THE KY FAN NORMS AND BEYOND: DUAL NORMS AND COMBINATIONS FOR MATRIX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In this article, we explore the use of matrix norms for optimizing functions of weight matrices, a crucial problem in training large language models. Moving beyond the spectral norm that underlies the Muon update, we leverage the Ky Fan k-norms to introduce a new family of Muon-like algorithms we name *Fanions*. Then, we consider their convex combinations with the Frobenius norm and corresponding duals to develop a new family of algorithms we name *F-Fanions*, one of them being *F-Muon*. We complement our theoretical analysis with an extensive empirical study of the algorithms across a wide range of tasks and settings, from which it follows that properly-tuned F-Muon is on par with Muon, which raises the question about Muon's optimality.

1 INTRODUCTION

Minimizing loss functions in unprecedentedly high-dimensional spaces has recently become an integral and crucial part in training large language models. Hence, new scalable, time- and memory-efficient algorithms have been demanded. Besides well-known Adam (Kingma & Ba, 2014) and AdamW (Loshchilov & Hutter, 2017), recently proposed Muon has shown promising results on training very large models (Jordan et al., 2024; Liu et al., 2025). Its key difference from Adam and AdamW is that it has been constructed specifically for optimizing functions of weight matrices, which are common in deep learning.

That is what can be said from a practical point of view. From the perspective of theory, Muon's main innovation was an intentional usage of matrix norms, i.e. the spectral norm, to derive the algorithm's update (Bernstein, 2025). Based upon recent theoretical advances that explain some theory behind Muon, Scion and Gluon (Bernstein, 2025; Kovalev, 2025; Pethick et al., 2025b; Riabinin et al., 2025), we explore application of other matrix norms to optimization of functions of matrices. As it has been done with Muon, we stipulate that our algorithms' updates be fast to compute.

In this article, we focus on the two most common matrix norms akin to the spectral, namely, the nuclear Norm and the Frobenius norm. Working in the linear minimization oracle (lmo) framework, which is equivalent to a factor to the trust region approach and the steepest descent under norm constraint, we derive Neon, our algorithm based on the nuclear norm. In the section *Matrix side of the updates*, we explain how Neon updates can be computed asymptotically faster than Muon updates by the Newton-Schulz iterations.

Noticing that Neon and Muon are diametrical in terms of the rank of the update matrix, we bridge the space by "regularizing" them by NormalizedSGD, which is derived in lmo with the Frobenius norm. We do this in the same lmo approach by considering a norm that is dual to the convex combination of the Frobenius norm and the spectral or the nuclear norms respectively. So we derive the algorithms we name F-Neon and F-Muon respectively.

Having faced the array of different Muon-like optimizers, according to the upper bounds from Kovalev (2025); Riabinin et al. (2025), with similar convergence behavior, we painstakingly compare them on a synthetic linear least squares problem with known Lipschitz constant. The efforts results in comparison of the algorithms by their convergence not in terms of the dual norms of their gradient, but in terms of the common spectral norm, which may strongly differ, especially in large matrices, from the initial dual norms. Thus, we compare the algorithms in a unified fashion.

Finally, we test Muon, Neon, NSGD, F-Muon and F-Neon on deep-learning tasks: training convolutional network on CIFAR-10 and fine-tuning NanoGPT. The results support the supremacy of Muon, but the most striking result of the tests is that F-Muon, only half of which is Muon, surpasses Muon's accuracy on the CIFAR tasks by a margin. The case of F-Muon answers in the affirmative to the question of feasibility of constructing a mixture of optimization algorithms to increase robustness of the composite algorithm.

2 Problem Statement

2.1 Function of a matrix

We consider the problem of minimizing a differentiable matrix function $F(\cdot)\colon \mathbb{R}^{m\times n} \to \mathbb{R}$

$$\min_{\boldsymbol{X}\in\mathbb{R}^{m\times n}} f(\boldsymbol{X}). \tag{1}$$

We equip the matrix space $\mathbb{R}^{m\times n}$ with a norm $\|\cdot\|\colon \mathbb{R}^{m\times n} \to \mathbb{R}_+$, which possibly does not coincide with the Frobenius norm $\|\cdot\|_{\mathrm{F}}$. $\|\cdot\|^\dagger\colon \mathbb{R}^{m\times n} \to \mathbb{R}_+$ is the dual norm associated with $\|\cdot\|$, i.e., $\|\boldsymbol{X}\|^\dagger = \sup_{\|\boldsymbol{X}'\|\leq 1}\langle \boldsymbol{X}, \boldsymbol{X}'\rangle$ for all $\boldsymbol{X} \in \mathbb{R}^{m\times n}$.

