; \xi_j^k)$

  if $k \bmod T = 0$ **then**

    // Update global momentum and shared subspaces

    for *each worker j in parallel* **do**

      $\Delta C_j^k \leftarrow \mathcal{C}_{\mathrm{a}}(M_j^k - C_j^{k-1})$, $C_j^k \leftarrow C_j^{k-1} + \Delta C_j^k$

      Send ΔC_j^k to the server

    Server $C^k \leftarrow C^{k-1} + \frac{1}{N}\sum_j \Delta C_j^k$

    $(Q_L^k, Q_R^k) \leftarrow \mathrm{SVD}_r(C^k)$, $\widehat{B}^k \leftarrow (Q_L^k)^\top C^k Q_R^k$

    Broadcast Q_L^k, Q_R^k and set $\widehat{B}_j^k \leftarrow (Q_L^k)^\top C_j^k Q_R^k$ on workers

  else

    // Track momentum within the current subspaces

    $C^k \leftarrow C^{k-1}$, $(Q_L^k, Q_R^k) \leftarrow (Q_L^{k-1}, Q_R^{k-1})$

    for *each worker j in parallel* **do**

      $C_j^k \leftarrow C_j^{k-1}$, $B_j^k \leftarrow (Q_L^k)^\top M_j^k Q_R^k$

      $R_j^k \leftarrow \mathcal{C}_{\mathrm{b}}(B_j^k - \widehat{B}_j^{k-1})$, $\widehat{B}_j^k \leftarrow \widehat{B}_j^{k-1} + R_j^k$

      Send R_j^k to the server

    Server $\widehat{B}^k \leftarrow \widehat{B}^{k-1} + \frac{1}{N}\sum_j R_j^k$

  Server $W^k \leftarrow \mathrm{LMO}(\widehat{B}^k)$, $O^k \leftarrow Q_L^k W^k (Q_R^k)^\top$

  $X^{k+1} \leftarrow X^k + \eta O^k$ and broadcast X^{k+1}

3.2 Convergence Analysis

To characterize the convergence upper bound of Algorithm 1, we make the following assumptions, similar to those used in EF21-Muon (Gruntkowska et al., 2026).

Assumption 1 (Lower Boundedness)**.** *There exists $f^\star \in \mathbb{R}$ such that $f(X) \ge f^\star$ for every X.*

Assumption 2 (Smoothness)**.** *There exist finite constants $L, L_1, \dots, L_N \ge 0$ such that, for all X, Y and every worker j, $\|\nabla f(X) - \nabla f(Y)\|_* \le L\|X - Y\|_{\mathrm{op}}$ and $\|\nabla f_j(X) - \nabla f_j(Y)\|_* \le L_j\|X - Y\|_{\mathrm{op}}$.*

Assumption 3 (Stochastic Gradients)**.** *For each worker j and fixed X, the stochastic gradient satisfies $\mathbb{E}_{\xi_j}\nabla f_j(X;\xi_j) = \nabla f_j(X)$ and $\mathbb{E}_{\xi_j}\|\nabla f_j(X;\xi_j) - \nabla f_j(X)\|_{\mathrm{F}}^2 \le \sigma^2$, for some constant $\sigma \ge 0$.*

Assumption 4 (Contractive Compressors)**.** *There exist $\alpha_{\mathrm{a}}, \alpha_{\mathrm{b}} \in (0,1]$ such that $\mathbb{E}\|\mathcal{C}_{\mathrm{a}}(A) - A\|_{\mathrm{F}}^2 \le (1-\alpha_{\mathrm{a}})\|A\|_{\mathrm{F}}^2$ for every $A \in \mathbb{R}^{m\times n}$ and $\mathbb{E}\|\mathcal{C}_{\mathrm{b}}(Y) - Y\|_{\mathrm{F}}^2 \le (1-\alpha_{\mathrm{b}})\|Y\|_{\mathrm{F}}^2$ for every $Y \in \mathbb{R}^{r\times r}$.*

Theorem 1 (Convergence of AdaSC-Muon)**.** *Suppose Assumptions 1–4 hold. Let $\{X^k\}$ be generated by Algorithm 1 with step size $\eta > 0$ and momentum parameter $\beta \in (0,1]$. Then, for any $K \ge 1$,*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(X^k)\|_* \le \frac{n\Delta}{rK\eta} + \eta\mathcal{L}_\beta + a_{n,r}\frac{\sigma}{\sqrt{N}}\sqrt{\frac{\beta}{2-\beta}} + \beta\sigma\left(1 + \sqrt{\frac{\beta}{2-\beta}}\right)\left[2n\Gamma_T + \frac{a_{n,r}(T-1)}{\sqrt{N}}\right], \tag{3}$$

where $\Delta = f(X^0) - f^\star$, $a_{n,r} = n/\sqrt{r} + \sqrt{n}$, and $\mathcal{L}_\beta = nL/(2r) + (n/r+1)L(1-\beta)/\beta + a_{n,r}(T-1)L + 2n\overline{L}\Gamma_T$, with $\overline{L} = \frac{1}{N}\sum_{j=1}^N L_j$. The compression coefficient is $\Gamma_T = 2T\rho_{\mathrm{a}}/(1-\rho_{\mathrm{a}}) + \frac{\rho_{\mathrm{b}} - \rho_{\mathrm{b}}^T}{1-\rho_{\mathrm{b}}}$, where $\rho_{\mathrm{a}} = \sqrt{1-\alpha_{\mathrm{a}}}$ and $\rho_{\mathrm{b}} = \sqrt{1-\alpha_{\mathrm{b}}}$.

Corollary 1 (Convergence Rate). *Choosing the constant parameters* $\eta = \sqrt{\frac{n\Delta}{rK\mathcal{L}_\beta}}$ *and* $\beta = \min\{1, 2\big[1 + \frac{a_{n,r}\sigma}{\sqrt{2N}}\sqrt{\frac{rK}{n\Delta(\frac{n}{r}+1)L}}\big]^{-1}, 2\big[1 + \big[\frac{2(1+\sqrt{2})^2 rK\sigma^2}{n\Delta(\frac{n}{r}+1)L}\big(2n\Gamma_T + \frac{a_{n,r}(T-1)}{\sqrt{N}}\big)^2\big]^{1/3}\big]^{-1}\}$*, we have*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(X^k)\|_* \le 2\sqrt{\frac{n\Delta\mathcal{L}_1}{rK}} + 2^{5/4}\left[\frac{n\Delta}{rK}\left(\frac{n}{r}+1\right)L\frac{a_{n,r}^2\sigma^2}{N}\right]^{1/4} + 3\left[\frac{(1+\sqrt{2})n\Delta\sigma}{2rK}\left(\frac{n}{r}+1\right)L\left(2n\Gamma_T + \frac{a_{n,r}(T-1)}{\sqrt{N}}\right)\right]^{1/3}, \quad (4)$$

where $\mathcal{L}_1 = \frac{nL}{2r} + a_{n,r}(T-1)L + 2n\overline{L}\Gamma_T$*. The convergence rate is* $\mathcal{O}(K^{-1/4})$ *for* $\sigma > 0$*.*

Discussion. For fixed $r, T, \alpha_{\mathrm{a}}, \alpha_{\mathrm{b}}$, Corollary 1 gives the same $\mathcal{O}(K^{-1/4})$ stochastic convergence rate as EF21-Muon (Gruntkowska et al., 2026, Theorem 5). The bound explains the different effects of r and T. A smaller r weakens the descent guarantee. At each basis update, the selected singular directions retain at least an r/n fraction of the corrected momentum's nuclear norm, which explains the factor n/r in the bound. A larger T reuses the same bases for more iterations. The analysis must then account for momentum changes over up to $T-1$ iterations, together with the compression errors. The dependence on T and the contraction parameters appears only in the $K^{-1/2}$ and $K^{-1/3}$ terms. These terms decay faster than the leading $K^{-1/4}$ term, whose coefficient still depends on r.

Proof sketch. The main difficulty is to derive a descent bound that accounts for subspace projection, momentum changes between basis updates, and the compression errors introduced by $\mathcal{C}_{\mathrm{a}}$ and $\mathcal{C}_{\mathrm{b}}$. From smoothness, it gives $f(X^k) - f(X^{k+1}) \ge -\eta\langle\nabla f(X^k), O^k\rangle - L\eta^2/2$. Let $\tau = T\lfloor k/T\rfloor$. The shared bases are selected from C^τ, whose singular values satisfy $\sum_{i=1}^r \sigma_i(C^\tau) \ge (r/n)\|C^\tau\|_*$. Using $-\langle\widehat{B}^k, W^k\rangle = \|\widehat{B}^k\|_*$ and the triangle inequality, we obtain $-\langle\overline{M}^k, O^k\rangle \ge (r/n)\|\overline{M}^k\|_* - (\sqrt{r} + r/\sqrt{n})\|\overline{M}^k - \overline{M}^\tau\|_{\mathrm{F}} - 2r\|\overline{M}^\tau - C^\tau\|_{\mathrm{op}} - 2r\|(Q_L^k)^\top\overline{M}^k Q_R^k - \widehat{B}^k\|_{\mathrm{op}}$. The momentum recursion gives $\mathbb{E}\|\overline{M}^k - \overline{M}^\tau\|_{\mathrm{F}} \le (T-1)\big[L\eta + \frac{\beta\sigma}{\sqrt{N}}(1 + \sqrt{\beta/(2-\beta)})\big]$. It also gives $\overline{M}^k - \nabla f(X^k) = D^k + Z^k$, where D^k is the contribution of successive gradient differences and Z^k is the accumulated stochastic gradient noise. These satisfy $\|D^k\|_* \le L\eta(1-\beta)/\beta$ and $\mathbb{E}\|Z^k\|_{\mathrm{F}} \le \frac{\sigma}{\sqrt{N}}\sqrt{\beta/(2-\beta)}$. Their inner products with the update are bounded using $\|O^k\|_{\mathrm{op}} \le 1$ and $\|O^k\|_{\mathrm{F}} \le \sqrt{r}$, respectively. For compression, the basis update satisfies $\widehat{B}_j^\tau - (Q_L^\tau)^\top M_j^\tau Q_R^\tau = (Q_L^\tau)^\top(C_j^\tau - M_j^\tau)Q_R^\tau$. This identity supplies the initial error for the projected error-feedback recursion. Applying contractivity of $\mathcal{C}_{\mathrm{a}}$ and $\mathcal{C}_{\mathrm{b}}$ gives $\mathbb{E}\|\overline{M}^\tau - C^\tau\|_{\mathrm{op}} + \mathbb{E}\|(Q_L^k)^\top\overline{M}^k Q_R^k - \widehat{B}^k\|_{\mathrm{op}} \le \Gamma_T\big[\overline{L}\eta + \beta\sigma(1 + \sqrt{\beta/(2-\beta)})\big]$. Substituting these estimates into the descent inequality, summing over k, and using $f(X^K) \ge f^\star$ proves Theorem 1. Details are provided in Appendix A.

3.3 Communication Complexity

We compare the communication cost of AdaSC-Muon and EF21-Muon at the same gradient accuracy $\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(X^k)\|_* \le \varepsilon$. Following EF21-P (Gruntkowska et al., 2023, Sections 1.2–1.4) and Natural Compression (Horvóth et al., 2022, Appendix D.7), the analysis combines the communication cost per iteration with the number of iterations sufficient to reach this accuracy. We count worker-to-server communication per worker and assume that server-to-worker broadcasting is free (Gruntkowska et al., 2026). Let ℓ_{a} and ℓ_{b} denote the bits used to transmit a compressed full-momentum difference and a compressed projected momentum difference, respectively. After initialization, each AdaSC-Muon worker sends one full-momentum difference and $T-1$ projected momentum differences every T iterations. Its average communication cost is therefore $\overline{\ell}_{\mathrm{A}}(T) = [\ell_{\mathrm{a}} + (T-1)\ell_{\mathrm{b}}]/T$ bits per iteration. In comparison, each EF21-Muon worker compresses a full-momentum difference at every iteration. We denote its communication cost per iteration by ℓ_{E}. To compare the iteration bounds, we apply the same convergence analysis to both methods. With exact initialization and exact LMOs, setting $r = n$, $T = 1$, and $\mathcal{C}_{\mathrm{a}} = \mathcal{C}_{\mathrm{E}}$ in Algorithm 1 gives normalized EF21-Muon with compressor $\mathcal{C}_{\mathrm{E}}$ and update $X^{k+1} = X^k + \eta\,\mathrm{LMO}(G^k)$. Hence Theorem 1 also applies to this baseline. Let $K_{\mathrm{A}}(\varepsilon, T)$ and $K_{\mathrm{E}}(\varepsilon)$ denote the resulting iteration upper bounds after choosing the constant step size and momentum to minimize each bound. Under the conditions of Proposition 1 in Appendix C,

they satisfy $K_{\mathrm{A}}(\varepsilon, T) \sim \frac{32\frac{n}{r}(\frac{n}{r}+1)a_{n,r}^2 \Delta L \sigma^2}{N\varepsilon^4}$ and $K_{\mathrm{E}}(\varepsilon) \sim \frac{256n\Delta L\sigma^2}{N\varepsilon^4}$, where $u(\varepsilon) \sim v(\varepsilon)$ means $\lim_{\varepsilon\to 0^+} u(\varepsilon)/v(\varepsilon) = 1$.

Let $\mathcal{B}_{\mathrm{A}}(\varepsilon, T)$ and $\mathcal{B}_{\mathrm{E}}(\varepsilon)$ denote the corresponding communication upper bounds. Multiplying the communication cost per iteration by the iteration bound gives $\mathcal{B}_{\mathrm{A}}(\varepsilon, T) \sim \overline{\ell}_{\mathrm{A}}(T) K_{\mathrm{A}}(\varepsilon, T)$ and $\mathcal{B}_{\mathrm{E}}(\varepsilon) \sim \ell_{\mathrm{E}} K_{\mathrm{E}}(\varepsilon)$. The exact totals include the initial full-matrix transmission and any remaining iterations after the last complete period. These contributions do not change the limiting ratio below. Appendix C provides the derivation and the bounds for finite ε.

