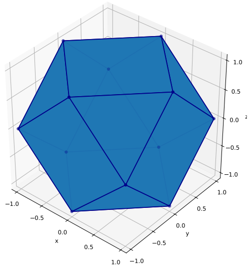
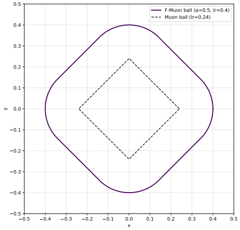
The Ky Fan Norms and Beyond: Dual Norms and Combinations for Matrix Optimization

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]AI4Science [5]AIRI



October 17, 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 $\|\cdot\|^\dagger_{\mathrm{KF-k}}$ 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)

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

LMO leads 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 utilize the SVD of $\boldsymbol{M}^t$: $\boldsymbol{M}^t = \boldsymbol{U}\Sigma\boldsymbol{V}^\top$.

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_∞	$-\eta\,\mathrm{sign}(g)$	(Bernstein et al., 2018, Thm. 1)
Signum	Ball in Max-norm l_∞	$-\eta\,\mathrm{sign}(g)$	(Bernstein et al., 2018, Thm. 3)
Muon	Ball in Spectral S_∞	$-\eta UV^\top$	(Jordan et al., 2024b)
Gauss-Southwell Coordinate Descent	Ball in l_1	$-\eta\sum_{i\in\arg\max\|g_i^t\|}\mathrm{sign}(g_i^t)e_i$	(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

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

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 obtain 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

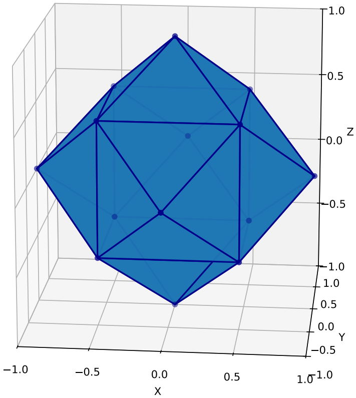
$\|\boldsymbol{M}^t\|^{\dagger\dagger}_{\mathrm{KF-k}} = \|\boldsymbol{M}^t\|_{\mathrm{KF-k}}$, and $\Delta = \sum_{i=1}^{k} u_i v_i^\top$ with $\|\Delta\|^{\dagger}_{\mathrm{KF-k}} = \max\{\frac{1}{k}\|\Delta\|_{\mathrm{nuc}}, \|\Delta\|_{\mathrm{op}}\} = \max\{\frac{1}{k}k, 1\} = 1$ delivers it:

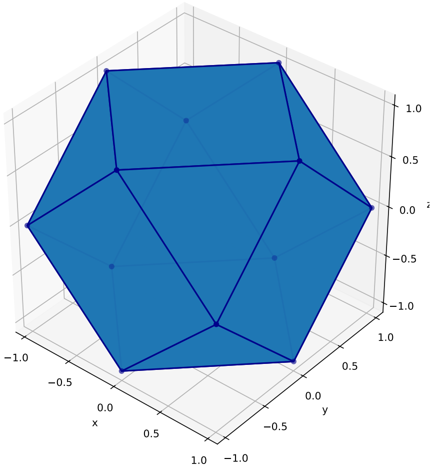
$$\langle \boldsymbol{M}^t, \Delta \rangle = \langle \boldsymbol{U\Sigma V}^\top, \sum_{i=1}^{k} u_i v_i^\top \rangle = \sum_{i,j=1}^{r,k} \langle u_i \sigma_i v_i^\top, u_j v_j^\top \rangle = \sum_{i=1}^{k} \sigma_i = \|\boldsymbol{M}^t\|_{\mathrm{KF-k}}$$

Hence,

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

Geometric intuition: Ky Fan norms





Ky Fan rank-2 ball and its dual in a 3D singular-value space

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 a convex combination of the Ky Fan rank-k norm and the Frobenius norm, which we call the 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 $\|\cdot\|^\dagger_{(i)}$.

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

Deriving F-Fanions by Recipe

$\|\boldsymbol{M}^t\|_{\mathrm{F-KF-k}}^{\dagger\dagger} = \|\boldsymbol{M}^t\|_{\mathrm{F-KF-k}}$, 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. $\|\Delta\|_{\mathrm{F-KF-k}}^\dagger \leq 1$ 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 the Frobenius norm:

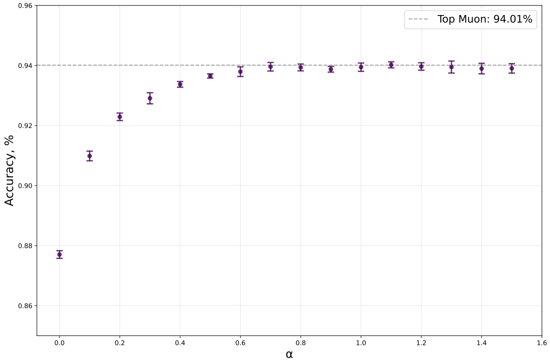
$$\|\boldsymbol{X}\|_{F*} = \alpha\|\boldsymbol{X}\|_* + (1-\alpha)\|\boldsymbol{X}\|_F, \qquad \|\boldsymbol{X}\|_{F2} = \alpha\|\boldsymbol{X}\|_{op} + (1-\alpha)\|\boldsymbol{X}\|_F.$$