2.2 Linear Minimization Oracle and Trust Region

We analyze the problem from the perspective of linear minimization oracle (lmo) and unconstrained stochastic conditional gradient descent (uSCG) (Pethick et al., 2025b). lmo is defined as:

$$\mathrm{lmo}(\boldsymbol{S}) \in \arg\min_{\boldsymbol{X}\in\mathcal{S}} \langle \boldsymbol{S}, \boldsymbol{X}\rangle, \tag{2}$$

where $\mathcal{S}$ is some set. We are interested in the case when $\mathcal{S}$ is a ball in our $\|\cdot\|$ norm:

$$\mathcal{S} := \mathcal{B}_\eta := \{\boldsymbol{X} \mid \|\boldsymbol{X}\| \leq \eta\}. \tag{3}$$

uSCG update is defined as: $\boldsymbol{X}^{k+1} = \boldsymbol{X}^k + \gamma_k lmo(\boldsymbol{M}^k)$, where $\boldsymbol{M}^{k+1} = (1-\alpha_{k+1})\boldsymbol{M}^k + \alpha_{k+1} g(\boldsymbol{X}^k, \xi_k)$ is a momentum.

It can be easily shown that the formula is equvalent to

$$\boldsymbol{X}^{k+1} = \boldsymbol{X}^k - \gamma_k \eta \arg\max_{\boldsymbol{X}\in\mathcal{B}_1} \langle \boldsymbol{S}, \boldsymbol{X}\rangle = \boldsymbol{X}^k - \gamma_k \eta \{\Delta \in \mathcal{B}_1 \mid \langle \boldsymbol{M}^k, \Delta\rangle = \|\boldsymbol{M}^k\|^\dagger\}. \tag{4}$$

Let us set $\gamma_k = 1$. This transforms algorithm defined by eq. (4) into Algorithm 1 from Kovalev (2025). Therefore, we can view eq. (4) both as an lmo-based algorithm and as a trust-region algorithm.

3 Different Norms $\|\cdot\|$ imply different updates

Based on different norms $\|\cdot\|$, we will simplify the update defined by the aforementioned equation:

$$\boldsymbol{X}^{k+1} = \boldsymbol{X}^k - \eta\{\Delta \in \mathcal{B}_1 \mid \langle \boldsymbol{M}^k, \Delta\rangle = \|\boldsymbol{M}^k\|^\dagger\} \tag{5}$$

Throughout this work, we denote the singular value decomposition (SVD) of $\boldsymbol{M}^k$ as $\boldsymbol{M}^k = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, where $\boldsymbol{U} = [u_1, u_2, \ldots, u_r]$, $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$, and $\boldsymbol{V} = [v_1, v_2, \ldots, v_r]$.

3.1 $\|\boldsymbol{M}^k\|_{\mathrm{F}}$ and Normalized SGD

Lemma 1. *When* $\|\cdot\| = \|\cdot\|_{\mathrm{F}}$*, eq.* (5) *turns into:*

$$\boldsymbol{X}^{k+1} = \boldsymbol{X}^k - \eta \frac{\boldsymbol{M}^k}{\|\boldsymbol{M}^k\|_{\mathrm{F}}} \tag{6}$$

The result is the same as in (Table 1, Pethick et al. (2025b)), but here we use the Euclidean norm as $\|\boldsymbol{M}^k\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2}$, which clearly is a matrix norm.

3.2 $\|\boldsymbol{M}^k\|_{\mathrm{op}}$ AND MUON

Lemma 2. *When* $\|\cdot\| = \|\cdot\|_{\mathrm{op}}$*, eq.* (5) *turns into:*

$$\boldsymbol{X}^{k+1} = \boldsymbol{X}^k - \eta UV^\top \tag{7}$$

Though the update is well-known, we provide a much simpler proof in the appendix, when compared to Bernstein & Newhouse (2024).

3.3 $\|\boldsymbol{M}^k\|_{\mathrm{nuc}}$ AND NEON

Lemma 3. *When* $\|\cdot\| = \|\cdot\|_{\mathrm{nuc}}$*, eq.* (5) *turns into:*

$$\boldsymbol{X}^{k+1} = \boldsymbol{X}^k - \eta u_1 v_1^\top \tag{8}$$

We name the derived algorithm *Neon*. In the section *The Matrix side of updates*, we will discuss how to compute the update efficiently.

3.4 $\|\boldsymbol{M}^k\|^\dagger_{\mathrm{F}*}$ AND F-MUON

We define $\|\cdot\|_{\mathrm{F}*}$ as a convex combination of $\|\cdot\|_{\mathrm{nuc}}$ and $\|\cdot\|_{\mathrm{F}}$:

$$\|\boldsymbol{X}\|_{\mathrm{F}*} = \alpha\|\boldsymbol{X}\|_{\mathrm{nuc}} + (1-\alpha)\|\boldsymbol{X}\|_{\mathrm{F}}, \tag{9}$$

where $\alpha \in [0,1]$ defines a specific norm from the $F*$-family. $\|\cdot\|^\dagger_{\mathrm{F}*}$ can be expressed as in eq. (18). However, its norm ball is a simple Minkowski sum of $\alpha\|\cdot\|_{\mathrm{nuc}}$ and $(1-\alpha)\|\cdot\|_{\mathrm{F}}$ balls fig. 2a.