Proposition 1 (Comparison under a common convergence analysis)**.** *Suppose Theorem 1 applies to both methods and $\Delta, L, \sigma > 0$. Fix m, n, N, r, T, the problem constants, the encoded bit lengths, and the positive contraction parameters. Then*

$$\lim_{\varepsilon\to 0^+} \frac{\mathcal{B}_{\mathrm{A}}(\varepsilon, T)}{\mathcal{B}_{\mathrm{E}}(\varepsilon)} = \frac{\overline{\ell}_{\mathrm{A}}(T)}{\ell_{\mathrm{E}}} \frac{\frac{n}{r}\left(\frac{n}{r}+1\right) a_{n,r}^2}{8n}. \tag{5}$$

If the right-hand side is less than one, AdaSC-Muon has a smaller communication upper bound for all sufficiently small ε.

The first factor in (5) is the communication cost per iteration relative to EF21-Muon. The second is the ratio of the iteration bounds as $\varepsilon \to 0^+$. For $r < n$, this second factor is greater than one. AdaSC-Muon therefore needs a lower communication cost per iteration to obtain a smaller total communication bound. Specifically, $\overline{\ell}_{\mathrm{A}}(T)/\ell_{\mathrm{E}} < 8n/[\frac{n}{r}(\frac{n}{r}+1)a_{n,r}^2]$ is sufficient for an improvement at all sufficiently small ε. A larger T reduces the average communication cost when $\ell_{\mathrm{a}} > \ell_{\mathrm{b}}$, since workers send full-momentum differences less often. However, it can also increase the iteration bound at a fixed ε because the terms involving $\mathcal{L}_\beta$ and Γ_T grow with T. The proposition compares communication upper bounds for fixed T. Appendix C gives the encoded bit counts and numerical examples. It also compares our bound with EF21-Muon's published theorem (Gruntkowska et al., 2026, Theorem 19) at the same gradient accuracy. This comparison uses different update rules, so the difference between the bounds cannot be attributed to subspace compression alone.

4 Experiments

We compare AdaSC-Muon, its diagonal option, and EF21-Muon under fixed tokens, fixed communication cost, fixed wall-clock time, and a target validation loss. These experiments evaluate whether AdaSC-Muon can improve the validation performance of EF21-Muon while reducing communication cost and wall-clock time. We further compare full and diagonal subspace compression to examine how their communication savings affect training progress and the resources required to reach a target validation loss. The code for our experiments is available at: `https://anonymous.4open.science/r/AdaSC-Muon-E553/`.

4.1 Experimental Setup

Training configuration. We build upon the EF21-Muon codebase (Gruntkowska et al., 2026) and pretrain NanoGPT-124M on FineWeb10B (Penedo et al., 2024). All experiments are conducted on a single node with 4 NVIDIA B200 GPUs in a Distributed Data Parallel (DDP) setup. Each GPU runs one worker, and one worker aggregates the compressed updates from all workers. The model has 12 transformer layers and hidden dimension 768. We use sequences of length 1024 and a global batch size of 256 sequences, corresponding to 262,144 tokens per step. All methods use the same initialization, random seed, validation set, and batch size. Following EF21-Muon, the learning rate remains constant for the first 40% of training and then decreases linearly to 10% of its initial value. This schedule covers the full number of training steps in each run. Additional model and training details are provided in Appendix D.1.

Methods. We use RankK 15% followed by Natural compression, the EF21-Muon configuration with the largest reported communication saving. Its quantized values are transmitted in 16-bit format. AdaSC-Muon uses a projection rank of 90% of $\min\{m, n\}$ for each $m \times n$ matrix. We compare different quantization bit widths and values of T. The best configuration uses four-bit quantization for both full and projected momentum differences and $T = 3$. We use this configuration in all comparisons. Workers communicate full-momentum differences to update the shared subspaces every

T iterations. Between subspace updates, workers communicate projected momentum differences. The diagonal option uses the same settings but compresses only the diagonal of each projected momentum difference. We use the reference learning rate 3.6×10^{-4} for EF21-Muon. For AdaSC-Muon, we test learning rates in $[1.0 \times 10^{-4}, 7.5 \times 10^{-4}]$ for 500 steps. The best learning rate is 1.8×10^{-4}, which we use for both options.

Evaluation. We report the minimum validation loss for comparisons with fixed tokens, communication costs, and wall-clock time. For the target validation loss of 3.350, we compare the tokens, communication cost, and wall-clock time needed to reach it. Following EF21-Muon, we assume that server-to-worker broadcasting is free and report the cumulative worker-to-server communication cost per worker. All runs use the same random seed. Further details are provided in Appendix D.

4.2 Experimental Results

Fixed tokens. AdaSC-Muon achieves substantial reductions in communication cost and wall-clock time compared with EF21-Muon at 5B tokens (19,073 steps), as shown in Figure 1 and Table 1. Communication cost decreases from 886.7 to 668.3 GiB per worker, a 24.6% reduction, and wall-clock time decreases from 17 to 8 hours, a 52.9% reduction. AdaSC-Muon also achieves a minimum validation loss of 3.274, compared with 3.278 for EF21-Muon. These results demonstrate substantial improvements in training efficiency with comparable validation performance at the same token count. AdaSC-Muon (diagonal) further reduces communication cost to 462.7 GiB per worker, achieving a 47.8% reduction relative to EF21-Muon. However, its minimum validation loss is 3.332, which is 0.058 higher than AdaSC-Muon and 0.054 higher than EF21-Muon. The diagonal option achieves the largest communication reduction, but its validation performance is worse at the same token count. Therefore, AdaSC-Muon retains comparable validation performance to EF21-Muon while providing substantial savings in both communication cost and wall-clock time.

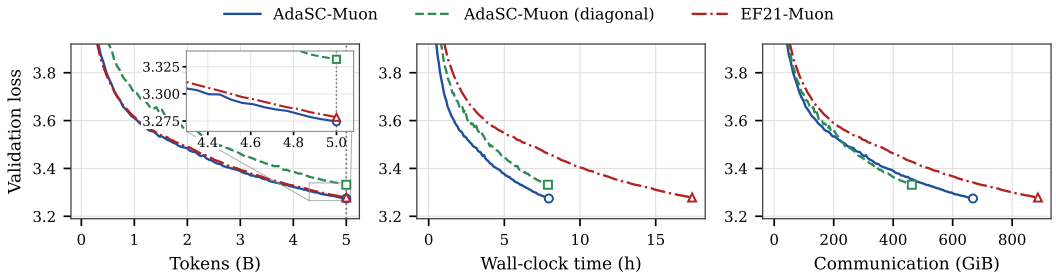

Figure 1: Comparison of AdaSC-Muon, AdaSC-Muon (diagonal), and EF21-Muon with 5B tokens. Validation loss versus tokens (left), wall-clock time (middle), and communication cost (right).

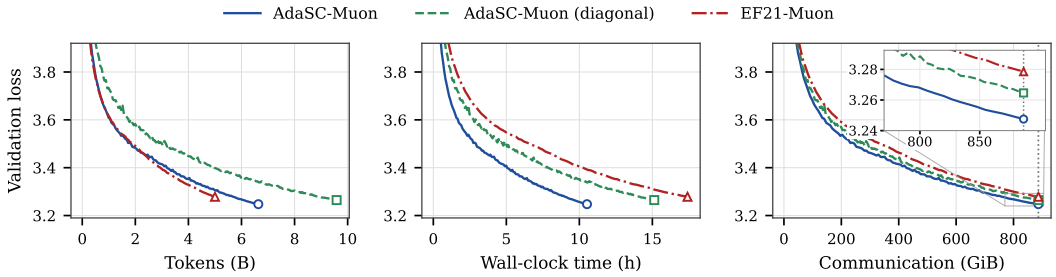

Figure 2: Comparison of AdaSC-Muon, AdaSC-Muon (diagonal), and EF21-Muon with a fixed communication cost of 886.7 GiB per worker. Validation loss versus tokens (left), wall-clock time (middle), and communication cost (right).

Fixed communication cost. We fix the cumulative worker-to-server communication cost at 886.7 GiB per worker, matching EF21-Muon after 5B training tokens. We choose the number of training steps for each AdaSC-Muon option to match this cost and compare validation performance and wall-clock

Table 1: Minimum validation loss, communication cost, and wall-clock time in the fixed-token (5B) and fixed-communication (886.7 GiB per worker) experiments.

Setting	Method	Steps	Tokens (B)	Minimum val. loss	Comm. cost (GiB)	Wall-clock time (h)
Fixed tokens (5B)	EF21-Muon	19,073	5.00	3.278	886.7	17
	AdaSC-Muon	19,073	5.00	**3.274**	668.3	**8**
	AdaSC-Muon (diagonal)	19,073	5.00	3.332	**462.7**	**8**
Fixed comm. cost (886.7 GiB)	EF21-Muon	19,073	5.00	3.278	886.7	17
	AdaSC-Muon	25,308	6.63	**3.248**	886.7	**11**
	AdaSC-Muon (diagonal)	**36,563**	**9.58**	3.265	886.7	15

time. Figure 2 and Table 1 show that AdaSC-Muon improves validation performance and substantially reduces wall-clock time compared with EF21-Muon. The number of training tokens increases from 5B for EF21-Muon to 6.63B for AdaSC-Muon, a 32.6% increase. AdaSC-Muon achieves a minimum validation loss of 3.248, compared with 3.278 for EF21-Muon, and requires 11 hours, reducing wall-clock time by 35.3%. The lower communication cost per step allows more training steps at the same total communication cost. For AdaSC-Muon (diagonal), the number of training tokens increases to 9.58B, 91.6% more than EF21-Muon. Its minimum validation loss of 3.265 improves upon EF21-Muon but remains 0.017 higher than AdaSC-Muon. The diagonal option completes 36,563 training steps, compared with 25,308 for AdaSC-Muon. This larger number of steps leads to a longer wall-clock time of 15 hours. Therefore, AdaSC-Muon achieves the lowest minimum validation loss and the shortest wall-clock time in this comparison, while the diagonal option processes the most tokens at the fixed communication cost.

Fixed wall-clock time and target validation loss. We compare different methods at 8 and 17 hours, matching the wall-clock times of AdaSC-Muon and EF21-Muon for 5B-token training, respectively. These comparisons evaluate whether EF21-Muon can achieve lower validation loss than AdaSC-Muon at the same wall-clock time. Table 2 reports these results and the target-loss comparison, with details in Appendices E.1 and E.2.

Table 2: Results at fixed wall-clock times of 8 and 17 hours and a target validation loss of 3.350.

Setting	Method	Steps	Tokens (B)	Minimum val. loss	Comm. cost (GiB)	Wall-clock time (h)
Fixed wall-clock time (8 hours)	EF21-Muon	8,679	2.28	3.376	**403.5**	8
	AdaSC-Muon	19,073	5.00	**3.274**	668.3	8
	AdaSC-Muon (diagonal)	**19,172**	**5.03**	3.330	465.1	8
Fixed wall-clock time (17 hours)	EF21-Muon	19,073	5.00	3.278	**886.7**	17
	AdaSC-Muon	41,961	11.00	**3.203**	1469.9	17
	AdaSC-Muon (diagonal)	**42,234**	**11.07**	3.246	1024.1	17
Fixed val. loss (3.350)	EF21-Muon	13,871	3.64	3.350	644.9	13
	AdaSC-Muon	**13,475**	**3.53**	3.350	472.2	**6**
	AdaSC-Muon (diagonal)	17,496	4.59	3.350	**424.4**	7

5 Conclusion

In this paper, we proposed AdaSC-Muon, a distributed Muon algorithm that uses shared subspaces to reduce communication and orthogonalization costs. By combining basis reuse with error feedback, AdaSC-Muon reduces repeated communication of local singular factors and performs orthogonalization on smaller matrices. Periodic full-momentum corrections update the subspaces as momentum changes during training. We established convergence guarantees and communication complexity bounds for AdaSC-Muon. NanoGPT-124M pretraining experiments on FineWeb show comparable validation loss at 5B tokens with lower worker-to-server communication cost and shorter wall-clock time than the evaluated EF21-Muon configuration. The diagonal variant provides further communication savings with a tradeoff in validation performance and training steps. Our findings suggest that combining shared subspace compression with error feedback is a promising direction for communication-efficient distributed Muon training.

AI USE STATEMENT

In this work, we used generative AI tools for mathematical check and code implementation assist. We have not used generative AI tools for the algorithm design and theoretical development, the problem formulation, the convergence analysis or its proofs, and synthetic data generation is not applicable to this work. Additionally, we used generative AI tools for figure and table generation, code editing, and readability editing of the manuscript. We have reviewed all AI-assisted work. The implementation was verified against an independent reference implementation of the same algorithms, every reported number is produced by the released code from stored per-seed records, and all AI-assisted text was read and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work contains no personally identifiable information and no human subjects. We see no direct ethical concerns arising from the method itself.

REPRODUCIBILITY STATEMENT

Algorithm 1 describes AdaSC-Muon. Section 4.1 and Appendix D.1 provide the dataset, model configuration, training hyperparameters, and compression settings. Appendix D.2 specifies communication cost and Appendix E.2 describes the evaluation at a target validation loss. The assumptions and theoretical results are stated in Sections 3.2 and 3.3, with proofs in Appendices A, B, and C. The experiment code is linked in Section 4.

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. *Advances in neural information processing systems*, 30, 2017.

Noah Amsel, David Persson, Christopher Musco, and Robert M Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. In *International Conference on Learning Representations*, volume 2026, pp. 138323–138360, 2026.

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

Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *Journal of Machine Learning Research*, 24(276):1–50, 2023.

Chuyan Chen, Yutong He, Pengrui Li, Weichen Jia, and Kun Yuan. Greedy low-rank gradient compression for distributed learning with convergence guarantees. *IEEE Transactions on Signal Processing*, 2026a.

Shaolong Chen, Youming Tao, Shuzhen Chen, Falko Dressler, Qingqing Ye, and Di Wang. Fedsub-muon: Communication-efficient federated llm fine-tuning via structured subspace muon. *arXiv preprint arXiv:2609.06073*, 2026b.

Sizhe Dang, Jiaqi Shao, Xiaodong Zheng, Guang Dai, Yan Song, and Haishan Ye. From $o(mn)$ to $o(r^2)$: Two-sided low-rank communication for adam in distributed training with memory efficiency, 2026. URL https://arxiv.org/abs/2602.08007.

Arthur Douillard, Qixuan Feng, Andrei A Rusu, Rachita Chhaparia, Yani Donchev, Adhiguna Kuncoro, Marc'Aurelio Ranzato, Arthur Szlam, and Jiajun Shen. Diloco: Distributed low-communication training of language models. *arXiv preprint arXiv:2311.08105*, 2023.

