The Ky Fan Norms and Beyond: Dual Norms and Combinations for Matrix Optimization

Alexey Kravatskiy[1] Ivan Kozyrev[1] Nikolai Kozlov[1] Alexander Vinogradov[1] Daniil Merkulov[1,2,3,4] Ivan Oseledets[5,2]

[1]MIPT

[2]Skoltech

[3]HSE

[4]SberAI

[5]AIRI

October 9, 2025

Moving beyond the spectral norm and Muon

Objective: $\min_{\boldsymbol{X} \in \mathbb{R}^{m \times n}} f(\boldsymbol{X})$

❓ What will happen if we change the spectral norm in the derivation of Muon?

F-Fanions: Muon, Neon, NSGD, Dion without EF, and so much more

From $\|\cdot\|_{\mathrm{op}}$ and Muon:

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta U V^\top$$

To $\|\cdot\|_{\mathrm{KF-k}}^\dagger$ and a general F-Fanion:

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta \left(\alpha \sum_{i=1}^{k} u_i v_i^\top + (1-\alpha) \frac{\boldsymbol{M}^t}{\|\boldsymbol{M}^t\|_{\mathrm{F}}} \right), \alpha \in [0,1]$$

Linear Minimization Oracle (LMO) and Trust Region

Let us equip $\mathbb{R}^{m\times n}$ with a norm $\|\cdot\|$. Its dual is $\|\boldsymbol{X}\|^\dagger = \sup_{\|\boldsymbol{X}'\|\leq 1}\langle \boldsymbol{X}, \boldsymbol{X}'\rangle$. $\langle\cdot,\cdot\rangle$ is a Frobenius product.

Both LMO and Trust Region lead to the update

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta \arg\max_{\boldsymbol{X}\in\mathcal{B}_1}\langle \boldsymbol{M}^t, \boldsymbol{X}\rangle = \boldsymbol{X}^t - \eta\{\Delta \in \mathcal{B}_1 \mid \langle \boldsymbol{M}^t, \Delta\rangle = \|\boldsymbol{M}^t\|^\dagger\}$$

Recipe: It means we seek Δ from the 1-norm ball that delivers $\langle \boldsymbol{M}^t, \Delta\rangle = \|\boldsymbol{M}^t\|^\dagger$. We will often return to the SVD $\boldsymbol{M}^t = \boldsymbol{U}\Sigma\boldsymbol{V}^\top$.

Frobenius $\|\boldsymbol{M}^k\|_{\mathrm{F}}$ and Normalized SGD

Deriving NSGD by Recipe

$\|\boldsymbol{M}^t\|_{\mathrm{F}}^{\dagger} = \|\boldsymbol{M}^t\|_{\mathrm{F}}$, and $\Delta = \frac{\boldsymbol{M}^t}{\|\boldsymbol{M}^t\|_{\mathrm{F}}}$ with $\|\Delta\|_{\mathrm{F}} = 1$ delivers it. Hence,

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta \frac{\boldsymbol{M}^t}{\|\boldsymbol{M}^t\|_{\mathrm{F}}}$$

Spectral $\|M^k\|_{\mathrm{op}}$ and Muon. Nuclear $\|M^k\|_{\mathrm{nuc}}$ and Neon.

Deriving Muon by Recipe

$\|\boldsymbol{M}^t\|_{\mathrm{op}}^\dagger = \|\boldsymbol{M}^t\|_{\mathrm{nuc}}$, and $\Delta = \boldsymbol{U}\boldsymbol{V}^\top$ delivers it: $\|\Delta\|_{\mathrm{op}} = 1$ and

$$\langle \boldsymbol{U}\Sigma\boldsymbol{V}^\top, \boldsymbol{U}\boldsymbol{V}^\top \rangle = \mathrm{tr}(\boldsymbol{V}\Sigma\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{V}^\top) = \mathrm{tr}\,\Sigma = \|\boldsymbol{M}^t\|_{\mathrm{nuc}}$$

Hence,

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta\boldsymbol{U}\boldsymbol{V}^\top$$

Deriving Neon by Recipe

$\|\boldsymbol{M}^t\|_{\mathrm{nuc}}^\dagger = \|\boldsymbol{M}^t\|_{\mathrm{op}}$, and $\Delta = u_1 v_1^\top$ delivers it: $\|\Delta\|_{\mathrm{nuc}} = 1$ and

$$\langle \boldsymbol{U}\Sigma\boldsymbol{V}^\top, u_1 v_1^\top \rangle = \mathrm{tr}(\boldsymbol{V}\Sigma\boldsymbol{U}^\top u_1 v_1^\top) = \mathrm{tr}\,\mathrm{diag}(\sigma_1, 0, \ldots, 0) = \sigma_1 = \|\boldsymbol{M}^t\|_{\mathrm{op}}$$

Hence,

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta u_1 v_1^\top$$

Of Matrix and Vector Algorithms

Table: lmo optimizers in Schatten S_p norms and in l_p norms.