Lemma 4. *When* $\|\cdot\| = \|\cdot\|^\dagger_{\mathrm{F}*}$*, eq.* (5) *turns into:*

$$\boldsymbol{X}^{k+1} = \boldsymbol{X}^k - \eta(\alpha UV^\top + (1-\alpha)\frac{\boldsymbol{M}^k}{\|\boldsymbol{M}^k\|_{\mathrm{F}}}) \tag{10}$$

We name the derived algorithm *F-Muon*. It turns out that F-Muon is a convex combination of Normalized SGD and Muon. The implications are significant and discussed in the following sections.

3.5 $\|\boldsymbol{M}^k\|^\dagger_{\mathrm{F2}}$ AND F-NEON

We define $\|\cdot\|_{\mathrm{F2}}$ as a convex combination of $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_{\mathrm{F}}$:

$$\|\boldsymbol{X}\|_{\mathrm{F2}} = \alpha\|\boldsymbol{X}\|_{\mathrm{op}} + (1-\alpha)\|\boldsymbol{X}\|_{\mathrm{F}}, \tag{11}$$

where $\alpha \in [0,1]$ defines a specific norm from the F2-family.$\|\cdot\|^\dagger_{\mathrm{F2}}$ can be expressed as in eq. (19). However, its norm ball is a simple Minkowski sum of $\alpha\|\cdot\|_{\mathrm{op}}$ and $(1-\alpha)\|\cdot\|_{\mathrm{F}}$ balls fig. 2b.

Lemma 5. *When* $\|\cdot\| = \|\cdot\|^\dagger_{\mathrm{F2}}$*, eq.* (5) *turns into:*

$$\boldsymbol{X}^{k+1} = \boldsymbol{X}^k - \eta(\alpha u_1 v_1^\top + (1-\alpha)\frac{\boldsymbol{M}^k}{\|\boldsymbol{M}^k\|_{\mathrm{F}}}) \tag{12}$$

We name the derived algorithm *F-Neon*. It turns out that F-Neon is a convex combination of Normalized SGD and Neon. The implications are significant and discussed in the following sections.

Table 1: lmo optimizers in Schatten S_p norms and in l_p norms. g is the gradient. When it is a matrix, $g = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$

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., 2024)
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

3.6 $\|\boldsymbol{M}^k\|^\dagger_{\mathrm{KF-k}}$ AND MUON, NEON, AND CENTRALIZED DION WITHOUT ERROR FEEDBACK

We remind the reader that the Ky Fan k-norm (Bhatia (2013), p. 92), which we denote as $\|\cdot\|_{\mathrm{KF-k}}$, is the sum of the k largest singular values of the matrix. It can be proved that $\|\cdot\|^\dagger_{\mathrm{KF-k}} = \max\{\frac{1}{k}\|\cdot\|_{\mathrm{nuc}}, \|\cdot\|_{\mathrm{op}}\}$ (see Bhatia (2013), p. 96). Special cases of the Ky Fan k-norm are the Ky Fan 1-norm, which is the sprectral norm, and the Ky Fan $\min\{m, n\}$-norm, which is the nuclear norm.

Lemma 6. *When* $\|\cdot\| = \|\boldsymbol{M}^k\|^\dagger_{\mathrm{KF-k}}$, *eq.* (5) *turns into:*

$$\boldsymbol{X}^{k+1} = \boldsymbol{X}^k - \eta\sum_{i=1}^{k} u_i v_i^\top \tag{13}$$

In the section *The Matrix side of updates*, we will discuss how to compute the updates efficiently.

We introduce the family of *Fanions*, which consists of k-*Fanions*, lmo-based algorithms under the $\|\boldsymbol{M}^k\|^\dagger_{\mathrm{KF-k}}$ norms. By this terminology and the lemma, Muon is a $\min\{n, m\}$-Fanion, while Neon is an 1-Fanion. Moreover, the centralized version of rank-r Dion (Ahn & Xu, 2025) without the error feedback and without scaling of the update, from the perspective of lmo, is actually a r-Fanion.

In addition, one can consider F-KF-k-norm = $\alpha\|\cdot\|_{\mathrm{KF-k}} + (1-\alpha)\|\cdot\|_{\mathrm{F}}$, the dual to which will produce algorithms like F-Dion without the error feedback. The resulting family can be named *F-Fanions*.

3.7 ALGORITHMS FOR MATRICES ↔ ALGORITHMS FOR VECTORS

lmo optimizers in Schatten S_p norms and in l_p norms with common p may be analogous to each other, as is illustrated by table 1. The analogies go beyond similarities in the updates. SignSGD is very close to Adam, as noted in Bernstein & Newhouse (2024), and both Adam and Muon perform well in training large models. NSGD is the same for both matrix and vector cases. Greedy Coordinate Descent methods are not applied to high-dimensional problems, from this perspective, it is not surprising that Neon underperforms on such problems.