Ilyas Fatkhullin, Alexander Tyurin, and Peter Richtárik. Momentum provably improves error feedback! *Advances in Neural Information Processing Systems*, 36:76444–76495, 2023.

Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. Ef21-p and friends: Improved theoretical communication complexity for distributed optimization with bidirectional compression. In *International conference on machine learning*, pp. 11761–11807. PMLR, 2023.

Kaja Gruntkowska, Alexander Gaponov, Zhirayr Tovmasyan, and Peter Richtárik. Error feedback for muon and friends. In *International Conference on Learning Representations*, volume 2026, pp. 94378–94449, 2026.

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

Chuan He, Shuyi Ren, Jingwei Mao, and Erik G Larsson. Demuon: A decentralized muon for matrix optimization over graphs. *arXiv preprint arXiv:2510.01377*, 2025.

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

Samuel Horvóth, Chen-Yu Ho, Ludovit Horvath, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. In *Mathematical and Scientific Machine Learning*, pp. 129–141. PMLR, 2022.

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

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *International conference on machine learning*, pp. 3252–3261. PMLR, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

Ahmed Khaled, Kaan Ozkara, Tao Yu, Mingyi Hong, and Youngsuk Park. Muonbp: Faster muon via block-periodic orthogonalization. In *International Conference on Learning Representations*, volume 2026, pp. 17389–17416, 2026.

Sarit Khirirat, Abdurakhmon Sadiev, Artem Riabinin, Eduard Gorbunov, and Peter Richtárik. Error feedback under (l_0, l_1)-smoothness: Normalization and momentum. *Advances in Neural Information Processing Systems*, 38:162969–163009, 2026.

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

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

Jiaxiang Li and Mingyi Hong. A note on the convergence of muon. *arXiv preprint arXiv:2502.02900*, 2025.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887*, 2017.

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

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

Neel Mishra, Kushagara Trivedi, and Pawan Kumar. Signmuon: Communication-efficient distributed muon optimization. *arXiv preprint arXiv:2605.16311*, 2026.

Ionut-Vlad Modoranu, Mher Safaryan, Erik Schultheis, Max Ryabinin, Artem Chumachenko, and Dan Alistarh. Fft-based dynamic subspace selection for low-rank adaptive optimization of large language models. *arXiv preprint arXiv:2505.17967*, 2025.

Tatsuhiro Nakamori, Laura Gomezjurado Gonzalez, Ganesh Talluri, Ansh Tiwari, Hideyuki Kawashima, Ioannis Mitliagkas, Guillaume Rabusseau, and Hiroki Naganuma. Orth-dion: Eliminating geometric mismatch in distributed low-rank spectral optimization. *arXiv preprint arXiv:2605.16341*, 2026.

Rui Pan, Yang Luo, Yuxing Liu, Yang You, and Tong Zhang. Unbiased gradient low-rank projection. *arXiv preprint arXiv:2510.17802*, 2025.

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

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.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: international conference for high performance computing, networking, storage and analysis*, pp. 1–16. IEEE, 2020.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making muon & scion great again!(bridging theory and practice of lmo-based optimizers for llms). *arXiv preprint arXiv:2505.13416*, 2025.

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: 4384–4396, 2021.

Thomas Robert, Mher Safaryan, Ionut-Vlad Modoranu, and Dan Alistarh. Ldadam: Adaptive optimization from low-dimensional gradient statistics. In *International Conference on Learning Representations*, volume 2025, pp. 4778–4814, 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 *Interspeech*, volume 2014, pp. 1058–1062. Singapore, 2014.

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.

Egor Shulgin, Sultan AlRashed, Francesco Orabona, and Peter Richtárik. Beyond the ideal: Analyzing the inexact muon update. *arXiv preprint arXiv:2510.19933*, 2025.

Maria Smirnova and Alexey Kravatskiy. Sign compression for muon: Signmuon, muonsign, and the limits of error feedback. *arXiv preprint arXiv:2607.29674*, 2026.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. *Advances in neural information processing systems*, 31, 2018.

Yuki Takezawa, Anastasia Koloskova, Xiaowen Jiang, and Sebastian Stich. Fedmuon: Federated learning with bias-corrected lmo-based optimization. In *International Conference on Learning Representations*, volume 2026, pp. 39538–39569, 2026.

Benjamin Thérien, Xiaolong Huang, Aaron Defazio, Irina Rish, and Eugene Belilovsky. Muloco: Muon is a practical inner optimizer for diloco. *arXiv preprint arXiv:2505.23725*, 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, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam for language modeling. In *International Conference on Learning Representations*, volume 2025, pp. 93423–93444, 2025.

Hengrui Zhang, Boao Kong, Jiahe Geng, and Zhengyang Huang. Suda-muon: Structural design principles and boundaries for fully decentralized muon. *arXiv preprint arXiv:2604.23980*, 2026.

Hanzhen Zhao, Xingyu Xie, Cong Fang, and Zhouchen Lin. Separate: A simple low-rank projection for gradient compression in modern large-scale model training process. In *International Conference on Learning Representations*, volume 2025, pp. 24343–24368, 2025.

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

Appendix

A Proof of Theorem 1 **15**

B Proof of Corollary 1 **23**

C Communication Complexity Analysis **25**

D Experimental Details **29**

D.1 Model Configuration 29

D.2 Communication Cost 29

E Additional Experimental Results **30**

E.1 Fixed Wall-Clock Time 30

E.2 Target Validation Loss 31

E.3 Diagonal Transmission 31

A Proof of Theorem 1

Descent inequality. By Assumption 2, the update $X^{k+1} = X^k + \eta O^k$ satisfies

$$f(X^{k+1}) \le f(X^k) + \langle \nabla f(X^k), X^{k+1} - X^k \rangle + \frac{L}{2}\|X^{k+1} - X^k\|_{\mathrm{op}}^2$$
$$= f(X^k) + \eta \langle \nabla f(X^k), O^k \rangle + \frac{L\eta^2}{2}\|O^k\|_{\mathrm{op}}^2. \tag{A.1}$$

Since $(Q_L^k)^\top Q_L^k = (Q_R^k)^\top Q_R^k = I_r$,

$$\|W^k\|_{\mathrm{op}} = \|(Q_L^k)^\top O^k Q_R^k\|_{\mathrm{op}}$$
$$\le \|O^k\|_{\mathrm{op}} \le \|Q_L^k\|_{\mathrm{op}} \|W^k\|_{\mathrm{op}} \|(Q_R^k)^\top\|_{\mathrm{op}} = \|W^k\|_{\mathrm{op}}. \tag{A.2}$$

The Frobenius norm satisfies

$$\|O^k\|_{\mathrm{F}}^2 = \mathrm{tr}\big(Q_R^k (W^k)^\top (Q_L^k)^\top Q_L^k W^k (Q_R^k)^\top\big)$$
$$= \mathrm{tr}\big((W^k)^\top W^k (Q_R^k)^\top Q_R^k\big) = \mathrm{tr}\big((W^k)^\top W^k\big) = \|W^k\|_{\mathrm{F}}^2. \tag{A.3}$$

The core LMO constrains $\|W^k\|_{\mathrm{op}} \le 1$. Its singular values are therefore at most one, so

$$\|O^k\|_{\mathrm{op}} = \|W^k\|_{\mathrm{op}} \le 1, \qquad \|O^k\|_{\mathrm{F}} = \|W^k\|_{\mathrm{F}} \le \sqrt{r}\|W^k\|_{\mathrm{op}} \le \sqrt{r}. \tag{A.4}$$

It follows that

$$f(X^{k+1}) \le f(X^k) + \eta \langle \nabla f(X^k), O^k \rangle + \frac{L\eta^2}{2}. \tag{A.5}$$

We next bound the negative inner product. Adding and subtracting the global momentum,

$$-\langle \nabla f(X^k), O^k \rangle = -\langle \overline{M}^k, O^k \rangle + \langle \overline{M}^k - \nabla f(X^k), O^k \rangle. \tag{A.6}$$

Write a singular-value decomposition of the $r \times r$ core as $\widehat{B}^k = \sum_{i=1}^r s_i(\widehat{B}^k) u_i v_i^\top$. For every W with $\|W\|_{\mathrm{op}} \le 1$,

$$-\langle \widehat{B}^k, W \rangle = -\sum_{i=1}^r s_i(\widehat{B}^k) u_i^\top W v_i$$
$$\le \sum_{i=1}^r s_i(\widehat{B}^k)\|W\|_{\mathrm{op}} \le \sum_{i=1}^r s_i(\widehat{B}^k) = \|\widehat{B}^k\|_*. \tag{A.7}$$

The exact LMO satisfies

$$\langle \widehat{B}^k, W^k \rangle = \min_{\|W\|_{\mathrm{op}} \le 1} \langle \widehat{B}^k, W \rangle = -\|\widehat{B}^k\|_*. \tag{A.8}$$

Therefore,

$$-\langle \overline{M}^k, O^k \rangle = -\langle (Q_L^k)^\top \overline{M}^k Q_R^k, W^k \rangle$$
$$= \|\widehat{B}^k\|_* - \langle (Q_L^k)^\top \overline{M}^k Q_R^k - \widehat{B}^k, W^k \rangle$$
$$\ge \|\widehat{B}^k\|_* - \|(Q_L^k)^\top \overline{M}^k Q_R^k - \widehat{B}^k\|_* \|W^k\|_{\mathrm{op}}$$
$$\ge \|(Q_L^k)^\top \overline{M}^k Q_R^k\|_* - 2\|(Q_L^k)^\top \overline{M}^k Q_R^k - \widehat{B}^k\|_*$$
$$\ge \|(Q_L^k)^\top \overline{M}^k Q_R^k\|_* - 2r\|(Q_L^k)^\top \overline{M}^k Q_R^k - \widehat{B}^k\|_{\mathrm{op}}. \tag{A.9}$$

The last inequality uses $\|A\|_* \le r\|A\|_{\mathrm{op}}$ for an $r \times r$ matrix. To bound the first term, write $k = sT + \ell$, $0 \le \ell < T$. The current bases are paired top-r singular bases of C^{sT} and remain fixed throughout this interval. If $s_1 \ge \cdots \ge s_n \ge 0$ are its singular values, then

$$\|(Q_L^k)^\top C^{sT} Q_R^k\|_* = \sum_{i=1}^r s_i \ge \frac{r}{n}\sum_{i=1}^n s_i = \frac{r}{n}\|C^{sT}\|_*. \tag{A.10}$$

For $r < n$, the ordering of the singular values gives

$$\sum_{i=r+1}^{n} s_i \le (n-r)s_r \le \frac{n-r}{r}\sum_{i=1}^{r} s_i,$$
$$\sum_{i=1}^{n} s_i \le \left(1+\frac{n-r}{r}\right)\sum_{i=1}^{r} s_i = \frac{n}{r}\sum_{i=1}^{r} s_i. \tag{A.11}$$

For every $A \in \mathbb{R}^{m\times n}$,

$$\|(Q_L^k)^\top A Q_R^k\|_{\mathrm{op}} \le \|(Q_L^k)^\top\|_{\mathrm{op}}\|A\|_{\mathrm{op}}\|Q_R^k\|_{\mathrm{op}} = \|A\|_{\mathrm{op}}. \tag{A.12}$$

Also, $0 \preceq Q_L^k(Q_L^k)^\top \preceq I_m$ and $0 \preceq Q_R^k(Q_R^k)^\top \preceq I_n$.

$$\begin{aligned}\|(Q_L^k)^\top A Q_R^k\|_{\mathrm{F}}^2 &= \mathrm{tr}\big((Q_R^k)^\top A^\top Q_L^k (Q_L^k)^\top A Q_R^k\big)\\ &\le \mathrm{tr}\big((Q_R^k)^\top A^\top A Q_R^k\big)\\ &= \mathrm{tr}\big(A^\top A Q_R^k (Q_R^k)^\top\big) \le \mathrm{tr}(A^\top A) = \|A\|_{\mathrm{F}}^2.\end{aligned} \tag{A.13}$$

For an $r \times r$ matrix Z,

$$\|Z\|_* = \sum_{i=1}^{r} s_i(Z) \le \sqrt{r}\left(\sum_{i=1}^{r} s_i(Z)^2\right)^{1/2} = \sqrt{r}\|Z\|_{\mathrm{F}},$$
$$\|Z\|_* \le r \max_{1\le i\le r} s_i(Z) = r\|Z\|_{\mathrm{op}}. \tag{A.14}$$

It gives

$$\begin{aligned}\|(Q_L^k)^\top A Q_R^k\|_{\mathrm{op}} &\le \|A\|_{\mathrm{op}}, & \|(Q_L^k)^\top A Q_R^k\|_{\mathrm{F}} &\le \|A\|_{\mathrm{F}},\\ \|(Q_L^k)^\top A Q_R^k\|_* &\le \sqrt{r}\|A\|_{\mathrm{F}}, & \|(Q_L^k)^\top A Q_R^k\|_* &\le r\|A\|_{\mathrm{op}},\\ \|A\|_* &\le \sqrt{n}\|A\|_{\mathrm{F}}, & \|A\|_* &\le n\|A\|_{\mathrm{op}}.\end{aligned} \tag{A.15}$$

Then we decompose

$$\overline{M}^k = C^{sT} + (\overline{M}^k - \overline{M}^{sT}) + (\overline{M}^{sT} - C^{sT}). \tag{A.16}$$

By the triangle inequality,

$$\begin{aligned}\|(Q_L^k)^\top \overline{M}^k Q_R^k\|_* &\ge \|(Q_L^k)^\top C^{sT} Q_R^k\|_* - \|(Q_L^k)^\top(\overline{M}^k - \overline{M}^{sT})Q_R^k\|_*\\ &\quad - \|(Q_L^k)^\top(\overline{M}^{sT} - C^{sT})Q_R^k\|_*\\ &\ge \frac{r}{n}\|C^{sT}\|_* - \sqrt{r}\|\overline{M}^k - \overline{M}^{sT}\|_{\mathrm{F}} - r\|\overline{M}^{sT} - C^{sT}\|_{\mathrm{op}},\end{aligned} \tag{A.17}$$

$$\begin{aligned}\|C^{sT}\|_* &\ge \|\overline{M}^k\|_* - \|\overline{M}^k - \overline{M}^{sT}\|_* - \|\overline{M}^{sT} - C^{sT}\|_*\\ &\ge \|\overline{M}^k\|_* - \sqrt{n}\|\overline{M}^k - \overline{M}^{sT}\|_{\mathrm{F}} - n\|\overline{M}^{sT} - C^{sT}\|_{\mathrm{op}}.\end{aligned} \tag{A.18}$$