Method	lmo constraint set $\mathcal{D}$	lmo	Reference
Normalized SGD	l_2-ball, S_2-ball	$-\eta \frac{g}{\|g\|_2} = -\eta \frac{g}{\|g\|_F}$	(Hazan et al., 2015)
Momentum Normalized SGD	Ball in l_2, or Ball in S_2	$-\eta \frac{g}{\|g\|_2} = -\eta \frac{g}{\|g\|_F}$	(Cutkosky et al., 2020)
SignSGD	Ball in Max-norm l_∞	$-\eta \operatorname{sign}(g)$	(Bernstein et al., 2018, Thm. 1)
Signum	Ball in Max-norm l_∞	$-\eta \operatorname{sign}(g)$	(Bernstein et al., 2018, Thm. 3)
Muon	Ball in Spectral S_∞	$-\eta U V^\top$	(Jordan et al., 2024b)
Gauss-Southwell Coordinate Descent	Ball in l_1	$-\eta \{i : g_i \geq g_k \forall k\}$	(Shi et al., 2016, p.19)
Neon	Ball in Nuclear S_1	$-\eta u_1 v_1^\top$	This work

Understanding Dion by Thomas Pethick

Without momentum, Dion is simplified to

$$\Delta \leftarrow g + e$$
$$e \leftarrow \Delta - \sum_{i=1}^{r} \sigma_i u_i v_i^\top$$
$$x \leftarrow x - \gamma \sum_{i=1}^{r} u_i v_i^\top$$

where $\gamma > 0$ and $\sum_{i=1}^{r} \sigma_i u_i v_i^\top$ is the rank-r truncated SVD of $\Delta \in \mathbb{R}^{m \times n}$.

❓ What will we get if we try $\|\boldsymbol{M}\|_{\mathrm{KF-k}} := \sum_{i=1}^{r} \sigma_i$?

Ky Fan k-rank $\|\boldsymbol{M}^k\|^{\dagger}_{\mathrm{KF-k}}$ and Fanions

Deriving Fanions by Recipe

$\|\boldsymbol{M}^t\|^{\dagger\dagger}_{\mathrm{KF-k}} = \|\boldsymbol{M}^t\|_{\mathrm{KF-k}}$, and $\Delta = \sum_{i=1}^{k} u_i v_i^\top$ with $\|\Delta\|^{\dagger}_{\mathrm{KF-k}} = \max\{\frac{1}{k}\|\Delta\|_{\mathrm{nuc}}, \|\Delta\|_{\mathrm{op}}\} = \max\{\frac{1}{k}k, 1\} = 1$ delivers it:

$$\langle \boldsymbol{M}^t, \Delta \rangle = \langle \boldsymbol{U\Sigma V}^\top, \sum_{i=1}^{k} u_i v_i^\top \rangle = \sum_{i,j=1}^{r,k} \langle u_i \sigma_i v_i^\top, u_j v_j^\top \rangle = \sum_{i=1}^{k} \sigma_i = \|\boldsymbol{M}^t\|_{\mathrm{KF-k}}$$

Hence,

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta \sum_{i=1}^{k} u_i v_i^\top$$

Computing the k-rank update: Lanczos method

Method	rtol	k	time (s)
Power Iterations	0.01	1	7.7
SVDS (TRLan)	0.01	1	**0.18**
PCA Low Rank (RSVD)	0.01	1	1.15
SVDS (TRLan)	0.01	10	**0.47**
PCA Low Rank (RSVD)	0.01	10	19.4
SVDS (TRLan)	0.01	100	**1.96**
PCA Low Rank (RSVD)	0.01	100	170

Comparison on a 5000×5000 matrix; rtol is the relative Frobenius error of $\sum_{i=1}^{k} u_i \sigma_i v_i^\top$

Frobeniusize the norms!