Despite such similarities, no theory has been yet proposed that would reduce the matrix case to the optimization of a function of a singular values vector. If such a theory is developed, analysis of matrix algorithms like Muon will be greatly simplified.

4 MATRIX SIDE OF UPDATES

To compute the algorithms' updates, we use thick-restarted Lanczos method for singular value problem (TRLan) on $\boldsymbol{M}^{k\top}\boldsymbol{M}^k$ or $\boldsymbol{M}^k\boldsymbol{M}^{k\top}$ matrices (the one with less size is picked), implemented in CuPy lSibrary (Preferred Infrastructure & Developers, 2025) and described in Simonz (1998).

This method is designed for efficiently approximating the largest singular values and vectors of large matrices. Its thick-restart strategy retains the most informative Ritz vectors at each cycle, which accelerates convergence while avoiding excessive memory growth. We are specifically interested in this algorithm because it allows us to extract several largest singular values and related singular

vectors of the matrix to make a Neon step. Moreover, TRLan is stable GPU-friendly algorithm because it mainly operates with matrix-vector multiplications (MVs), which are higly-parallel, and does not require full reorthogonalization against the whole Krylov basis by managing short recurrent formulas and incorparating thick restarting.

Per-cycle complexity is $O(mn^2 + n^2k + nk^2)$, where m and n are dimensions of target matrix and n is the smallest one, k is retained subspace's size.

In table 2, we compare performance of TRLan, RSVD, and power iterations on calculating k-rank update, which is used in k-Fanion. The results highlight that TrLan is much faster that its competitors.

During the research it was noted that RSVD can give good and fast approximation for singular values, but the matrix of approximation is far from the one given by truncated SVD, while TRLan gives good and fast approximation of a matrix, but not so good approximation for singular values. That means that TRLan may be not a perfect choice for algorithms like Dion, where σ_i are required for error feedback.

Method	rtol	k	time,s
Power Iterations	0.01	1	7.7
SVDS (thick-restart Lanczos method)	0.01	1	0.18
PCA Low Rank (RSVD)	0.01	1	1.15
SVDS (thick-restart Lanczos method)	0.01	10	0.47
PCA Low Rank (RSVD)	0.01	10	19.4
SVDS (thick-restart Lanczos method)	0.01	100	1.96
PCA Low Rank (RSVD)	0.01	100	170

Table 2: k-rank updated comparison
Comparison of different numerical methods to calculate k-rank update on 5000×5000 matrix of real numbers, rtol is an error in Frobenius norm relative to the k-rank approximation.

5 Trust Region Bounds for L-smooth functions

First, we analyze the problem in the unstochastic case. From Corollary 1 of Kovalev (2025), we directly get the following result that matches lower bounds, as was noted in that article.

Lemma 7. *To reach the precision $\min_{k=1\dots K}\|\nabla f(\boldsymbol{X}_k)\|^\dagger \leq \varepsilon$ by the iterations equation 5 under the conditions of Assumption (A1), it is sufficient to choose the stepsize η and the number of iterations K as follows:*

$$\eta = \mathcal{O}\left(\frac{\varepsilon}{L}\right), \qquad K = \mathcal{O}\left(\frac{L\Delta_0}{\varepsilon^2}\right). \tag{14}$$

In the stochastic case, from Corollary of 2 of Kovalev (2025), we directly get the following result that once more matches lower bounds:

Lemma 8. *To reach the precision $\mathbb{E}\min_{k=1\dots K}\|\nabla f(\boldsymbol{X}_k)\|^\dagger \leq \varepsilon$ by equation 5 under the assumptions Assumption (A2), Assumption (A1), Assumption (A3), it is sufficient to choose the parameters as follows:*

$$\eta = \mathcal{O}\left(\min\left\{\frac{\varepsilon}{L}, \frac{\varepsilon^3}{\rho^2\sigma^2 L}\right\}\right), \qquad \alpha = \mathcal{O}\left(\min\left\{1, \frac{\varepsilon^2}{\rho^2\sigma^2}\right\}\right), \tag{15}$$

$$K = \mathcal{O}\left(\max\left\{\frac{\rho\sigma}{\varepsilon}, \frac{\rho^3\sigma^3}{\varepsilon^3}, \frac{L\Delta_0}{\varepsilon^2}, \frac{L\Delta_0\rho^2\sigma^2}{\varepsilon^4}\right\}\right). \tag{16}$$

As the norms $\|\cdot\|_{\mathrm{F}}$, $\|\cdot\|_{\mathrm{nuc}}$, $\|\cdot\|_{\mathrm{F}}$ are almost proportional to each other when $m, n \to \infty$ (with high probability for random matrices), the expected convergence guarantees in terms of $\|\cdot\|_{\mathrm{F}}$ are the same (it can be easily shown by noting that $\|\boldsymbol{X}\| \sim \alpha\|\boldsymbol{X}\|_{\mathrm{F}}$, $\|\nabla f(\boldsymbol{X})\|^\dagger \sim \frac{1}{\alpha}\|\nabla f(\boldsymbol{X})\|_{\mathrm{F}}$, and expressing L-constant via L_F-constant for the Frobenius norm).