Substituting (A.18) into (A.17) gives

$$\begin{aligned}\|(Q_L^k)^\top \overline{M}^k Q_R^k\|_* \ge \frac{r}{n}\|\overline{M}^k\|_* &- \left(\sqrt{r}+\frac{r}{\sqrt{n}}\right)\|\overline{M}^k - \overline{M}^{sT}\|_{\mathrm{F}}\\ &- 2r\|\overline{M}^{sT} - C^{sT}\|_{\mathrm{op}}.\end{aligned} \tag{A.19}$$

The reverse triangle inequality gives

$$\|\overline{M}^k\|_* \ge \|\nabla f(X^k)\|_* - \|\overline{M}^k - \nabla f(X^k)\|_*. \tag{A.20}$$

Using (A.9), (A.19), and (A.20) in (A.6),

$$\begin{aligned}-\langle \nabla f(X^k), O^k\rangle \ge \frac{r}{n}\|\nabla f(X^k)\|_* &- \frac{r}{n}\|\overline{M}^k - \nabla f(X^k)\|_* + \langle \overline{M}^k - \nabla f(X^k), O^k\rangle\\ &- \left(\sqrt{r}+\frac{r}{\sqrt{n}}\right)\|\overline{M}^k - \overline{M}^{sT}\|_{\mathrm{F}} - 2r\|\overline{M}^{sT} - C^{sT}\|_{\mathrm{op}}\\ &- 2r\|(Q_L^k)^\top \overline{M}^k Q_R^k - \widehat{B}^k\|_{\mathrm{op}}.\end{aligned} \tag{A.21}$$

Substituting into (A.5) and rearranging,

$$\begin{aligned}\frac{r\eta}{n}\|\nabla f(X^k)\|_* \le f(X^k) - f(X^{k+1}) + \frac{L\eta^2}{2} \\ + \eta\left[\frac{r}{n}\|\overline{M}^k - \nabla f(X^k)\|_* - \langle \overline{M}^k - \nabla f(X^k), O^k\rangle\right] \\ + \eta\left(\sqrt{r} + \frac{r}{\sqrt{n}}\right)\|\overline{M}^k - \overline{M}^{sT}\|_{\mathrm{F}} \\ + 2r\eta\|\overline{M}^{sT} - C^{sT}\|_{\mathrm{op}} + 2r\eta\|(Q_L^k)^\top \overline{M}^k Q_R^k - \widehat{B}^k\|_{\mathrm{op}}. \end{aligned} \tag{A.22}$$

We first bound the momentum error and then the two compression errors.

Momentum and compression errors. Let $\varepsilon_j^k = \nabla f_j(X^k;\xi_j^k) - \nabla f_j(X^k)$ for $k \ge 1$. Let $\mathcal{F}_{k-1}$ contain all information before the round-k gradient samples, including X^k. Fresh sampling and Assumption 3 give

$$\mathbb{E}[\varepsilon_j^k \mid \mathcal{F}_{k-1}] = 0, \qquad \mathbb{E}[\|\varepsilon_j^k\|_{\mathrm{F}}^2 \mid \mathcal{F}_{k-1}] \le \sigma^2. \tag{A.23}$$

Subtracting $\nabla f_j(X^k)$ from the momentum recursion,

$$\begin{aligned} M_j^k - \nabla f_j(X^k) &= (1-\beta)(M_j^{k-1} - \nabla f_j(X^k)) + \beta\varepsilon_j^k \\ &= (1-\beta)(M_j^{k-1} - \nabla f_j(X^{k-1})) + (1-\beta)(\nabla f_j(X^{k-1}) - \nabla f_j(X^k)) + \beta\varepsilon_j^k. \end{aligned} \tag{A.24}$$

Iterating (A.24) gives

$$\begin{aligned} M_j^k - \nabla f_j(X^k) = (1-\beta)^k\big(M_j^0 - \nabla f_j(X^0)\big) \\ + \sum_{t=1}^{k}(1-\beta)^{k-t}\begin{bmatrix}(1-\beta)\big(\nabla f_j(X^{t-1}) - \nabla f_j(X^t)\big) \\ + \beta\varepsilon_j^t\end{bmatrix}. \end{aligned} \tag{A.25}$$

The initialization makes the first term zero. Separating the two sums,

$$\begin{aligned} M_j^k - \nabla f_j(X^k) &= D_j^k + Z_j^k, \\ D_j^k &= \sum_{t=1}^{k}(1-\beta)^{k-t+1}(\nabla f_j(X^{t-1}) - \nabla f_j(X^t)), \\ Z_j^k &= \beta\sum_{t=1}^{k}(1-\beta)^{k-t}\varepsilon_j^t. \end{aligned} \tag{A.26}$$

Averaging over workers,

$$\begin{aligned} \overline{M}^k - \nabla f(X^k) &= D^k + Z^k, \\ D^k &= \frac{1}{N}\sum_j D_j^k = \sum_{t=1}^{k}(1-\beta)^{k-t+1}(\nabla f(X^{t-1}) - \nabla f(X^t)), \\ Z^k &= \frac{1}{N}\sum_j Z_j^k = \frac{\beta}{N}\sum_{t=1}^{k}(1-\beta)^{k-t}\sum_j \varepsilon_j^t. \end{aligned} \tag{A.27}$$

From (A.4), $\|X^t - X^{t-1}\|_{\mathrm{op}} \le \eta$. Then we get

$$\begin{aligned} \|D_j^k\|_* &\le \sum_{t=1}^{k}(1-\beta)^{k-t+1}L_j\|X^{t-1} - X^t\|_{\mathrm{op}} \\ &\le L_j\eta\sum_{h=1}^{k}(1-\beta)^h = \frac{(1-\beta)[1-(1-\beta)^k]}{\beta}L_j\eta \le \frac{(1-\beta)L_j\eta}{\beta}. \end{aligned} \tag{A.28}$$

Using the global smoothness constant in (A.27),

$$\|D^k\|_* \le L\eta \sum_{h=1}^{k} (1-\beta)^h \le \frac{(1-\beta)L\eta}{\beta}. \tag{A.29}$$

For $s < t$, ε_j^s is $\mathcal{F}_{t-1}$-measurable, so

$$\mathbb{E}\langle \varepsilon_j^s, \varepsilon_l^t \rangle = \mathbb{E}\langle \varepsilon_j^s, \mathbb{E}[\varepsilon_l^t \mid \mathcal{F}_{t-1}] \rangle = 0. \tag{A.30}$$

For $j \neq l$, conditional independence within round t also gives

$$\mathbb{E}[\langle \varepsilon_j^t, \varepsilon_l^t \rangle \mid \mathcal{F}_{t-1}] = \langle \mathbb{E}[\varepsilon_j^t \mid \mathcal{F}_{t-1}], \mathbb{E}[\varepsilon_l^t \mid \mathcal{F}_{t-1}] \rangle = 0. \tag{A.31}$$

Expanding the squared Frobenius norm and using (A.30),

$$\begin{aligned}
\mathbb{E}\|Z_j^k\|_{\mathrm{F}}^2 &= \beta^2 \sum_{t=1}^{k} (1-\beta)^{2(k-t)} \mathbb{E}\|\varepsilon_j^t\|_{\mathrm{F}}^2 \\
&\le \beta^2 \sigma^2 \sum_{h=0}^{k-1} (1-\beta)^{2h} \\
&= \frac{\beta^2 [1-(1-\beta)^{2k}]}{1-(1-\beta)^2} \sigma^2 \\
&= \frac{\beta [1-(1-\beta)^{2k}]}{2-\beta} \sigma^2.
\end{aligned} \tag{A.32}$$

$$\begin{aligned}
\mathbb{E}\|Z^k\|_{\mathrm{F}}^2 &= \frac{\beta^2}{N^2} \sum_{t=1}^{k} \sum_{v=1}^{k} (1-\beta)^{2k-t-v} \sum_{j=1}^{N} \sum_{l=1}^{N} \mathbb{E}\langle \varepsilon_j^t, \varepsilon_l^v \rangle \\
&= \frac{\beta^2}{N^2} \sum_{t=1}^{k} (1-\beta)^{2(k-t)} \sum_{j=1}^{N} \mathbb{E}\|\varepsilon_j^t\|_{\mathrm{F}}^2 \\
&\le \frac{\beta^2 \sigma^2}{N} \sum_{h=0}^{k-1} (1-\beta)^{2h} = \frac{\beta [1-(1-\beta)^{2k}]}{N(2-\beta)} \sigma^2.
\end{aligned} \tag{A.33}$$

We apply Jensen's inequality to give

$$\begin{aligned}
\mathbb{E}\|Z_j^k\|_{\mathrm{F}} &\le \left(\mathbb{E}\|Z_j^k\|_{\mathrm{F}}^2\right)^{1/2} \le \sigma \sqrt{\frac{\beta[1-(1-\beta)^{2k}]}{2-\beta}} \le \sigma \sqrt{\frac{\beta}{2-\beta}}, \\
\mathbb{E}\|Z^k\|_{\mathrm{F}} &\le \left(\mathbb{E}\|Z^k\|_{\mathrm{F}}^2\right)^{1/2} \le \frac{\sigma}{\sqrt{N}} \sqrt{\frac{\beta[1-(1-\beta)^{2k}]}{2-\beta}} \le \frac{\sigma}{\sqrt{N}} \sqrt{\frac{\beta}{2-\beta}}.
\end{aligned} \tag{A.34}$$

For the two inner products, Hölder's inequality and Frobenius Cauchy–Schwarz give, respectively,

$$\begin{aligned}
|\langle D^k, O^k \rangle| &\le \|D^k\|_* \|O^k\|_{\mathrm{op}} \le \|D^k\|_*, \\
|\langle Z^k, O^k \rangle| &\le \|Z^k\|_{\mathrm{F}} \|O^k\|_{\mathrm{F}} \le \sqrt{r} \|Z^k\|_{\mathrm{F}}.
\end{aligned} \tag{A.35}$$

Using these inequalities and $\|Z^k\|_* \le \sqrt{n}\|Z^k\|_{\mathrm{F}}$ in (A.27),

$$\begin{aligned}
&\frac{r}{n} \|\overline{M}^k - \nabla f(X^k)\|_* - \langle \overline{M}^k - \nabla f(X^k), O^k \rangle \\
&\le \frac{r}{n} \left(\|D^k\|_* + \|Z^k\|_*\right) + |\langle D^k, O^k \rangle| + |\langle Z^k, O^k \rangle| \\
&\le \frac{r}{n} \left(\|D^k\|_* + \sqrt{n}\|Z^k\|_{\mathrm{F}}\right) + \|D^k\|_* + \sqrt{r}\|Z^k\|_{\mathrm{F}} \\
&= \left(1 + \frac{r}{n}\right) \|D^k\|_* + \left(\sqrt{r} + \frac{r}{\sqrt{n}}\right) \|Z^k\|_{\mathrm{F}}.
\end{aligned} \tag{A.36}$$

Taking expectations and using (A.29) and (A.34),

$$\mathbb{E}\left[\frac{r}{n}\|\overline{M}^k - \nabla f(X^k)\|_* - \langle \overline{M}^k - \nabla f(X^k), O^k\rangle\right] \leq \left(1+\frac{r}{n}\right)\frac{(1-\beta)L\eta}{\beta} + \left(\sqrt{r}+\frac{r}{\sqrt{n}}\right)\frac{\sigma}{\sqrt{N}}\sqrt{\frac{\beta}{2-\beta}}. \tag{A.37}$$

Subtracting M_j^{k-1} from the momentum recursion,

$$\begin{aligned} M_j^k - M_j^{k-1} &= \beta(g_j^k - M_j^{k-1}) \\ &= \beta\big[\nabla f_j(X^k) - \nabla f_j(X^{k-1}) - D_j^{k-1} - Z_j^{k-1} + \varepsilon_j^k\big]. \end{aligned} \tag{A.38}$$

Since $\|A\|_{\mathrm{F}} \leq \|A\|_*$ and $\mathbb{E}\|\varepsilon_j^k\|_{\mathrm{F}} \leq \sigma$,

$$\begin{aligned} \mathbb{E}\|M_j^k - M_j^{k-1}\|_{\mathrm{F}} &\leq \beta L_j\eta + (1-\beta)L_j\eta + \beta\sigma\sqrt{\frac{\beta}{2-\beta}} + \beta\sigma \\ &= L_j\eta + \beta\sigma\left(1+\sqrt{\frac{\beta}{2-\beta}}\right). \end{aligned} \tag{A.39}$$

$$\overline{M}^k - \overline{M}^{k-1} = \beta\left[\nabla f(X^k) - \nabla f(X^{k-1}) - D^{k-1} - Z^{k-1} + \frac{1}{N}\sum_j \varepsilon_j^k\right]. \tag{A.40}$$

By (A.31),

$$\mathbb{E}\|\frac{1}{N}\sum_j \varepsilon_j^k\|_{\mathrm{F}}^2 = \frac{1}{N^2}\sum_j \mathbb{E}\|\varepsilon_j^k\|_{\mathrm{F}}^2 \leq \frac{\sigma^2}{N}, \qquad \mathbb{E}\|\frac{1}{N}\sum_j \varepsilon_j^k\|_{\mathrm{F}} \leq \frac{\sigma}{\sqrt{N}}. \tag{A.41}$$

$$\begin{aligned} \mathbb{E}\|\overline{M}^k - \overline{M}^{k-1}\|_{\mathrm{F}} &\leq \beta L\eta + (1-\beta)L\eta + \frac{\beta\sigma}{\sqrt{N}}\sqrt{\frac{\beta}{2-\beta}} + \frac{\beta\sigma}{\sqrt{N}} \\ &= L\eta + \frac{\beta\sigma}{\sqrt{N}}\left(1+\sqrt{\frac{\beta}{2-\beta}}\right). \end{aligned} \tag{A.42}$$

For $k = sT + \ell$, $0 \leq \ell < T$, it follows that

$$\begin{aligned} \mathbb{E}\|\overline{M}^k - \overline{M}^{sT}\|_{\mathrm{F}} &\leq \sum_{t=sT+1}^{sT+\ell} \mathbb{E}\|\overline{M}^t - \overline{M}^{t-1}\|_{\mathrm{F}} \\ &\leq \ell\left[L\eta + \frac{\beta\sigma}{\sqrt{N}}\left(1+\sqrt{\frac{\beta}{2-\beta}}\right)\right] \\ &\leq (T-1)\left[L\eta + \frac{\beta\sigma}{\sqrt{N}}\left(1+\sqrt{\frac{\beta}{2-\beta}}\right)\right]. \end{aligned} \tag{A.43}$$

The difference is zero when $\ell = 0$.