❗ $\|\cdot\|_{\mathrm{F}}^{\dagger} = \|\cdot\|_{\mathrm{F}}$ is not a Ky Fan norm, so NSGD is not a Fanion.

Let us consider the convex combination of the Ky Fan rank-k norm and the Frobenius norm, which we call F-KF-k-norm:

$$\|\cdot\|_{\mathrm{F-KF-k}} = \alpha\|\cdot\|_{\mathrm{KF-k}} + (1-\alpha)\|\cdot\|_{\mathrm{F}}, \alpha \in [0,1]$$

Balls of Duals to Convex Combinations of Norms

Lemma

Let $\|\cdot\|_{(1)}$ *and* $\|\cdot\|_{(2)}$ *be norms on a finite-dimensional Euclidean space, and let* $\alpha, \beta \geq 0$*. Define*

$$\|x\| := \alpha\|x\|_{(1)} + \beta\|x\|_{(2)}.$$

Then the dual unit ball of $\|\cdot\|$ *satisfies*

$$B_{\|\cdot\|^\dagger} = \alpha B_{\|\cdot\|^\dagger_{(1)}} + \beta B_{\|\cdot\|^\dagger_{(2)}},$$

where $+$ *denotes the Minkowski sum and* $B_{\|\cdot\|^\dagger_{(i)}}$ *is the unit ball of the dual norm* $\|\cdot\|^\dagger_{(i)}$*.*

$\|\boldsymbol{M}^k\|^{\dagger}_{\mathrm{F-KF-k}}$ and F-Fanions

Deriving F-Fanions by Recipe

$\|\boldsymbol{M}^t\|^{\dagger\dagger}_{\mathrm{F-KF-k}} = \|\boldsymbol{M}^t\|_{\mathrm{F-KF-k}}$, and $\Delta = \alpha \sum_{i=1}^k u_i v_i^\top + (1-\alpha)\frac{\boldsymbol{M}^k}{\|\boldsymbol{M}^k\|_{\mathrm{F}}}$ delivers it:

1. $\|\Delta\|^{\dagger}_{\mathrm{F-KF-k}} \leq 1$ because $\sum_{i=1}^k u_i v_i^\top$ lies in $B_{\|\cdot\|_{\mathrm{KF-k}}}$ and $\frac{\boldsymbol{M}^k}{\|\boldsymbol{M}^k\|_{\mathrm{F}}}$ lies in $B_{\|\cdot\|_{\mathrm{F}}}$, so Δ lies in the Minkowski sum
2. $\langle \boldsymbol{M}^t, \Delta\rangle = \alpha\langle \boldsymbol{U\Sigma V}^\top, \sum_{i=1}^k u_i v_i^\top\rangle + (1-\alpha)\langle \boldsymbol{M}^t, \frac{\boldsymbol{M}^t}{\|\boldsymbol{M}^t\|_{\mathrm{F}}}\rangle =$ $\alpha \sum_{i=1}^k \sigma_i + (1-\alpha)\|\boldsymbol{M}^t\|_{\mathrm{F}} = \|\boldsymbol{M}^t\|_{\mathrm{F-KF-k}}$

Hence,

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta\Big(\alpha \sum_{i=1}^k u_i v_i^\top + (1-\alpha)\frac{\boldsymbol{M}^t}{\|\boldsymbol{M}^t\|_{\mathrm{F}}}\Big)$$

F-Muon and F-Neon

Convex combinations with Frobenius norm:

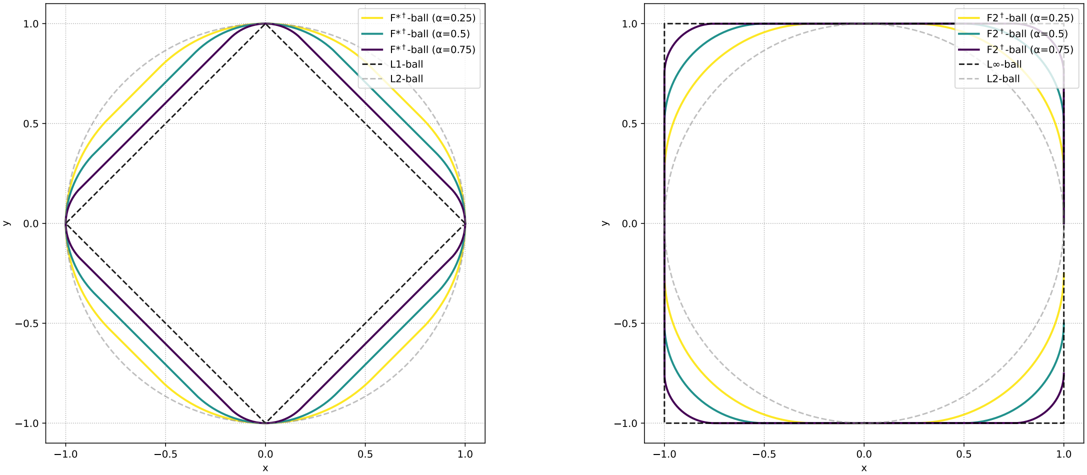
$$\|\boldsymbol{X}\|_{F*} = \alpha\|\boldsymbol{X}\|_* + (1-\alpha)\|\boldsymbol{X}\|_F, \qquad \|\boldsymbol{X}\|_{F2} = \alpha\|\boldsymbol{X}\|_{op} + (1-\alpha)\|\boldsymbol{X}\|_F.$$

Dual-induced updates:

$$\text{F-Muon:}\;\; \boldsymbol{X}^{k+1} = \boldsymbol{X}^k - \eta\Big(\alpha\,\boldsymbol{U}\boldsymbol{V}^\top + (1-\alpha)\frac{\boldsymbol{M}^k}{\|\boldsymbol{M}^k\|_F}\Big),$$

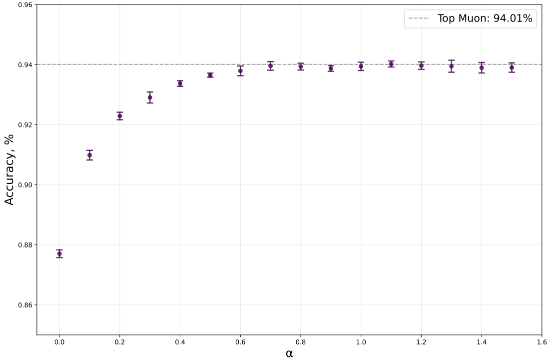
$$\text{F-Neon:}\;\; \boldsymbol{X}^{k+1} = \boldsymbol{X}^k - \eta\Big(\alpha\,u_1 v_1^\top + (1-\alpha)\frac{\boldsymbol{M}^k}{\|\boldsymbol{M}^k\|_F}\Big).$$

Geometric intuition: Dual balls

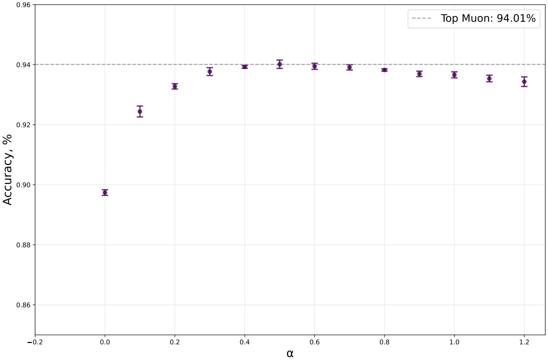


Dual balls for F-Muon and F-Neon across α in a 2D singular-value space

F-Muon on CIFAR-10 airbench



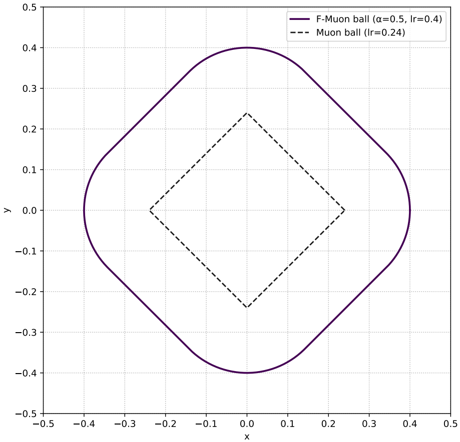
Params tuned for Muon (Keller, 2023)



Params tuned for F-Muon

F-Muon with $\alpha \approx 0.5$ matches Muon after tuning

LMO balls on CIFAR-10



LMO balls for learning rates from the experiment

❓ How should we interpret cases with $\alpha > 1$?