From the theory of random martices and the Marchenko-Pastur law, we get that random $\boldsymbol{M} \in \mathbb{R}^{m\times n} : \boldsymbol{M} \sim \mathcal{N}(0, \sigma^2\mathbb{R}^{m\times n})$ has the following asymptotics of its norms:

Nuclear: $\sigma n\sqrt{m}$

Frobenius: $\sigma\sqrt{mn}$

Spectral: $\sigma(\sqrt{m}+\sqrt{n})$

This means that for square random matrices $n \times n$ the following asymptotics take place: $\|\cdot\|_{\mathrm{F}} \sim \frac{\sqrt{n}}{2}\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_{\mathrm{nuc}} \sim \frac{n}{2}\|\cdot\|_{\mathrm{op}}$.

6 EXPERIMENTS

6.1 CIFAR-10 AIRBENCH

We adapt Keller Jordan's code to test F-Muon, Neon, and F-Neon on the CIFAR-10 airbench (Keller, 2023). First, we run F-Muon for different α with the same `lr=0.24(1 - step/total_steps factor), momentum=0.6, nesterov=True`, as have been finetuned by Jordan. We record the accuracy after 8 epochs of training.

Then, we tune F-Muon with $\alpha = 0.5$. Tuned parameters are `lr=0.4(1 - step/total_steps factor), momentum=0.65 - 0.1(1 - step/total_steps factor), nesterov = True`. While Muon reaches $94.00 \pm 0.13\%$ accuracy after 8 epochs, tuned F-Muon reaches $93.97 \pm 0.14\%$ after 8 epochs.

Finally, we take the set of tuned parameters and test on different α. We notice that even when $\alpha = 0.1$, the accuracy is much higher than in case of the pure NSGD.

The results are curious and could be represented by fig. 1: lmo ball, which we plotted in a 2D space of singular values, has drastically changed, but the observed convergence after tuning has not degraded. These observations raise the question of how much lmo-based algorithms are sensitive to the constraint area, i.e. what will happen if the ball is corrupted. In this particular example, we have had a blurred ball, which proved as robust as the original ball.

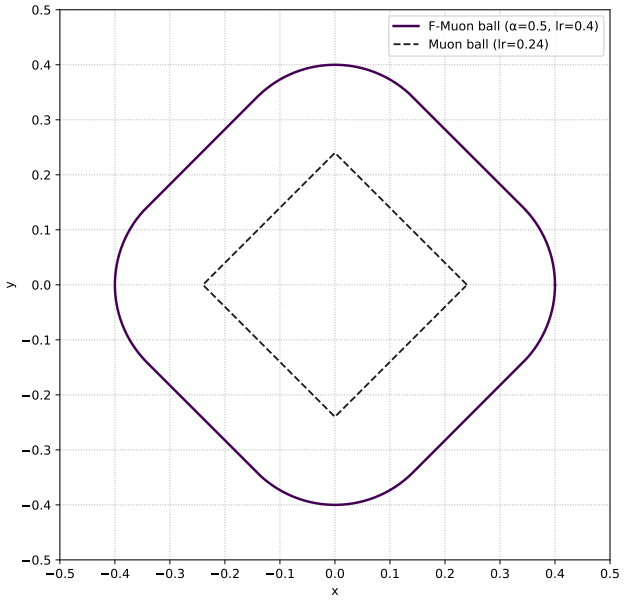


Figure 1: Visualization of lmo balls for Muon and F-Muon.

6.2 MODDED NANOGPT

NanoGPT benchmark

7 Related Work

As Muon (Jordan et al., 2024) is a very successful and popular optimizer for functions of weight matrices, a lot of reseach has been put into, first, further improving its performance, and, second, in explaining its success.

Improvements of Muon. Regarding the first point, in less than a year, a large number of applications and improvements of Muon has been proposed. Liu et al. (2025) adapted the algorthm for training language models larger than NanoGPT. Shah et al. (2025) organized efficient hyperparameter transfer by combining Muon with maximal update parametrization. To construct their COSMOS optimizer, Chen et al. (2025) applied computationally intensive updates of SOAP optimizer to a low-dimensional "leading eigensubspace" while using memory-efficient methods like Muon for the remaining parameters. Amsel et al. (2025) proposed a more efficient alternative to Newton-Schulz operations. Si et al. (2025) introduced AdaMuon which combines element-wise adaptivity with orthogonal updates. We suppose that the described above or similar techniques can be applied to our optimizers as well. For example, F-Muon also benefits from faster alternatives to Newton-Schulz iterations, and Neon may be a great substitute to Muon in COSMOS, because, as we have shown in *the Matrix side of the updates*, Lanczos algorithms is much faster than Newton-Schulz iterations on large matrices.