The server and worker full-momentum approximations satisfy $C^k = \frac{1}{N}\sum_j C_j^k$. The identity holds at initialization and is preserved at a full-momentum correction round:

$$\begin{aligned} C^{sT} &= C^{sT-1} + \frac{1}{N}\sum_j \Delta C_j^{sT} \\ &= \frac{1}{N}\sum_j (C_j^{sT-1} + \Delta C_j^{sT}) = \frac{1}{N}\sum_j C_j^{sT}. \end{aligned} \tag{A.44}$$

Let $\mathcal{H}$ contain the preceding history and current gradients and compression inputs, before any current compression outputs. For an $\mathcal{H}$-measurable input V, contractivity and conditional Jensen give

$$\begin{aligned}\mathbb{E}[\|\mathcal{C}_{\mathrm{a}}(V) - V\|_{\mathrm{F}} \mid \mathcal{H}] &\le \sqrt{\mathbb{E}[\|\mathcal{C}_{\mathrm{a}}(V) - V\|_{\mathrm{F}}^2 \mid \mathcal{H}]} \\ &\le \sqrt{1-\alpha_{\mathrm{a}}}\|V\|_{\mathrm{F}}. \end{aligned} \tag{A.45}$$

At a full-momentum correction round, $\mathcal{H}$ excludes the new bases, which are selected after compression. The conditional bound allows correlated worker seeds.

For $s \ge 1$, $C_j^{sT-1} = C_j^{(s-1)T}$, so

$$C_j^{sT} - M_j^{sT} = \mathcal{C}_{\mathrm{a}}^{sT,j}(M_j^{sT} - C_j^{(s-1)T}) - (M_j^{sT} - C_j^{(s-1)T}). \tag{A.46}$$

Taking expectations and applying (A.45),

$$\begin{aligned}\mathbb{E}\|C_j^{sT} - M_j^{sT}\|_{\mathrm{F}} &\le \sqrt{1-\alpha_{\mathrm{a}}}\mathbb{E}\|M_j^{sT} - C_j^{(s-1)T}\|_{\mathrm{F}} \\ &\le \sqrt{1-\alpha_{\mathrm{a}}}\mathbb{E}\|C_j^{(s-1)T} - M_j^{(s-1)T}\|_{\mathrm{F}} \\ &\quad + \sqrt{1-\alpha_{\mathrm{a}}} \sum_{t=(s-1)T+1}^{sT} \mathbb{E}\|M_j^t - M_j^{t-1}\|_{\mathrm{F}} \\ &\le \sqrt{1-\alpha_{\mathrm{a}}}\mathbb{E}\|C_j^{(s-1)T} - M_j^{(s-1)T}\|_{\mathrm{F}} \\ &\quad + T\sqrt{1-\alpha_{\mathrm{a}}}\left[L_j\eta + \beta\sigma\left(1 + \sqrt{\frac{\beta}{2-\beta}}\right)\right]. \end{aligned} \tag{A.47}$$

We further get

$$\begin{aligned}&\sqrt{1-\alpha_{\mathrm{a}}}\sum_{v=1}^{s-1}(1-\alpha_{\mathrm{a}})^{v/2} + \sqrt{1-\alpha_{\mathrm{a}}} \\ &= \sum_{v=2}^{s}(1-\alpha_{\mathrm{a}})^{v/2} + (1-\alpha_{\mathrm{a}})^{1/2} = \sum_{v=1}^{s}(1-\alpha_{\mathrm{a}})^{v/2}. \end{aligned} \tag{A.48}$$

Its closed form follows by subtracting the shifted sum

$$\begin{aligned}&\left(1-\sqrt{1-\alpha_{\mathrm{a}}}\right)\sum_{v=1}^{s}(1-\alpha_{\mathrm{a}})^{v/2} \\ &= \sum_{v=1}^{s}(1-\alpha_{\mathrm{a}})^{v/2} - \sum_{v=1}^{s}(1-\alpha_{\mathrm{a}})^{(v+1)/2} \\ &= \sqrt{1-\alpha_{\mathrm{a}}} - (1-\alpha_{\mathrm{a}})^{(s+1)/2}. \end{aligned} \tag{A.49}$$

Consequently,

$$\begin{aligned}\mathbb{E}\|C_j^{sT} - M_j^{sT}\|_{\mathrm{F}} &\le T\left[L_j\eta + \beta\sigma\left(1 + \sqrt{\frac{\beta}{2-\beta}}\right)\right]\sum_{v=1}^{s}(1-\alpha_{\mathrm{a}})^{v/2} \\ &= \frac{T\sqrt{1-\alpha_{\mathrm{a}}}[1-(1-\alpha_{\mathrm{a}})^{s/2}]}{1-\sqrt{1-\alpha_{\mathrm{a}}}}\left[L_j\eta + \beta\sigma\left(1 + \sqrt{\frac{\beta}{2-\beta}}\right)\right] \\ &\le \frac{T\sqrt{1-\alpha_{\mathrm{a}}}}{1-\sqrt{1-\alpha_{\mathrm{a}}}}\left[L_j\eta + \beta\sigma\left(1 + \sqrt{\frac{\beta}{2-\beta}}\right)\right]. \end{aligned} \tag{A.50}$$

Thus

$$\begin{aligned}
&\frac{1}{N}\sum_{j=1}^{N}\left[L_j\eta+\beta\sigma\left(1+\sqrt{\frac{\beta}{2-\beta}}\right)\right]\\
&=\frac{\eta}{N}\sum_{j=1}^{N}L_j+\frac{1}{N}\left(\sum_{j=1}^{N}1\right)\beta\sigma\left(1+\sqrt{\frac{\beta}{2-\beta}}\right)\\
&=\frac{\eta}{N}\sum_{j=1}^{N}L_j+\beta\sigma\left(1+\sqrt{\frac{\beta}{2-\beta}}\right). \qquad \text{(A.51)}
\end{aligned}$$

To use the Frobenius estimates in the descent bound, we apply $\|A\|_{\mathrm{op}}\le\|A\|_{\mathrm{F}}$. Using (A.51) and (A.44),

$$\begin{aligned}
\mathbb{E}\|\overline{M}^{sT}-C^{sT}\|_{\mathrm{op}}&\le\mathbb{E}\|\frac{1}{N}\sum_{j=1}^{N}(M_j^{sT}-C_j^{sT})\|_{\mathrm{F}}\\
&\le\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\|M_j^{sT}-C_j^{sT}\|_{\mathrm{F}}\\
&\le\frac{T\sqrt{1-\alpha_{\mathrm{a}}}}{1-\sqrt{1-\alpha_{\mathrm{a}}}}\left[\frac{\eta}{N}\sum_{j=1}^{N}L_j+\beta\sigma\left(1+\sqrt{\frac{\beta}{2-\beta}}\right)\right]. \qquad \text{(A.52)}
\end{aligned}$$

At sT, the algorithm resets the core to the projected full-momentum approximation. Hence

$$\begin{aligned}
\widehat{B}_j^{sT}-(Q_L^{sT})^\top M_j^{sT}Q_R^{sT}&=(Q_L^{sT})^\top(C_j^{sT}-M_j^{sT})Q_R^{sT},\\
\mathbb{E}\|\widehat{B}_j^{sT}-(Q_L^{sT})^\top M_j^{sT}Q_R^{sT}\|_{\mathrm{F}}&\le\mathbb{E}\|C_j^{sT}-M_j^{sT}\|_{\mathrm{F}}. \qquad \text{(A.53)}
\end{aligned}$$

Inside the interval, $k=sT+\ell$ with $1\le\ell<T$, the bases are fixed. Subtracting the projected momentum from the core update,

$$\begin{aligned}
&\widehat{B}_j^{k}-(Q_L^{k})^\top M_j^{k}Q_R^{k}\\
&=\mathcal{C}_{\mathrm{b}}^{k,j}\big((Q_L^{k})^\top M_j^{k}Q_R^{k}-\widehat{B}_j^{k-1}\big)-\big((Q_L^{k})^\top M_j^{k}Q_R^{k}-\widehat{B}_j^{k-1}\big). \qquad \text{(A.54)}
\end{aligned}$$

For an $\mathcal{H}$-measurable core input Z, the same conditional argument gives

$$\mathbb{E}[\|\mathcal{C}_{\mathrm{b}}(Z)-Z\|_{\mathrm{F}}\mid\mathcal{H}]\le\sqrt{\mathbb{E}[\|\mathcal{C}_{\mathrm{b}}(Z)-Z\|_{\mathrm{F}}^2\mid\mathcal{H}]}\le\sqrt{1-\alpha_{\mathrm{b}}}\|Z\|_{\mathrm{F}}. \qquad \text{(A.55)}$$

Since $Q_L^k=Q_L^{k-1}$ and $Q_R^k=Q_R^{k-1}$,

$$\begin{aligned}
&\mathbb{E}\|\widehat{B}_j^{k}-(Q_L^{k})^\top M_j^{k}Q_R^{k}\|_{\mathrm{F}}\\
&\le\sqrt{1-\alpha_{\mathrm{b}}}\mathbb{E}\|(Q_L^{k})^\top M_j^{k}Q_R^{k}-\widehat{B}_j^{k-1}\|_{\mathrm{F}}\\
&\le\sqrt{1-\alpha_{\mathrm{b}}}\mathbb{E}\|\widehat{B}_j^{k-1}-(Q_L^{k-1})^\top M_j^{k-1}Q_R^{k-1}\|_{\mathrm{F}}\\
&\quad+\sqrt{1-\alpha_{\mathrm{b}}}\mathbb{E}\|(Q_L^{k})^\top(M_j^{k}-M_j^{k-1})Q_R^{k}\|_{\mathrm{F}}\\
&\le\sqrt{1-\alpha_{\mathrm{b}}}\mathbb{E}\|\widehat{B}_j^{k-1}-(Q_L^{k-1})^\top M_j^{k-1}Q_R^{k-1}\|_{\mathrm{F}}\\
&\quad+\sqrt{1-\alpha_{\mathrm{b}}}\left[L_j\eta+\beta\sigma\left(1+\sqrt{\frac{\beta}{2-\beta}}\right)\right]. \qquad \text{(A.56)}
\end{aligned}$$

Iterating within the current interval and using (A.53),

$$\begin{aligned}
&\mathbb{E}\|\widehat{B}_j^{sT+\ell} - (Q_L^{sT+\ell})^\top M_j^{sT+\ell} Q_R^{sT+\ell}\|_{\mathrm{F}} \\
&\le (1-\alpha_{\mathrm{b}})^{\ell/2}\mathbb{E}\|\widehat{B}_j^{sT} - (Q_L^{sT})^\top M_j^{sT} Q_R^{sT}\|_{\mathrm{F}} \\
&\qquad + \sum_{h=1}^{\ell}(1-\alpha_{\mathrm{b}})^{(\ell-h+1)/2}\left[L_j\eta + \beta\sigma\left(1+\sqrt{\frac{\beta}{2-\beta}}\right)\right] \\
&\le (1-\alpha_{\mathrm{b}})^{\ell/2}\mathbb{E}\|C_j^{sT} - M_j^{sT}\|_{\mathrm{F}} \\
&\qquad + \left[\sum_{v=1}^{\ell}(1-\alpha_{\mathrm{b}})^{v/2}\right]\left[L_j\eta + \beta\sigma\left(1+\sqrt{\frac{\beta}{2-\beta}}\right)\right] \\
&\le \left[\frac{T\sqrt{1-\alpha_{\mathrm{a}}}}{1-\sqrt{1-\alpha_{\mathrm{a}}}} + \sum_{v=1}^{T-1}(1-\alpha_{\mathrm{b}})^{v/2}\right]\left[L_j\eta + \beta\sigma\left(1+\sqrt{\frac{\beta}{2-\beta}}\right)\right] \\
&= \left[\frac{T\sqrt{1-\alpha_{\mathrm{a}}}}{1-\sqrt{1-\alpha_{\mathrm{a}}}} + \frac{\sqrt{1-\alpha_{\mathrm{b}}} - (1-\alpha_{\mathrm{b}})^{T/2}}{1-\sqrt{1-\alpha_{\mathrm{b}}}}\right]\left[L_j\eta + \beta\sigma\left(1+\sqrt{\frac{\beta}{2-\beta}}\right)\right]. \qquad \text{(A.57)}
\end{aligned}$$

In the second bound, use (A.53) and set $v = \ell - h + 1$. We get

$$\left(1-\sqrt{1-\alpha_{\mathrm{b}}}\right)\sum_{v=1}^{T-1}(1-\alpha_{\mathrm{b}})^{v/2} = \sqrt{1-\alpha_{\mathrm{b}}} - (1-\alpha_{\mathrm{b}})^{T/2}. \qquad \text{(A.58)}$$

The identity holds initially and is preserved at a projected-momentum round:

$$\widehat{B}^k = \frac{1}{N}\sum_j \widehat{B}_j^{k-1} + \frac{1}{N}\sum_j R_j^k = \frac{1}{N}\sum_j \widehat{B}_j^k. \qquad \text{(A.59)}$$

At a full-momentum correction round, the common new bases give

$$\begin{aligned}
\frac{1}{N}\sum_j \widehat{B}_j^{sT} &= (Q_L^{sT})^\top\left(\frac{1}{N}\sum_j C_j^{sT}\right)Q_R^{sT} \\
&= (Q_L^{sT})^\top C^{sT} Q_R^{sT} = \widehat{B}^{sT}. \qquad \text{(A.60)}
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\mathbb{E}\|(Q_L^k)^\top \overline{M}^k Q_R^k - \widehat{B}^k\|_{\mathrm{op}} \\
&\le \mathbb{E}\|\frac{1}{N}\sum_j\left((Q_L^k)^\top M_j^k Q_R^k - \widehat{B}_j^k\right)\|_{\mathrm{F}} \\
&\le \frac{1}{N}\sum_j \mathbb{E}\|(Q_L^k)^\top M_j^k Q_R^k - \widehat{B}_j^k\|_{\mathrm{F}} \\
&\le \left[\frac{T\sqrt{1-\alpha_{\mathrm{a}}}}{1-\sqrt{1-\alpha_{\mathrm{a}}}} + \frac{\sqrt{1-\alpha_{\mathrm{b}}} - (1-\alpha_{\mathrm{b}})^{T/2}}{1-\sqrt{1-\alpha_{\mathrm{b}}}}\right]\left[\frac{\eta}{N}\sum_{j=1}^{N} L_j + \beta\sigma\left(1+\sqrt{\frac{\beta}{2-\beta}}\right)\right]. \qquad \text{(A.61)}
\end{aligned}$$

