Error Feedback for Muon and Friends

Kaja Gruntkowska Alexander Gaponov Zhirayr Tovmasyan Peter Richtárik

King Abdullah University of Science and Technology (KAUST)

arXiv:2510.00643

*EF21-Muon brings the power of **Muon/Scion/Gluon** to distributed training, offering compressed communication-efficient non-Euclidean LMO updates with convergence guarantees under generalized smoothness and delivering significant practical communication savings.*

Optimization Problem

We consider the nonconvex *distributed* optimization problem:

$$\min_{X\in\mathcal{S}}\left\{f(X)=\frac{1}{n}\sum_{j=1}^{n}f_j(X)\right\}, \qquad f_j(X)=\mathbb{E}_{\xi_j\sim\mathcal{D}_j}\left[f_j(X;\xi_j)\right]$$

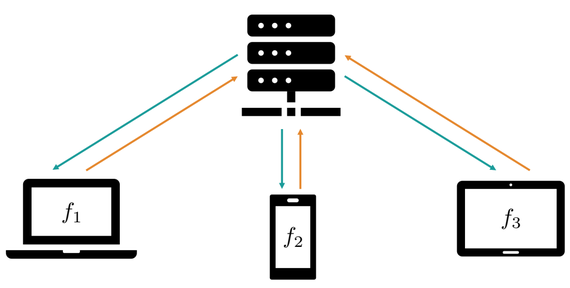
- $f_j(X)$ – loss of the model X on the data stored on worker j
- **Goal:** find $\hat{X}$ such that $\mathbb{E}\left[\|\nabla f(\hat{X})\|_\star^2\right]\leq\varepsilon$

How to Distribute Muon?

The basic update of **Muon** is

$$\begin{aligned}X^{k+1}&=X^k-t^kU^k(V^k)^T\\&=X^k+\mathsf{LMO}_{\mathcal{B}^{2\to2}(0,t^k)}(M^k),\end{aligned}$$

where $M^k=U^k\Sigma^k(V^k)^T$.



Contractive compressor

$$\mathbb{E}\left[\|\mathcal{C}(X)-X\|^2\right]\leq(1-\alpha)\|X\|^2 \qquad \forall X\in\mathcal{S}$$

Assumptions

A1: There exist $f^\star\in\mathbb{R}$ such that $f(X)\geq f^\star$ for all $X\in\mathcal{S}$.

A2: There exist $f_j^\star\in\mathbb{R}$ such that $f_j(X)\geq f_j^\star$ for all $X\in\mathcal{S}$.

A3: f is L-smooth, i.e.,

$$\|\nabla f(X)-\nabla f(Y)\|_\star\leq L\|X-Y\|$$

for all $X,Y\in\mathcal{S}$. Moreover, the functions f_j are L_j–smooth for all $j\in[n]$. We define $\tilde{L}^2=\frac{1}{n}\sum_{j=1}^n L_j^2$.

A4: $f:\mathcal{S}\mapsto\mathbb{R}$ is (L^0,L^1)–smooth, i.e.,

$$\|\nabla f(X)-\nabla f(Y)\|_\star\leq\left(L^0+L^1\|\nabla f(X)\|_\star\right)\|X-Y\|$$

for all $X,Y\in\mathcal{S}$. Moreover, the functions f_j, $j\in[n]$, are (L_j^0,L_j^1)–smooth. We define $L_{\max}^1=\max_{j\in[n]}L_j^1$ and $\bar{L}^0=\frac{1}{n}\sum_{j=1}^n L_j^0$.

A5: $\mathbb{E}_{\xi_j\sim\mathcal{D}_j}\left[\nabla f_j(X;\xi_j)\right]=\nabla f_j(X)$ and $\exists\,\sigma\geq0$ such that $\mathbb{E}_{\xi_j\sim\mathcal{D}_j}\left[\left\|\nabla f_j(X;\xi_j)-\nabla f_j(X)\right\|_2^2\right]\leq\sigma^2$ for all $X\in\mathcal{S}$.