Theory behind Muon. Regarding the second point, there has been a prolonged gap in theory behind Muon, simplistic derivation of Bernstein (2025) based on Bernstein & Newhouse (2024) excluded. This gap, as it seems to us, is not even now completely closed. For example, Kovalev (2025) has provided convergence guarantees of Muon in various settings, from which, however, Muon's supremacy cannot be recovered. Indeed, although the obtained bounds depend on the norm choice, the asymptotics of the convergence remain the same as for NSGD and other optimizers, $K = \mathcal{O}(\varepsilon^{-4})$ in a L-smooth stochastic case.

Similar drawback has a recent article Riabinin et al. (2025), where L-smoothness assumption is replaced with a more practical (L_0, L_1)-smoothness. The authors derived from their theorems optimal stepsizes for Muon and Scion that match fine-tuned stepsizes reported by Pethick et al. (2025b). But still, they did not explain why, for instance, NSGD is inferior to Muon in training large-language models.

We suppose that the reason for the recorded by us discrepancy between Neon and Muon performance, both of which are described by Scion or Gluon frameworks, lies in the structure of the norm ball, which must be an object of further research.

The nuclear norm in lmo. As we found out only when writing this article, the nuclear norm has been already explored in the context of the linear minimization oracle. Pethick et al. (2025a) applied it to create νSAM, a new sharpness-aware minimization technique. The update from their Lemma 3.1 coincides with the update of Neon, but is used for completely different purposes. In addition, Pethick et al. (2025a) use power iterations to find u_1 and $v_1^\top$, while we suggest utilizing much more efficient and precise Lanczos algorithm.

The Ky Fan Norm and Dion. Rank-k Centralized Dion, Algorithm 1 from Ahn & Xu (2025), without an error feedback and scaling of the update, turns out to be an lmo-based algorithm under the $\|\cdot\|^\dagger_{\mathrm{KF-k}}$ norm, which we described in *Different norms imply different updates*. Reported by the authors necessity of using error feedback to obtain satisfactory accuracy may take place in the cases of our algorithms as well, for example, F-Neon. This we leave to future research.

8 Conclusion

In this article, we have generalized several successful algorithms, like Muon and Dion, to the lmo-based algorithms in the $\|\cdot\|^\dagger_{\mathrm{KF-k}}$ norm. Also we have proposed the technique of "regularizing" the updates with NSGD, a trick to increase the robustness of the algorithms and motivated by the consideration of the $\|\cdot\|_{\mathrm{F}*}$ and $\|\cdot\|_{\mathrm{F}2}$ norms. Generalizations of well-known norms and subsequent combination of them may further improve performance of lmo-based algorithms. If a theory is developed that explains the discrepancy between performance of different algorithms based on the matrix norms, one will probably be able to apply it to the generalizations and combinations of the

norms as well, which leaves an open space to construct norms specific for given architectures and probably even their parameters.

9 Author Contributions

IO suggested using the nuclear norm in the Bernstein & Newhouse (2024) framework. DM supervised the project and helped with editing the article. IK, NK, and AV participated mainly on the first stage of research, when it has been a project in the optimization course at MIPT. IK suggested using composite norms (though not F2 and F*) and KKT conditions to find the resulting updates. NK suggested Lanczos algorithm as the fastest tool to compute Neon's updates. AV tested Neon on the finetuning of NanoGPT. All other work was done by AK.

References

Kwangjun Ahn and Byron Xu. Dion: A communication-efficient optimizer for large models, 2025. URL https://arxiv.org/abs/2504.05295.

Noah Amsel, David Persson, Christopher Musco, and Robert M Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. *arXiv preprint arXiv:2505.16932*, 2025.

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

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

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

Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Weizhu Chen, Chen Liang, Tuo Zhao, Zixuan Zhang, Hao Kang, Liming Liu, Zichong Li, and Zhenghao Xu. Cosmos: A hybrid adaptive optimizer for memory-efficient training of llms, 2025. URL https://arxiv.org/abs/2502.17410.

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Momentum-based variance reduction in nonconvex sgd. *Advances in Neural Information Processing Systems*, 2020.

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

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/.

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

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.

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.

Thomas Pethick, Parameswaran Raman, Lenon Minorics, Mingyi Hong, Shoham Sabach, and Volkan Cevher. νSAM: Memory-efficient sharpness-aware minimization via nuclear norm constraints. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=V6ia5hWIMD.

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*, 2025b.

Inc. Preferred Infrastructure and CuPy Developers. CuPy: cupyx.scipy.sparse.linalg.svds — api reference. https://docs.cupy.dev/en/stable/reference/generated/cupyx.scipy.sparse.linalg.svds.html, 2025. Accessed: 2025-08-24.