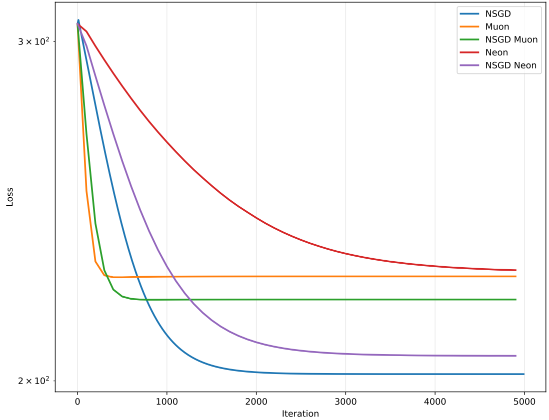
NanoGPT speedrun

- Setup from Jordan et al. (2024a): 1750 iterations and cross-entropy loss lower than 3.28
- Muon: `lr=0.05, momentum=0.95`, 3.279
- F-Muon with $\alpha = 0.5$: `lr=0.07, momentum=0.95`, 3.281

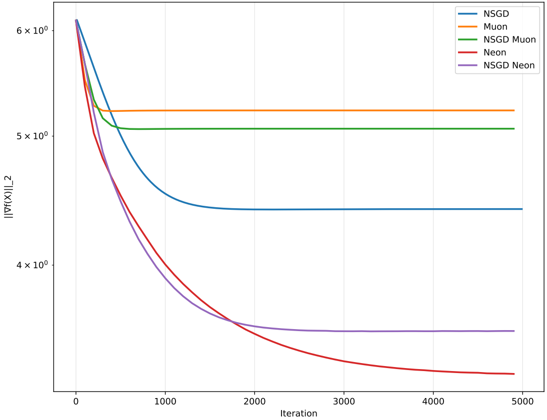
Trust-region bounds (L-smooth)

- Deterministic: choose $\eta = \mathcal{O}(\varepsilon/L)$, $K = \mathcal{O}(L\Delta_0/\varepsilon^2)$ (Kovalev, 2025).
- Stochastic (with variance σ^2, link between norms ρ): detailed schedules for η, α, K.
- Asymptotically, norms are proportional for large random matrices (Marchenko–Pastur), aligning guarantees across choices.

Experiment: Random linear least squares

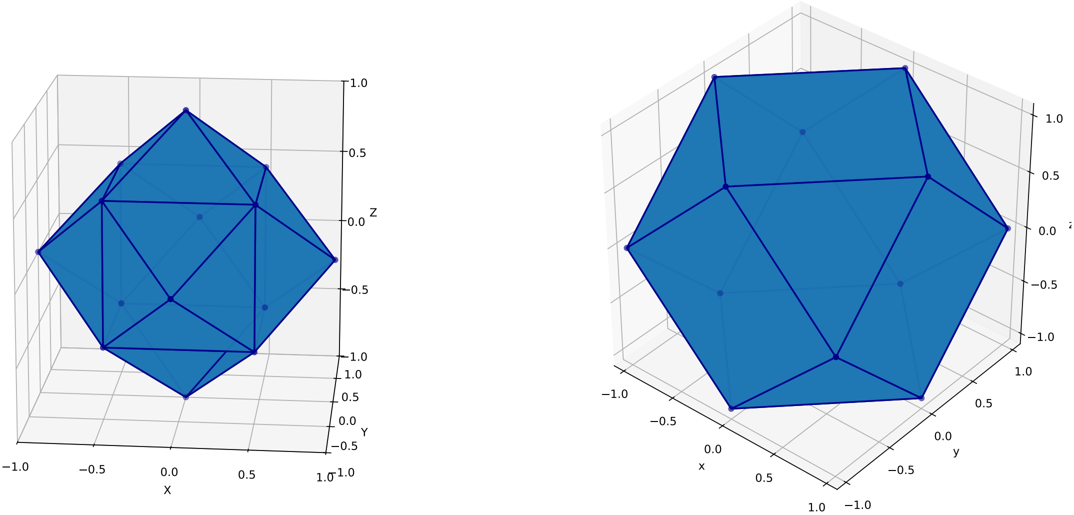


Loss vs iteration



Spectral norm of gradient vs iteration

Setup informed by Kovalev (2025).