Taking expectations in (A.22), substitute (A.37), (A.43), (A.52), and (A.61) to obtain

$$\begin{aligned}
\frac{r\eta}{n}\mathbb{E}\|\nabla f(X^k)\|_* \le{}& \mathbb{E}f(X^k) - \mathbb{E}f(X^{k+1}) + \frac{L\eta^2}{2} \\
&+ \left(1+\frac{r}{n}\right)\frac{L\eta^2(1-\beta)}{\beta} + \eta\left(\sqrt{r}+\frac{r}{\sqrt{n}}\right)\frac{\sigma}{\sqrt{N}}\sqrt{\frac{\beta}{2-\beta}} \\
&+ \eta\left(\sqrt{r}+\frac{r}{\sqrt{n}}\right)(T-1)\left[L\eta + \frac{\beta\sigma}{\sqrt{N}}\left(1+\sqrt{\frac{\beta}{2-\beta}}\right)\right]
\end{aligned}$$

$$+ 2r\eta \left[\frac{2T\sqrt{1-\alpha_{\mathrm{a}}}}{1-\sqrt{1-\alpha_{\mathrm{a}}}} + \frac{\sqrt{1-\alpha_{\mathrm{b}}} - (1-\alpha_{\mathrm{b}})^{T/2}}{1-\sqrt{1-\alpha_{\mathrm{b}}}} \right] \times \left[\frac{\eta}{N} \sum_{j=1}^{N} L_j + \beta\sigma \left(1 + \sqrt{\frac{\beta}{2-\beta}} \right) \right]. \tag{A.62}$$

Summing (A.62) over $k = 0, \ldots, K-1$, the objective terms satisfy

$$\sum_{k=0}^{K-1} \left[\mathbb{E} f(X^k) - \mathbb{E} f(X^{k+1}) \right] = f(X^0) - \mathbb{E} f(X^K) \le f(X^0) - f^\star. \tag{A.63}$$

Divide by $rK\eta/n$. This gives

$$\begin{aligned} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(X^k)\|_* \le{}& \frac{n[f(X^0) - f^\star]}{rK\eta} + \frac{nL\eta}{2r} + \left(\frac{n}{r} + 1 \right) \frac{L\eta(1-\beta)}{\beta} \\ &+ \left(\frac{n}{\sqrt{r}} + \sqrt{n} \right) \frac{\sigma}{\sqrt{N}} \sqrt{\frac{\beta}{2-\beta}} \\ &+ \left(\frac{n}{\sqrt{r}} + \sqrt{n} \right) (T-1) \left[L\eta + \frac{\beta\sigma}{\sqrt{N}} \left(1 + \sqrt{\frac{\beta}{2-\beta}} \right) \right] \\ &+ 2n \left[\frac{2T\sqrt{1-\alpha_{\mathrm{a}}}}{1-\sqrt{1-\alpha_{\mathrm{a}}}} + \frac{\sqrt{1-\alpha_{\mathrm{b}}} - (1-\alpha_{\mathrm{b}})^{T/2}}{1-\sqrt{1-\alpha_{\mathrm{b}}}} \right] \\ &\times \left[\frac{\eta}{N} \sum_{j=1}^{N} L_j + \beta\sigma \left(1 + \sqrt{\frac{\beta}{2-\beta}} \right) \right]. \end{aligned} \tag{A.64}$$

The two compression fractions combine as Γ_T. Collecting the terms proportional to η gives $\mathcal{L}_\beta$, and the remaining terms give exactly (3).

B Proof of Corollary 1

Parameter selection. We give the constant learning rate and momentum for a horizon K. Assume first that $\Delta > 0$ and $L > 0$. We define

$$\begin{aligned} A &= \frac{n\Delta}{r}, & B &= \left(\frac{n}{r} + 1 \right) L, & H &= \mathcal{L}_1, \\ S &= a_{n,r} \frac{\sigma}{\sqrt{N}}, & D &= 2\sigma \left[2n\Gamma_T + \frac{a_{n,r}(T-1)}{\sqrt{N}} \right]. \end{aligned} \tag{B.1}$$

Thus $A, B, H > 0$ and $S, D \ge 0$. Theorem 1 states

$$\begin{aligned} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(X^k)\|_* \le{}& \frac{A}{K\eta} + \eta \left[H + B\frac{1-\beta}{\beta} \right] + S\sqrt{\frac{\beta}{2-\beta}} \\ &+ \frac{D}{2}\beta \left(1 + \sqrt{\frac{\beta}{2-\beta}} \right). \end{aligned} \tag{B.2}$$

The parameters used in the corollary are

$$\beta = \frac{2}{1 + \max\left\{ 1, S/p_K, \left((1+\sqrt{2})D/p_K \right)^{2/3} \right\}}, \qquad \eta = \sqrt{\frac{A}{K[H + B(1-\beta)/\beta]}}, \tag{B.3}$$

where $p_K = \sqrt{2AB/K}$. The maximum is finite and at least one, so $\beta \in (0, 1]$ and $\eta > 0$.

For a fixed β, let $Q_\beta = H + B(1-\beta)/\beta$. The terms depending on the learning rate satisfy

$$\frac{A}{K\eta} + Q_\beta \eta - 2\sqrt{\frac{AQ_\beta}{K}} = \left(\sqrt{\frac{A}{K\eta}} - \sqrt{Q_\beta \eta}\right)^2 \geq 0,$$

$$\frac{A}{K\eta} = Q_\beta \eta \iff \eta = \sqrt{\frac{A}{KQ_\beta}}. \tag{B.4}$$

Thus (B.3) minimizes the upper bound over positive learning rates for its selected β.

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(X^k)\|_* \leq 2\sqrt{\frac{A}{K}\left[H + B\frac{1-\beta}{\beta}\right]} + S\sqrt{\frac{\beta}{2-\beta}} + \frac{D}{2}\beta\left(1 + \sqrt{\frac{\beta}{2-\beta}}\right). \tag{B.5}$$

Balancing the momentum and noise terms. Set $u = \sqrt{\beta/(2-\beta)} \in (0, 1]$. We get

$$\beta = \frac{2u^2}{1+u^2}, \qquad \frac{1-\beta}{\beta} = \frac{u^{-2}-1}{2}, \qquad \frac{D}{2}\beta\left(1 + \sqrt{\frac{\beta}{2-\beta}}\right) = Du^2\frac{1+u}{1+u^2}. \tag{B.6}$$

Let $\rho = (1+\sqrt{2})/2$. Then

$$\rho(1+u^2) - (1+u) = \rho\big(u - (\sqrt{2}-1)\big)^2 \geq 0 \tag{B.7}$$

implies $(1+u)/(1+u^2) \leq \rho$. Applying this inequality and $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ to (B.5) yields

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(X^k)\|_* &\leq 2\sqrt{\frac{A}{K}\left[H + \frac{B}{2}(u^{-2}-1)\right]} + Su + Du^2\frac{1+u}{1+u^2} \\ &\leq 2\sqrt{\frac{AH}{K}} + p_K\sqrt{u^{-2}-1} + Su + \rho D u^2.\end{aligned} \tag{B.8}$$

To obtain the two balancing scales, consider the derivatives

$$\frac{\mathrm{d}}{\mathrm{d}u}\left(\frac{p_K}{u} + Su\right) = -\frac{p_K}{u^2} + S, \qquad \frac{\mathrm{d}}{\mathrm{d}u}\left(\frac{p_K}{u} + \rho D u^2\right) = -\frac{p_K}{u^2} + 2\rho D u. \tag{B.9}$$

For positive noise coefficients, these expressions vanish at $(p_K/S)^{1/2}$ and $(p_K/(2\rho D))^{1/3}$, respectively. Taking the smaller of the two and capping it at one gives

$$u = \Gamma_K^{-1/2}, \qquad \Gamma_K = \max\left\{1, \frac{S}{p_K}, \left(\frac{2\rho D}{p_K}\right)^{2/3}\right\}. \tag{B.10}$$

Since $\beta = 2/(1+\Gamma_K)$ and $2\rho = 1+\sqrt{2}$, this is exactly (B.3). We next verify the bound for every case of the maximum. First,

$$\begin{aligned}u^{-2} - 1 &= \max\left\{0, \frac{S}{p_K} - 1, \left(\frac{2\rho D}{p_K}\right)^{2/3} - 1\right\} \\ &\leq \frac{S}{p_K} + \left(\frac{2\rho D}{p_K}\right)^{2/3}, \\ p_K\sqrt{u^{-2}-1} &\leq \sqrt{p_K S} + (2\rho D p_K^2)^{1/3}.\end{aligned} \tag{B.11}$$

The other entries of the maximum give

$$\begin{aligned}Su &= \frac{S}{\sqrt{\Gamma_K}} \leq \sqrt{p_K S}, \\ \rho D u^2 &= \frac{\rho D}{\Gamma_K} \leq \rho D\left(\frac{p_K}{2\rho D}\right)^{2/3} = 2^{-2/3}(\rho D p_K^2)^{1/3}.\end{aligned} \tag{B.12}$$

The last displayed calculation is used when $D > 0$. If $D = 0$, the term is zero. The first inequality similarly holds trivially when $S = 0$.

Substituting (B.11) and (B.12) into (B.8),

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(X^k)\|_* \le 2\sqrt{\frac{AH}{K}} + 2\sqrt{p_K S} + \left(2^{1/3}+2^{-2/3}\right)(\rho D p_K^2)^{1/3}$$
$$= 2\sqrt{\frac{AH}{K}} + 2^{5/4}\left(\frac{ABS^2}{K}\right)^{1/4} + 3\left(\frac{(1+\sqrt{2})ABD}{4K}\right)^{1/3}. \quad \text{(B.13)}$$

Here $p_K^2 = 2AB/K$ and $2^{1/3}+2^{-2/3} = 3\cdot 2^{-2/3}$. We get

$$\frac{AH}{K} = \frac{n\Delta\mathcal{L}_1}{rK},$$
$$\frac{ABS^2}{K} = \frac{n\Delta}{rK}\left(\frac{n}{r}+1\right)L\frac{a_{n,r}^2\sigma^2}{N},$$
$$\frac{(1+\sqrt{2})ABD}{4K} = \frac{(1+\sqrt{2})n\Delta\sigma}{2rK}\left(\frac{n}{r}+1\right)L\left[2n\Gamma_T + \frac{a_{n,r}(T-1)}{\sqrt{N}}\right]. \quad \text{(B.14)}$$

Inserting these identities into (B.13) proves (4).

C Communication Complexity Analysis

We derive sufficient communication bounds for the expected nuclear norm of the gradient. The primary comparison applies Theorem 1 to AdaSC-Muon and its full-rank, every-iteration correction case, which is normalized EF21-Muon. We then retain the comparison with EF21-Muon's published stochastic theorem. Following EF21-P (Gruntkowska et al., 2023) and Natural Compression (Horvóth et al., 2022), each bound combines a sufficient iteration count with the transmitted bits.

We consider Option I and a single $m \times n$ parameter matrix with $m \ge n$. Communication is counted per worker, with free server broadcasts. Both methods use exact gradient initialization and one dense initial transmission. We assume fresh independent gradient samples across workers and iterations, the stated compressor conditions, and the compact-polar LMO defined in (1). The contraction inequalities hold conditionally on the current input and past history, as ensured by fresh compressor randomness. All problem and compression parameters are fixed as $\varepsilon \to 0^+$, with $\Delta, L, \sigma > 0$. The target $\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(X^k)\|_* \le \varepsilon$ is equivalently the expected nuclear norm at an iterate drawn uniformly from the first K iterates.

Normalized EF21-Muon as a special case. Set $r = n$, $T = 1$, and $\mathcal{C}_{\mathrm{a}} = \mathcal{C}_{\mathrm{E}}$, where $\mathcal{C}_{\mathrm{E}}$ has Frobenius contraction parameter $\alpha_{\mathrm{E}} \in (0, 1]$. The projected compressor is unused and may be set to the identity. Every iteration updates

$$C_j^k = C_j^{k-1} + \mathcal{C}_{\mathrm{E}}(M_j^k - C_j^{k-1}), \qquad C^k = \frac{1}{N}\sum_{j=1}^{N} C_j^k. \quad \text{(C.1)}$$

These are exactly the EF21-Muon momentum approximations G_j^k and G^k under matched initialization and the same gradient and compression samples. The full thin singular bases give $C^k = Q_L^k \widehat{B}^k (Q_R^k)^\top$. By the compact-polar convention, including zero singular values,

$$O^k = -Q_L^k \operatorname{polar}(\widehat{B}^k)(Q_R^k)^\top = -\operatorname{polar}(C^k) = \mathrm{LMO}(G^k). \quad \text{(C.2)}$$

This identity holds for rectangular and rank-deficient matrices. Thus the parameter update is $X^{k+1} = X^k + \eta\,\mathrm{LMO}(G^k)$. The bases are auxiliary in this special case and need not be formed or broadcast by EF21-Muon.