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.

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.

Hao-Jun Michael Shi, Shenyinying Tu, Yangyang Xu, and Wotao Yin. A primer on coordinate descent algorithms. *arXiv preprint arXiv:1610.00040*, 2016.

Chongjie Si, Debing Zhang, and Wei Shen. Adamuon: Adaptive muon optimizer. *arXiv preprint arXiv:2507.11005*, 2025.

Kesheng Wuzand Horst Simonz. Thick-restart lanczos method for symmetric eigenvalue problemsy. 1998.

Yao-Liang Yu. Arithmetic duality for norms, 2012. URL https://cs.uwaterloo.ca/~y328yu/notes/normduality.pdf.

A FORMAL ASSUMPTIONS

Here we restate assumptions from Kovalev (2025) that are used in the theorems from the main part of the article. **Lipschitz continuous gradient.** We assume that the gradient $\nabla f(\cdot)$ is Lipschitz continuous with respect to the norm $\|\cdot\|$, that is, the following inequality holds:

$$\|\nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{X}')\|^\dagger \leq L\|\boldsymbol{X} - \boldsymbol{X}'\| \quad \text{for all } \boldsymbol{X}, \boldsymbol{X}' \in \mathbb{R}^{m\times n}, \tag{A1}$$

where $L > 0$ is the gradient Lipschitz constant.

Stochastic gradient estimator. We assume access to a stochastic estimator $g(\cdot\,;\xi)\colon \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ of the gradient $\nabla f(\cdot)$, where $\xi \sim \mathcal{D}$ is a random variable sampled from a probability distribution $\mathcal{D}$. We assume that the stochastic gradient estimator $g(\cdot;\xi)$ is unbiased and has bounded variance, that is, the following relations hold:

$$\mathbb{E}_{\xi\sim\mathcal{D}}\left[g(\boldsymbol{X};\xi)\right] = \nabla f(\boldsymbol{X}) \quad \text{and} \quad \mathbb{E}_{\xi\sim\mathcal{D}}\left[\|g(\boldsymbol{X};\xi) - \nabla f(\boldsymbol{X})\|_{\mathrm{F}}^2\right] \leq \sigma^2 \quad \text{for all } \boldsymbol{X} \in \mathbb{R}^{m\times n}, \tag{A2}$$

where $\sigma > 0$ is a positive variance parameter, and $\|\cdot\|_{\mathrm{F}}$ is the standard Euclidean, i.e. Frobenius, norm induced by the inner product $\langle\cdot,\cdot\rangle$, i.e., $\|\boldsymbol{X}\|_{\mathrm{F}} = \sqrt{\langle \boldsymbol{X}, \boldsymbol{X}\rangle} = \sqrt{\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X})}$.

It is important to highlight that while Assumption (A1) uses the dual norm $\|\cdot\|^\dagger$ to measure the difference between the gradients, the variance in Assumption (A2) is measured with respect to the Frobenius norm $\|\cdot\|_{\mathrm{F}}^2$, which is necessary to properly utilize the unbiasedness property of the stochastic gradient estimator $g(\cdot;\xi)$. Therefore, we need to provide a connection between these norms using the following inequality:

$$\|\boldsymbol{X}\|^\dagger \leq \rho \cdot \|\boldsymbol{X}\|_{\mathrm{F}} \quad \text{for all } \boldsymbol{X} \in \mathbb{R}^{m\times n}, \tag{A3}$$

where $\rho > 0$ is a positive constant.

B NORMS $\|\cdot\|_{\mathrm{F}*}^{\dagger}$ AND $\|\cdot\|_{\mathrm{F}2}^{\dagger}$

First, we need a well-known fact mentioned, for example, in Yu (2012, Table 1). For the sake of completeness, we provide a proof of the fact.

Lemma 9. *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

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

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

Proof. Write $f(x) = \alpha\|x\|_{(1)}$ and $g(x) = \beta\|x\|_{(2)}$, so

$$\|x\| = f(x) + g(x).$$

Recall two standard facts:

1. For any norm $\|\cdot\|$ and $\lambda > 0$,
$$(\lambda\|\cdot\|)^*(y) = \sup_x\big(\langle y, x\rangle - \lambda\|x\|\big) = \delta_{\lambda B_{\|\cdot\|^{\dagger}}}(y),$$
i.e., the indicator function of the scaled dual ball.
2. The Fenchel conjugate of a sum satisfies
$$(f+g)^*(y) = \inf_{u+v=y}\big(f^*(u) + g^*(v)\big).$$

Applying these to f and g, we have

$$f^*(u) = \delta_{\alpha B_{\|\cdot\|_{(1)}^{\dagger}}}(u), \quad g^*(v) = \delta_{\beta B_{\|\cdot\|_{(2)}^{\dagger}}}(v).$$