Algorithm 1 EF21-Muon

Parameters: radii $t^k>0$; momentum parameter $\beta\in(0,1]$; initial iterate $X^0\in\mathcal{S}$ (stored on the server); initial iterate shift $W^0=X^0$ (stored on the server and the workers); initial gradient estimators G_j^0 (stored on the workers); $G^0=\frac{1}{n}\sum_{j=1}^n G_j^0$ (stored on the server); initial momentum M_j^0 (stored on the workers); worker compressors $\mathcal{C}_j^k$; server compressors $\mathcal{C}^k$

for $k=0,1,\ldots,K-1$ **do**

  $X^{k+1}=\mathsf{LMO}_{\mathcal{B}(X^k,t^k)}(G^k)$ — Take LMO-type step

  $S^k=\mathcal{C}^k(X^{k+1}-W^k)$ — Compress shifted model on the server

  $W^{k+1}=W^k+S^k$ — Update model shift

  Broadcast S^k to all workers

  for $j=1,\ldots,n$ **in parallel do**

    $W^{k+1}=W^k+S^k$ — Update model shift

    $M_j^{k+1}=(1-\beta)M_j^k+\beta\nabla f_j(W^{k+1};\xi_j^{k+1})$ — Compute momentum

    $R_j^{k+1}=\mathcal{C}_j^k(M_j^{k+1}-G_j^k)$ — Compress shifted gradient

    $G_j^{k+1}=G_j^k+R_j^{k+1}$

    Broadcast R_j^{k+1} to the server

  end for

  $G^{k+1}=\frac{1}{n}\sum_{j=1}^n G_j^{k+1}=G^k+\frac{1}{n}\sum_{j=1}^n R_j^{k+1}$ — Compute gradient estimator

end for

Convergence under L-smoothness

Let Assumptions A1, A3 and A5 hold. Let $\{X^k\}_{k=0}^{K-1}$, $K\geq1$, be the iterates of **EF21-Muon** initialized with $X^0=W^0$, $G_j^0=M_j^0=\nabla f_j(X^0;\xi_j^0)$, $j\in[n]$, and run with $\mathcal{C}^k\in\mathbb{B}(\alpha_P)$, $\mathcal{C}_j^k\in\mathbb{B}_2(\alpha_D)$, $\beta=\min\left\{1,\left(\frac{\delta^0Ln}{\underline{\rho}^2\sigma^2K}\right)^{1/2},\left(\frac{\delta^0L\alpha_D}{\underline{\rho}^2\sigma^2K}\right)^{1/3},\left(\frac{\delta^0L\alpha_D^2}{\underline{\rho}^2\sigma^2K}\right)^{1/4}\right\}$ and $\gamma=\left(2\sqrt{\zeta}+2L\right)^{-1}$, where $\zeta=\frac{\bar{\rho}^2}{\underline{\rho}^2}\left(\frac{12}{\beta^2}L^2+\frac{24(\beta+2)}{\alpha_P^2}L^2+\frac{36(\beta^2+4)}{\alpha_D^2}\tilde{L}^2+\frac{144\beta^2(2\beta+5)}{\alpha_P^2\alpha_D^2}\tilde{L}^2\right)$. Then

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla f(X^k)\|_\star^2\right]=\mathcal{O}\left(\frac{\delta^0\bar{\rho}^2\tilde{L}^0}{\underline{\rho}^2\alpha_P\alpha_DK}+\left(\frac{\delta^0\bar{\rho}^4\sigma^2L}{\underline{\rho}^2nK}\right)^{1/2}+\left(\frac{\delta^0\bar{\rho}^3\sigma L}{\underline{\rho}^2\sqrt{\alpha_D}K}\right)^{2/3}+\left(\frac{\delta^0\underline{\rho}^{8/3}\sigma^{2/3}L}{\bar{\rho}^2\alpha_D^{2/3}K}\right)^{3/4}\right)$$

where $\delta^0=f(X^0)-f^\star$.

Convergence under (L^0,L^1)-smoothness

Let Assumptions A1, A2, A4 and A5 hold. Let $\{X^k\}_{k=0}^{K-1}$, $K\geq1$, be the iterates of **EF21-Muon** initialized with $M_j^0=\nabla f_j(X^0;\xi_j^0)$, $G_j^0=\mathcal{C}_j^0(\nabla f_j(X^0;\xi_j^0))$, $j\in[n]$, and run with $\mathcal{C}^k\equiv\mathcal{I}$, $\mathcal{C}_j^k\in\mathbb{B}_2(\alpha_D)$, $\beta=1/(K+1)^{1/2}$ and $0\leq t^k\equiv t=\eta/(K+1)^{3/4}$, where $\eta^2\leq\min\left\{\frac{(K+1)^{1/2}}{6(L^1)^2},\frac{(K+1)^{1/2}(1-\sqrt{1-\alpha_D})\underline{\rho}}{24\sqrt{1-\alpha_D}\bar{\rho}(L_{\max}^1)^2},\frac{\underline{\rho}}{24\bar{\rho}(L_{\max}^1)^2},1\right\}$. Then