Let $\rho_{\mathrm{E}} = \sqrt{1-\alpha_{\mathrm{E}}}$. Substituting $r = n$ and $T = 1$ gives $a_{n,n} = 2\sqrt{n}$ and $\Gamma_1 = 2\rho_{\mathrm{E}}/(1-\rho_{\mathrm{E}})$. Theorem 1 consequently yields

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(X^k)\|_* &\le \frac{\Delta}{K\eta} + \eta\mathcal{L}_{\mathrm{E},\beta} + F_{\mathrm{E}}(\beta),\\ \mathcal{L}_{\mathrm{E},\beta} &= \frac{L}{2} + 2L\frac{1-\beta}{\beta} + \frac{4n\overline{L}\rho_{\mathrm{E}}}{1-\rho_{\mathrm{E}}},\\ F_{\mathrm{E}}(\beta) &= \frac{2\sqrt{n}\,\sigma}{\sqrt{N}}\sqrt{\frac{\beta}{2-\beta}} + \frac{4n\sigma\rho_{\mathrm{E}}}{1-\rho_{\mathrm{E}}}\beta\left(1+\sqrt{\frac{\beta}{2-\beta}}\right).\end{aligned} \tag{C.3}$$

Sufficient iteration counts. Write the noise term in Theorem 1 as

$$\begin{aligned}F_{\mathrm{A}}(\beta,T) &= \frac{a_{n,r}\sigma}{\sqrt{N}}\sqrt{\frac{\beta}{2-\beta}}\\ &\quad + \beta\sigma\left(1+\sqrt{\frac{\beta}{2-\beta}}\right)\left[2n\Gamma_T + \frac{a_{n,r}(T-1)}{\sqrt{N}}\right].\end{aligned} \tag{C.4}$$

The theorem guarantees

$$\begin{aligned}\frac{n\Delta}{rK} &\le \eta[\varepsilon - F_{\mathrm{A}}(\beta,T)] - \eta^2\mathcal{L}_\beta\\ &= \frac{[\varepsilon - F_{\mathrm{A}}(\beta,T)]^2}{4\mathcal{L}_\beta} - \mathcal{L}_\beta\left(\eta - \frac{\varepsilon - F_{\mathrm{A}}(\beta,T)}{2\mathcal{L}_\beta}\right)^2.\end{aligned} \tag{C.5}$$

The maximizing choice is $\eta = [\varepsilon - F_{\mathrm{A}}(\beta,T)]/(2\mathcal{L}_\beta)$. Optimizing over momentum gives

$$K_{\mathrm{A}}(\varepsilon,T) = \min_{\substack{0<\beta\le 1\\ F_{\mathrm{A}}(\beta,T)<\varepsilon}} \max\left\{1, \left\lceil \frac{4n\Delta\mathcal{L}_\beta}{r[\varepsilon - F_{\mathrm{A}}(\beta,T)]^2}\right\rceil\right\}, \tag{C.6}$$

$$K_{\mathrm{E}}(\varepsilon) = \min_{\substack{0<\beta\le 1\\ F_{\mathrm{E}}(\beta)<\varepsilon}} \max\left\{1, \left\lceil \frac{4\Delta\mathcal{L}_{\mathrm{E},\beta}}{[\varepsilon - F_{\mathrm{E}}(\beta)]^2}\right\rceil\right\}. \tag{C.7}$$

For the baseline, the corresponding step size is $\eta = [\varepsilon - F_{\mathrm{E}}(\beta)]/(2\mathcal{L}_{\mathrm{E},\beta})$. Each method selects its own parameters, which remain constant during the run. Admissible choices exist because both noise terms vanish as $\beta \downarrow 0$. These counts minimize the sufficient bounds supplied by the theorem, not the actual iterations needed by the methods.

Cumulative communication at finite accuracy. Let ℓ_0 be the dense initialization length in bits. Among the subsequent $K-1$ updates, AdaSC-Muon sends $q_K = \lfloor (K-1)/T \rfloor$ full corrections and $K-1-q_K$ projected corrections. The total bits per worker are

$$\begin{aligned}U_{\mathrm{A}}(K,T) &= \ell_0 + q_K\ell_{\mathrm{a}} + (K-1-q_K)\ell_{\mathrm{b}},\\ U_{\mathrm{E}}(K) &= \ell_0 + (K-1)\ell_{\mathrm{E}}.\end{aligned} \tag{C.8}$$

The sufficient communication bounds in Proposition 1 are

$$\mathcal{B}_{\mathrm{A}}(\varepsilon,T) = U_{\mathrm{A}}(K_{\mathrm{A}}(\varepsilon,T),T), \qquad \mathcal{B}_{\mathrm{E}}(\varepsilon) = U_{\mathrm{E}}(K_{\mathrm{E}}(\varepsilon)). \tag{C.9}$$

Their relative difference is $1 - \mathcal{B}_{\mathrm{A}}(\varepsilon,T)/\mathcal{B}_{\mathrm{E}}(\varepsilon)$. This includes initialization and the incomplete final period. In particular,

$$U_{\mathrm{A}}(K,T) = \ell_0 + (K-1)\overline{\ell}_{\mathrm{A}}(T) - \left\{\frac{K-1}{T}\right\}(\ell_{\mathrm{a}} - \ell_{\mathrm{b}}), \tag{C.10}$$

where $\{x\} = x - \lfloor x \rfloor$. The remainder has magnitude at most $|\ell_{\mathrm{a}} - \ell_{\mathrm{b}}|$. A numerical comparison at a specified finite ε requires the problem and contraction constants in (C.6)–(C.7).

Proof of Proposition 1. Let $A = n\Delta/r$, $B = (n/r+1)L$, and $S = a_{n,r}\sigma/\sqrt{N}$. For fixed T and contraction parameters,

$$\mathcal{L}_\beta = \frac{B}{\beta} + \mathcal{L}_1 - B, \qquad F_{\mathrm{A}}(\beta, T) = S\sqrt{\frac{\beta}{2}} + \mathcal{O}(\beta). \tag{C.11}$$

For every admissible β, $F_{\mathrm{A}}(\beta,T) \geq S\sqrt{\beta/2}$ implies $\beta < 2\varepsilon^2/S^2$. Set $x = S\sqrt{\beta}/(\sqrt{2}\varepsilon) \in (0,1)$. Uniformly over this range, it satisfies

$$\begin{aligned} \frac{4A\mathcal{L}_\beta}{[\varepsilon - F_{\mathrm{A}}(\beta,T)]^2} &\geq \frac{2ABS^2[1-\mathcal{O}(\varepsilon^2)]}{x^2(1-x)^2\varepsilon^4} \\ &\geq \frac{32ABS^2[1-\mathcal{O}(\varepsilon^2)]}{\varepsilon^4}, \end{aligned} \tag{C.12}$$

where $x(1-x) \leq 1/4$. Conversely, choosing $\beta = \varepsilon^2/(2S^2)$ gives $F_{\mathrm{A}} = \varepsilon/2 + \mathcal{O}(\varepsilon^2)$ and $\mathcal{L}_\beta = 2BS^2/\varepsilon^2 + \mathcal{O}(1)$. This choice is admissible for sufficiently small ε and attains the same leading coefficient. Therefore

$$K_{\mathrm{A}}(\varepsilon, T) \sim \frac{32\frac{n}{r}\left(\frac{n}{r}+1\right)a_{n,r}^2 \Delta L \sigma^2}{N\varepsilon^4}. \tag{C.13}$$

Here $u(\varepsilon) \sim v(\varepsilon)$ means $\lim_{\varepsilon\to 0^+} u(\varepsilon)/v(\varepsilon) = 1$.

For normalized EF21-Muon, the corresponding coefficients in (C.3) are Δ, $2L$, and $2\sqrt{n}\,\sigma/\sqrt{N}$. Substitution into $32ABS^2/\varepsilon^4$ gives

$$K_{\mathrm{E}}(\varepsilon) \sim \frac{256n\Delta L\sigma^2}{N\varepsilon^4}. \tag{C.14}$$

Both iteration counts diverge. Integer rounding, initialization, and the bounded remainder in (C.10) therefore vanish in the communication ratio. Dividing (C.13) by (C.14) and multiplying by $\overline{\ell}_{\mathrm{A}}(T)/\ell_{\mathrm{E}}$ proves (5).

The ratio equals one for $r = n$, $T = 1$, and $\ell_{\mathrm{a}} = \ell_{\mathrm{E}}$, as expected for the same method under the same analysis. For $r < n$, the leading iteration factor is greater than one. Consequently, the smaller corrections must compensate for this increase before they improve the communication bound. The condition is

$$\frac{\ell_{\mathrm{a}} + (T-1)\ell_{\mathrm{b}}}{T\ell_{\mathrm{E}}} < \frac{8n}{\frac{n}{r}\left(\frac{n}{r}+1\right)a_{n,r}^2}. \tag{C.15}$$

This gives a sufficient condition for an asymptotically smaller stated upper bound. It is not a lower bound on the communication required by EF21-Muon.

Dependence on the correction period. The proposition fixes T. To display its finite-accuracy effect, define $D_T = \sigma[2n\Gamma_T + a_{n,r}(T-1)/\sqrt{N}]$. Corollary 1 can be written as $2\sqrt{A\mathcal{L}_1/K} + 2^{5/4}(ABS^2/K)^{1/4} + 3[(1+\sqrt{2})ABD_T/(2K)]^{1/3}$. Making each term at most $\varepsilon/3$ gives a sufficient count

$$K \geq \left\lceil \max\left\{ \frac{36A\mathcal{L}_1}{\varepsilon^2}, \frac{2592ABS^2}{\varepsilon^4}, \frac{729(1+\sqrt{2})ABD_T}{2\varepsilon^3} \right\} \right\rceil. \tag{C.16}$$

For fixed contraction parameters, $\Gamma_T = \mathcal{O}(T)$, $\mathcal{L}_1 = \mathcal{O}(T)$, and $D_T = \mathcal{O}(T)$. Hence $K_{\mathrm{A}} = \mathcal{O}(\varepsilon^{-4} + T\varepsilon^{-3} + T\varepsilon^{-2})$ is sufficient. A larger T reduces full transmissions but enlarges the terms due to momentum changes and compression errors. The finite-accuracy comparison is determined by (C.9), not the average transmission alone. The fixed-T limit does not cover a period growing with $1/\varepsilon$ without accounting for these terms.

Encoded bit counts and examples. For row-wise b-bit packing with one FP32 scale per row and an eight-byte length field, consider

$$\begin{aligned} \ell_{\mathrm{a}} &= 8\left\lceil \frac{b_{\mathrm{a}}mn}{8} \right\rceil + 32m + 64, \\ \ell_{\mathrm{b}} &= 8\left\lceil \frac{b_{\mathrm{b}}r^2}{8} \right\rceil + 32r + 64, \qquad \ell_{\mathrm{E}} = 16s(m+n+1). \end{aligned} \tag{C.17}$$

The EF21-Muon expression includes both singular-vector matrices and the s singular values, each transmitted in 16-bit format. With $n = 768$, $r = 691$, $s = 115$, $b_{\mathrm{a}} = b_{\mathrm{b}} = 4$, and $T = 3$, the leading iteration ratio is 1.23787. The average bit ratios are 0.73644 for $m = n$ and 0.63199 for $m = 4n$. Substituting in Proposition 1 gives

$$\lim_{\varepsilon \to 0^+} \frac{\mathcal{B}_{\mathrm{A}}(\varepsilon, 3)}{\mathcal{B}_{\mathrm{E}}(\varepsilon)} = \begin{cases} 0.91162, & m = n = 768, \\ 0.78233, & m = 3072,\ n = 768, \end{cases} \tag{C.18}$$

where the displayed ratios are rounded. The leading sufficient communication bounds decrease by approximately 8.8% and 21.8%, respectively. These are conditional single-matrix comparisons, not whole-model training measurements. They assume that the compressors satisfy contraction with fixed positive parameters. Bit width alone does not establish contraction, and the parameter s/n of exact rank-s truncation cannot automatically be assigned to its composition with Natural compression.

Comparison with the published EF21-Muon theorem. We additionally specialize Theorem 19, equation (15), of EF21-Muon (Gruntkowska et al., 2026) to one matrix and identity server compression. This theorem analyzes the norm-scaled update $X^{k+1} = X^k + \gamma_{\mathrm{E}}(\beta)\|G^k\|_* \,\mathrm{LMO}(G^k)$. It differs from the normalized update used in (C.3). Define $\widetilde{L}^2 = \frac{1}{N}\sum_{j=1}^N L_j^2$. Since $\|V\|_{\mathrm{op}} \le \|V\|_{\mathrm{F}} \le \sqrt{n}\|V\|_{\mathrm{op}}$, the norm-equivalence constants are 1 and $\sqrt{n}$, and the permitted step size is

$$\begin{aligned} \gamma_{\mathrm{E}}(\beta) &= \frac{1}{2L + 2\sqrt{\zeta_{\mathrm{E}}(\beta)}}, \\ \zeta_{\mathrm{E}}(\beta) &= n\left[\left(\frac{12}{\beta^2} + 48 + 24\beta\right) L^2 + \frac{144 + 756\beta^2 + 288\beta^3}{\alpha_{\mathrm{E}}^2}\widetilde{L}^2\right]. \end{aligned} \tag{C.19}$$

With exact initialization $G_j^0 = M_j^0 = \nabla f_j(X^0)$ and identical initial server and worker parameters, it gives

$$\begin{aligned} \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(X^k)\|_*^2 &\le \frac{4\Delta}{K\gamma_{\mathrm{E}}(\beta)} + F_{\mathrm{E}}^{\mathrm{pub}}(\beta), \\ F_{\mathrm{E}}^{\mathrm{pub}}(\beta) &= 24n\sigma^2\beta\left[\frac{1}{N} + \frac{(1-\alpha_{\mathrm{E}})\beta}{\alpha_{\mathrm{E}}} + \frac{12\beta^2}{\alpha_{\mathrm{E}}^2}\right]. \end{aligned} \tag{C.20}$$

To compare the same accuracy criterion, Jensen's inequality gives

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(X^k)\|_* \le \sqrt{\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla f(X^k)\|_*^2}. \tag{C.21}$$

We therefore require the bound in (C.20) to be at most ε^2. Squaring a first-moment bound would not give the reverse implication. The resulting sufficient counts and communication bound are

$$\begin{aligned} K_{\mathrm{E}}^{\mathrm{pub}}(\varepsilon) &= \min_{\substack{0<\beta\le 1 \\ F_{\mathrm{E}}^{\mathrm{pub}}(\beta)<\varepsilon^2}} \max\left\{1, \left\lceil \frac{4\Delta}{\gamma_{\mathrm{E}}(\beta)[\varepsilon^2 - F_{\mathrm{E}}^{\mathrm{pub}}(\beta)]} \right\rceil\right\}, \\ \mathcal{B}_{\mathrm{E}}^{\mathrm{pub}}(\varepsilon) &= U_{\mathrm{E}}(K_{\mathrm{E}}^{\mathrm{pub}}(\varepsilon)). \end{aligned} \tag{C.22}$$

Let $a = 16\sqrt{3n}\,\Delta L$ and $c = 24n\sigma^2/N$. Equations (C.19)–(C.20) imply

$$\frac{4\Delta}{\gamma_{\mathrm{E}}(\beta)} = \frac{a}{\beta} + \mathcal{O}(1), \qquad F_{\mathrm{E}}^{\mathrm{pub}}(\beta) = c\beta + \mathcal{O}(\beta^2). \tag{C.23}$$