Dual-induced updates:

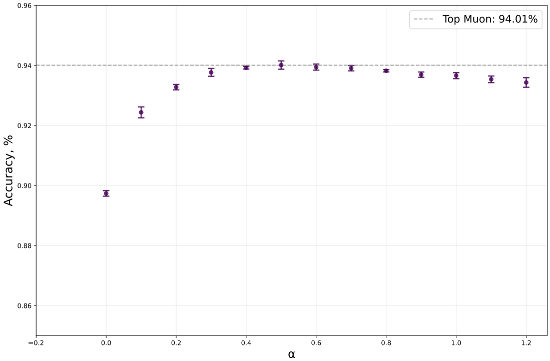
$$\text{F-Muon: } \boldsymbol{X}^{k+1} = \boldsymbol{X}^k - \eta\Big(\alpha\,\boldsymbol{U}\boldsymbol{V}^\top + (1-\alpha)\frac{\boldsymbol{M}^k}{\|\boldsymbol{M}^k\|_F}\Big),$$

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

F-Muon on CIFAR-10 airbench



With params tuned for Muon, as in (Keller, 2023)

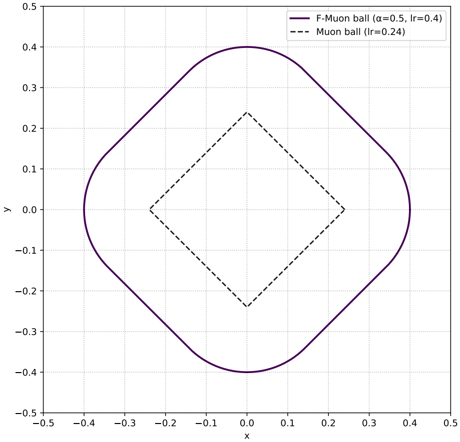


With params tuned for $\alpha = 0.5$ F-Muon

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

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

LMO balls on CIFAR-10



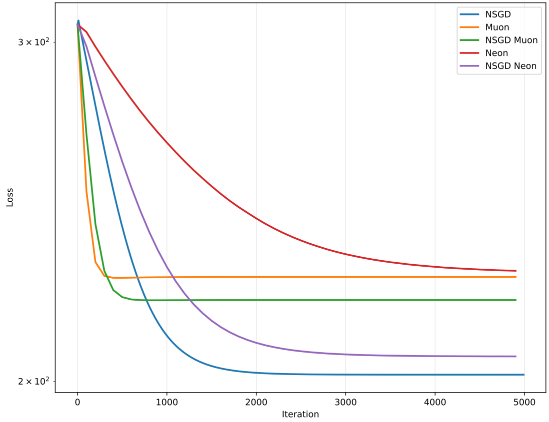
LMO balls for learning rates from the experiment

NanoGPT speedrun

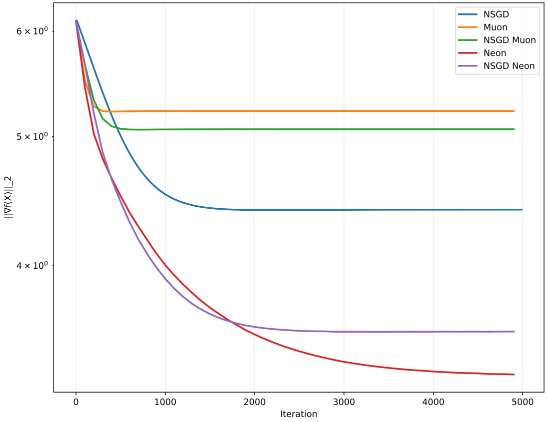
Setup from Jordan et al. (2024a): 1750 iterations and cross-entropy loss lower than 3.28

- Muon: `lr=0.05, momentum=0.95`, final loss = 3.279
- F-Muon with $\alpha = 0.5$: `lr=0.07, momentum=0.95`, final loss = 3.281
- NSGD: `lr=0.07, momentum=0.96`, final loss = 3.4651
- F-Muon with $\alpha = 0.5$: `lr=0.07, momentum=0.96`, final loss = 3.2824!

Random linear least squares: Loss and the Spectral Norm of the Gradient



Loss vs iteration



Spectral norm of gradient vs iteration

Linear least squares problem for a 50 x 50 matrix. Common `lr=0.01`.

Conclusion

- Muon and NSGD are special cases of F-Fanions
- A combination of norm-based updates is a norm-based update
- No existing bound describes the superiority of the spectral norm

For future experiments: Exploration of the matrix and vector algorithms correspondence

References I

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

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

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 II

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

Shi, Hao-Jun Michael et al. (2016). "A primer on coordinate descent algorithms". In: *arXiv preprint arXiv:1610.00040*.