$$\begin{aligned}&\min_{k=0,\ldots,K}\mathbb{E}\left[\left\|\nabla f(X^k)\right\|_\star\right]\\&\leq\frac{3\left(f(X^0)-f^\star\right)}{\eta(K+1)^{1/4}}+\frac{\eta L^0}{(K+1)^{3/4}}+\frac{16\sqrt{1-\alpha_D}\bar{\rho}\sigma}{(1-\sqrt{1-\alpha_D})(K+1)^{1/2}}+\frac{8\bar{\rho}\sigma}{\sqrt{n}(K+1)^{1/4}}\\&+\frac{\eta\bar{\rho}}{\underline{\rho}}\left(\frac{8}{(K+1)^{1/4}}+\frac{8\sqrt{1-\alpha_D}}{(1-\sqrt{1-\alpha_D})(K+1)^{3/4}}\right)\left(\frac{1}{n}\sum_{j=1}^n(L_j^1)^2\left(f^\star-f_j^\star\right)+\bar{L}^0\right).\end{aligned}$$

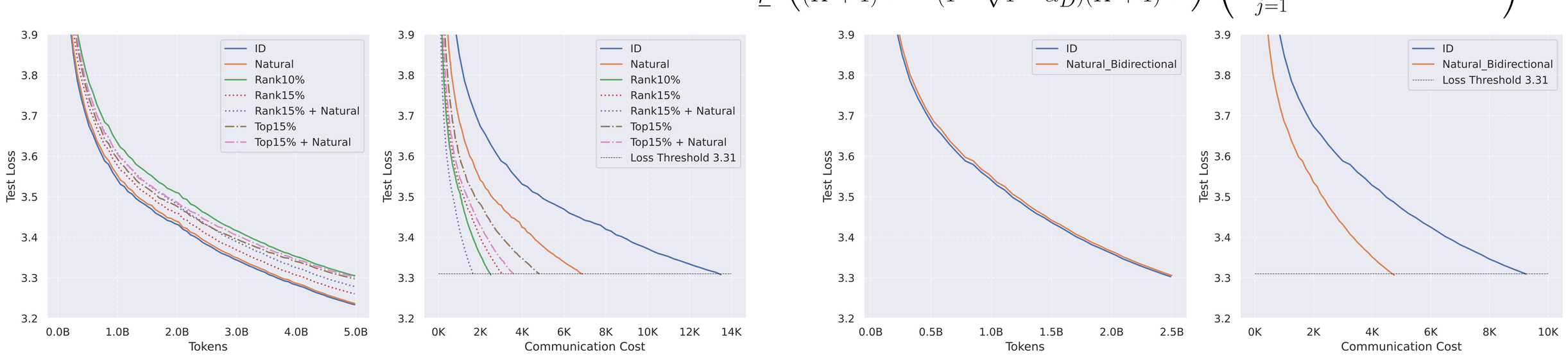


Figure 1. **Training NanoGPT-124M on FineWeb10B.** Panels 1 and 3: Test loss vs. # of tokens processed. Panels 2 and 4: Test loss vs. # of bytes sent to the server from each worker normalized by model size to reach test loss 3.31. Rank$X$%/Top$X$% = Rank$K$/Top$K$ compressor with sparsification level X%; ID = no compression.

References

[1] K. Jordan, Y. Jin, V. Boza, J. You, F. Cesista, L. Newhouse, and J. Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/

[2] A. Riabinin, E. Shulgin, K. Gruntkowska, and P. Richtárik. Gluon: Making Muon & Scion great again! *arXiv:2505.13416*, 2025.

[3] T. Pethick, W. Xie, K. Antonakopoulos, Z. Zhu, A. Silveti-Falls, and V. Cevher. Training deep learning models with norm-constrained LMOs. *arXiv:2502.07529*, 2025.

[4] P. Richtárik, I. Sokolov, and I. Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. *NeurIPS*, 2021.

[5] K. Gruntkowska, A. Tyurin, and P. Richtárik. EF21-P and friends: Improved theoretical communication complexity for distributed optimization with bidirectional compression. *ICML*, PMLR, pp. 11761–11807, 2023.