Moreover, $4\Delta/\gamma_{\mathrm{E}}(\beta) \ge a/\beta$ and $F_{\mathrm{E}}^{\mathrm{pub}}(\beta) \ge c\beta$. Hence $y = c\beta/\varepsilon^2 \in (0,1)$ for admissible β, and

$$\frac{4\Delta}{\gamma_{\mathrm{E}}(\beta)[\varepsilon^2 - F_{\mathrm{E}}^{\mathrm{pub}}(\beta)]} \ge \frac{ac}{y(1-y)\varepsilon^4} \ge \frac{4ac}{\varepsilon^4}. \tag{C.24}$$

Choosing $\beta = \varepsilon^2/(2c)$ attains the same leading coefficient. Consequently,

$$K_{\mathrm{E}}^{\mathrm{pub}}(\varepsilon) \sim \frac{1536\sqrt{3}\,n^{3/2}\Delta L\sigma^2}{N\varepsilon^4} \sim 6\sqrt{3n}\,K_{\mathrm{E}}(\varepsilon). \tag{C.25}$$

Dividing (C.13) by this bound and using (C.10) gives

$$\lim_{\varepsilon \to 0^+} \frac{\mathcal{B}_{\mathrm{A}}(\varepsilon, T)}{\mathcal{B}_{\mathrm{E}}^{\mathrm{pub}}(\varepsilon)} = \frac{\overline{\ell}_{\mathrm{A}}(T)}{\ell_{\mathrm{E}}} \frac{\frac{n}{r}\left(\frac{n}{r}+1\right) a_{n,r}^2}{48\sqrt{3}\, n^{3/2}}. \tag{C.26}$$

For the two encodings in (C.18), this limit is approximately 0.003165 and 0.002716, respectively. The corresponding reductions in the leading sufficient communication bounds are approximately 99.68% and 99.73%. These compare theorems at the same gradient accuracy but retain their different update rules and proofs. At $n = 768$, the published baseline bound is asymptotically 288 times the normalized baseline bound. Even at $r = n$ and $T = 1$, the leading iteration ratio is $1/(6\sqrt{3n})$. The smaller coefficient therefore cannot be attributed entirely to subspace compression.

D Experimental Details

D.1 Model Configuration

We follow the NanoGPT-124M architecture used by EF21-Muon (Gruntkowska et al., 2026), as summarized in Table D.1. All methods use the same model initialization, data order, validation shard, and random seed (52). Table D.2 gives the training and optimizer hyperparameters shared across methods.

Table D.1: NanoGPT-124M model configuration.

Model	Layers	Heads	Hidden dim.	FFN dim.	Vocabulary
NanoGPT-124M	12	6	768	3,072	50,304

Table D.2: Training and optimizer hyperparameters shared across methods.

Hyperparameter	Value
Sequence length	1024
Global batch size	256 sequences
Momentum coefficient	0.9 ($\beta = 0.1$)
Weight decay	0
Hidden-layer norm	Spectral norm
Hidden-layer scale	50
Embedding and output norm	ℓ_∞ norm
Embedding and output scale	3000
Newton–Schulz iterations	5 (bfloat16)

For a hidden weight matrix of shape $d_{\mathrm{out}} \times d_{\mathrm{in}}$, the update scale is $50\sqrt{d_{\mathrm{out}}/d_{\mathrm{in}}}$ (Pethick et al., 2025). Newton–Schulz uses the polynomial $3.4445x - 4.7750x^3 + 2.0315x^5$ and normalizes its input matrix Y by $\|Y\|_{\mathrm{F}} + 10^{-7}$. All methods use the EF21-Muon update with RankK 15% followed by Natural compression for the shared embedding and output matrix. AdaSC-Muon rounds $0.9 \min\{m, n\}$ to the nearest integer for its projection rank, while RankK rounds $0.15 \min\{m, n\}$ down. The learning rates and compression settings are given in Section 4.1.

D.2 Communication Cost

We count cumulative worker-to-server communication per worker, following EF21-Muon (Gruntkowska et al., 2026). The count includes compressed tensors, the metadata needed to decode them, and one uncompressed initialization for AdaSC-Muon. Network protocol overhead and additional traffic from collective operations are excluded. We convert bytes to GiB by dividing by 2^{30}.

Bytes per transmission. For a $d_1 \times d_2$ matrix quantized to b bits, AdaSC-Muon sends packed codes, one FP32 scale per row, and an eight-byte length field. The number of bytes is

$$P_b(d_1, d_2) = \underbrace{\left\lceil \frac{b d_1 d_2}{8} \right\rceil}_{\text{quantized values}} + \underbrace{4d_1}_{\text{row scales}} + \underbrace{8}_{\text{length field}}. \tag{D.1}$$

For each $m \times n$ momentum matrix, a full-momentum correction sends $P_{b_a}(m, n)$ bytes. A projected correction sends $P_{b_b}(r, r)$ bytes for AdaSC-Muon and $P_{b_b}(1, r)$ bytes for the diagonal option, whose r values share one scale. EF21-Muon sends the two singular-vector matrices and the singular values of a rank-s approximation. With Natural compression stored in 16-bit format, its cost is $2s(m+n+1)$ bytes, where $s = \lfloor 0.15 \min\{m, n\} \rfloor$.

Our model has 72 hidden weight matrices. Let (m_ℓ, n_ℓ) be the shape of matrix ℓ. At $r = 691$ and $b_a = b_b = 4$, the total bytes for a full correction, a projected correction, a diagonal correction, and initialization are, respectively,

$$P_{\text{full}} = \sum_{\ell=1}^{72} P_4(m_\ell, n_\ell) = 42{,}799{,}680, \qquad P_{\text{proj}} = 72P_4(691, 691) = 17{,}388{,}936,$$

$$P_{\text{diag}} = 72P_4(1, 691) = 25{,}776, \qquad P_0 = 4\sum_{\ell=1}^{72} m_\ell n_\ell = 339{,}738{,}624. \tag{D.2}$$

All methods also send $P_e = 2 \cdot 115 \cdot (50{,}304 + 768 + 1) = 11{,}746{,}790$ bytes per step for the shared embedding and output matrix. Summing the RankK cost over this matrix and the 72 hidden matrices gives $P_{\text{EF}} = 49{,}917{,}590$ bytes per step for EF21-Muon.

Cumulative communication. For AdaSC-Muon, initialization provides the first hidden-layer update. A run with $K \geq 1$ completed updates therefore sends $K - 1$ subsequent hidden-layer corrections. Of these, a full correction is sent every T steps, giving

$$n_{\text{full}} = \left\lfloor \frac{K-1}{T} \right\rfloor, \qquad n_{\text{proj}} = K - 1 - n_{\text{full}}. \tag{D.3}$$

The cumulative bytes uploaded by one worker are

$$\begin{aligned} B_K^{\text{AdaSC}} &= P_0 + n_{\text{full}} P_{\text{full}} + n_{\text{proj}} P_{\text{proj}} + K P_e, \\ B_K^{\text{EF}} &= K P_{\text{EF}}. \end{aligned} \tag{D.4}$$

For the diagonal option, replace P_{proj} with P_{diag}. The corresponding communication costs are $B_K^{\text{AdaSC}}/2^{30}$ and $B_K^{\text{EF}}/2^{30}$ GiB per worker.

E Additional Experimental Results

E.1 Fixed Wall-Clock Time

We compare the methods at 8 and 17 hours, matching the reported wall-clock times of AdaSC-Muon and EF21-Muon for 5B-token training, respectively. These comparisons evaluate whether the improvements at fixed tokens also yield better validation performance at the same wall-clock time. Table 2 reports the results, and Figures E.1 and E.2 show the training curves.

8 hours. AdaSC-Muon achieves a minimum validation loss of 3.274, compared with 3.376 for EF21-Muon, an improvement of 0.102. The number of training tokens increases from 2.28B to 5.00B, a 119.3% increase within the same reported wall-clock time. AdaSC-Muon (diagonal) completes 5.03B tokens and achieves a minimum validation loss of 3.330, improving upon EF21-Muon by 0.046. Although the two AdaSC-Muon options complete similar token counts, the full option achieves a loss 0.056 lower than the diagonal option. Its communication cost is higher at 668.3 GiB per worker, compared with 465.1 GiB for the diagonal option. Therefore, full projected compression provides the best validation performance in this comparison, while diagonal compression reduces communication relative to the full option.

17 hours. AdaSC-Muon retains its advantage when the comparison uses the wall-clock time required by EF21-Muon for 5B-token training. It achieves a minimum validation loss of 3.203, compared with 3.278 for EF21-Muon, an improvement of 0.075. The number of training tokens increases from 5.00B to 11.00B, a 120.0% increase. The diagonal option completes 11.07B tokens and achieves a minimum validation loss of 3.246. This loss is 0.032 lower than EF21-Muon but remains 0.043 higher than AdaSC-Muon. AdaSC-Muon therefore achieves the lowest minimum validation loss at both wall-clock times, with the full option consistently outperforming the diagonal option at similar token counts.

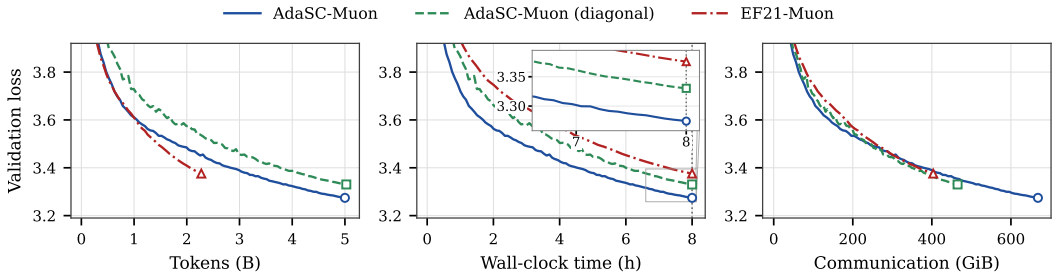

Figure E.1: Validation loss with 8 hours of wall-clock time.

The fixed-time comparisons allow each method to complete a different number of training steps. EF21-Muon has the lowest cumulative communication cost at both 8 and 17 hours, using 403.5 and 886.7 GiB per worker. At 17 hours, AdaSC-Muon and its diagonal option use 1469.9 and 1024.1 GiB, respectively. Their lower communication cost per step allows communication savings at fixed tokens, but their larger number of steps increases total communication in the fixed-time comparisons.

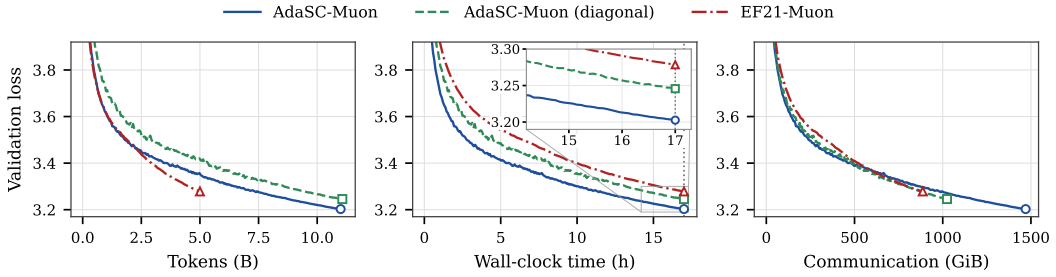

Figure E.2: Validation loss with 17 hours of wall-clock time.

E.2 Target Validation Loss

We compare the tokens, communication cost, and wall-clock time needed to reach validation loss 3.350. This comparison evaluates the resource savings of AdaSC-Muon at the same validation performance. Table 2 reports the results, and Figure E.3 shows the corresponding training curves.

AdaSC-Muon reaches the target with substantially lower communication cost and wall-clock time than EF21-Muon. Communication cost decreases from 644.9 to 472.2 GiB per worker, a 26.8% reduction, and wall-clock time decreases from 13 to 6 hours, a 53.8% reduction. AdaSC-Muon also requires fewer training steps, 13,475 compared with 13,871 for EF21-Muon, with the token count decreasing from 3.64B to 3.53B. These results show that the communication and wall-clock time savings are retained when the methods are compared at the same target loss.

AdaSC-Muon (diagonal) reduces communication cost further to 424.4 GiB per worker, 34.2% below EF21-Muon. It reaches the target in 7 hours, reducing wall-clock time by 46.2% relative to EF21-Muon. However, the diagonal option requires 17,496 steps and 4.59B tokens, compared with 13,475 steps and 3.53B tokens for AdaSC-Muon. Its smaller projected transmissions reduce total communication despite the additional steps, while the larger number of steps leads to a longer wall-clock time than the full option. Therefore, AdaSC-Muon reaches the target with the fewest steps and shortest wall-clock time, while the diagonal option achieves the lowest communication cost.

E.3 Diagonal Transmission

With exactly paired singular bases, the server's projected tracker is diagonal immediately after a full correction. Diagonal transmission preserves this form until the bases change. Writing $\widehat{B}^k =$

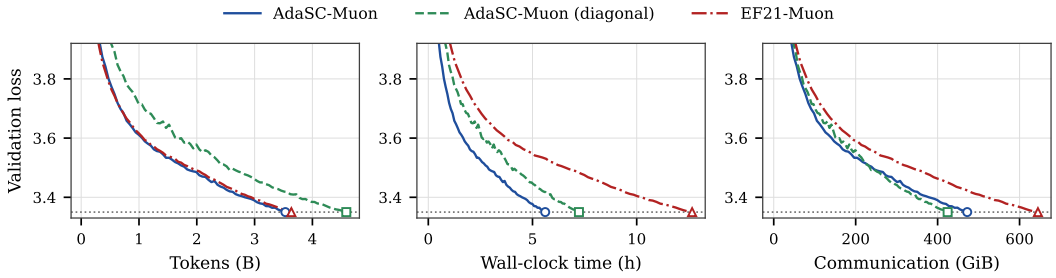

Figure E.3: Training curves up to target validation loss 3.350.

$\mathrm{Diag}(d^k)$, its exact update is

$$O^k = -Q_L^k \,\mathrm{Diag}(\mathrm{sign}(d^k))(Q_R^k)^\top. \tag{E.1}$$

Within a period, the signs can change, while the update's left and right directions remain fixed by the selected basis pairs. Finite Newton–Schulz iterations can also change the diagonal magnitudes.