−1.0
−0.5
0.0
0.5
1.0
X
Y
Z
x
y

Conclusions

- Fanions unify Muon/Neon via Ky Fan norms; F-Fanions add robustness via Frobenius mixing.
- TRLan enables efficient k-rank updates; practical speed benefits over power/RSVD.
- Empirically, F-Muon remains competitive on CIFAR-10 and NanoGPT.

References I

Ahn, Kwangjun and Byron Xu (2025). *Dion: A Communication-Efficient Optimizer for Large Models*. arXiv: 2504.05295 [cs.LG]. URL: https://arxiv.org/abs/2504.05295.

Amsel, Noah et al. (2025). "The Polar Express: Optimal Matrix Sign Methods and Their Application to the Muon Algorithm". In: *arXiv preprint arXiv:2505.16932*.

Bernstein, Jeremy (2025). *Deriving Muon*. URL: https://jeremybernste.in/writing/deriving-muon.

Bernstein, Jeremy et al. (2018). "signSGD: Compressed optimisation for non-convex problems". In: *International conference on machine learning*. PMLR, pp. 560–569.

Bhatia, Rajendra (2013). *Matrix analysis*. Vol. 169. Springer Science & Business Media.

Cutkosky, Ashok, Harsh Mehta, and Francesco Orabona (2020). "Momentum-based variance reduction in nonconvex SGD". In: *Advances in Neural Information Processing Systems*.

References II

Fan, Ky (1951). "Maximum properties and inequalities for the eigenvalues of completely continuous operators". In: *Proceedings of the National Academy of Sciences* 37.11, pp. 760–766.

Gupta, Vineet, Tomer Koren, and Yoram Singer (2018). *Shampoo: Preconditioned Stochastic Tensor Optimization*. arXiv: 1802.09568 [cs.LG]. URL: https://arxiv.org/abs/1802.09568.

Hazan, Elad, Kfir Levy, and Shai Shalev-Shwartz (2015). "Beyond convexity: Stochastic quasi-convex optimization". In: *Advances in neural information processing systems* 28.

Jordan, Keller et al. (2024a). *modded-nanogpt: Speedrunning the NanoGPT baseline*. URL: https://github.com/KellerJordan/modded-nanogpt.

Jordan, Keller et al. (2024b). *Muon: An optimizer for hidden layers in neural networks*. URL: https://kellerjordan.github.io/posts/muon/.

Keller, Jordan (2023). *cifar10-airbench*. URL: https://github.com/KellerJordan/cifar10-airbench.

References III

- Kovalev, Dmitry (2025). "Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization". In: *arXiv preprint arXiv:2503.12645*.
- Liu, Jingyuan et al. (2025). "Muon is scalable for llm training". In: *arXiv preprint arXiv:2502.16982*.
- Morwani, Depen et al. (2025). "A New Perspective on Shampoo's Preconditioner". In: *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=c6zI3Cp8c6.
- Pethick, Thomas et al. (2025a). "νSAM: Memory-Efficient Sharpness-Aware Minimization via Nuclear Norm Constraints". In: *Transactions on Machine Learning Research*. ISSN: 2835-8856. URL: https://openreview.net/forum?id=V6ia5hWIMD.
- Pethick, Thomas et al. (2025b). "Training deep learning models with norm-constrained lmos". In: *arXiv preprint arXiv:2502.07529*.

References IV

- Preferred Infrastructure, Inc. and CuPy Developers (2025). *CuPy: `cupyx.scipy.sparse.linalg.svds` — API Reference*. https://docs.cupy.dev/en/stable/reference/generated/cupyx.scipy.sparse.linalg.svds.html. Accessed: 2025-08-24.
- Riabinin, Artem et al. (2025). "Gluon: Making Muon & Scion Great Again!(Bridging Theory and Practice of LMO-based Optimizers for LLMs)". In: *arXiv preprint arXiv:2505.13416*.
- Shah, Ishaan et al. (2025). "Practical efficiency of muon for pretraining". In: *arXiv preprint arXiv:2505.02222*.
- Shi, Hao-Jun Michael et al. (2016). "A primer on coordinate descent algorithms". In: *arXiv preprint arXiv:1610.00040*.
- Si, Chongjie, Debing Zhang, and Wei Shen (2025). "AdaMuon: Adaptive Muon Optimizer". In: *arXiv preprint arXiv:2507.11005*.

Simonz, Kesheng Wuzand Horst (1998). "Thick-Restart Lanczos Method for Symmetric Eigenvalue Problemsy". In.