Thus

$$\|\cdot\|^*(y) = (f+g)^*(y) = \inf_{u+v=y}\big(\delta_{\alpha B_{\|\cdot\|_{(1)}^{\dagger}}}(u) + \delta_{\beta B_{\|\cdot\|_{(2)}^{\dagger}}}(v)\big) = \delta_{\alpha B_{\|\cdot\|_{(1)}^{\dagger}} + \beta B_{\|\cdot\|_{(2)}^{\dagger}}}(y).$$

But by definition, the conjugate of a norm is exactly the indicator of its dual unit ball:

$$\|\cdot\|^*(y) = \delta_{B_{\|\cdot\|^{\dagger}}}(y).$$

Therefore,

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

□

Consequently,

$$\|y\|^{\dagger} = \inf\big\{t \geq 0 : y \in t\big(\alpha B_{\|\cdot\|_{(1)}^{\dagger}} + \beta B_{\|\cdot\|_{(2)}^{\dagger}}\big)\big\} = \inf_{z,t}\big\{t \geq 0 : z \in t\alpha B_{\|\cdot\|_{(1)}^{\dagger}}, y - z \in t\beta B_{\|\cdot\|_{(1)}^{\dagger}}\big\} \tag{17}$$

Thus, we immediately find $\|\cdot\|_{\mathrm{F}*}^{\dagger}$, which is related to F-Muon update. Indeed, after setting $\beta = 1-\alpha$ and remembering that for smooth and bounded cases we can use $\min$ instead of $\inf$, we get

$$\|\boldsymbol{Y}\|_{\mathrm{F}*}^{\dagger} = \min_{\boldsymbol{Z}} \min_{t}\{t, s.t. \|\boldsymbol{Z}\|_{\mathrm{op}} \leq \alpha t, \|\boldsymbol{Y} - \boldsymbol{Z}\|_{\mathrm{F}} \leq (1-\alpha)t\} \tag{18}$$

If $\alpha = 1$, then $\boldsymbol{Z} = \boldsymbol{Y}$, and we get $\|\boldsymbol{Y}\|_{\mathrm{F}*}^{\dagger} = \|\boldsymbol{Y}\|_{\mathrm{op}}$. If $\alpha = 0$, then $\boldsymbol{Z} = 0$, and we get $\|\boldsymbol{Y}\|_{\mathrm{F}*}^{\dagger} = \|\boldsymbol{Y}\|_{\mathrm{F}}$.

Similarly, we find $\|\cdot\|_{\mathrm{F}2}^{\dagger}$, which is related to F-Neon update:

$$\|\boldsymbol{Y}\|_{\mathrm{F}2}^{\dagger} = \min_{\boldsymbol{Z}} \min_{t}\{t, s.t. \|\boldsymbol{Z}\|_{\mathrm{nuc}} \leq \alpha t, \|\boldsymbol{Y} - \boldsymbol{Z}\|_{\mathrm{F}} \leq (1-\alpha)t\} \tag{19}$$

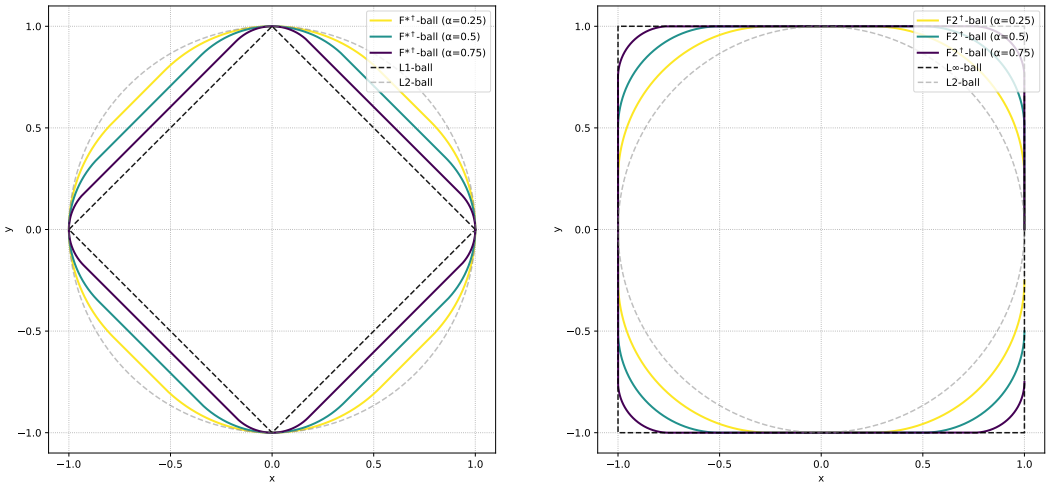
If $\alpha = 1$, then $\boldsymbol{Z} = \boldsymbol{Y}$, and we get $\|\boldsymbol{Y}\|_{\mathrm{F}2}^{\dagger} = \|\boldsymbol{Y}\|_{\mathrm{nuc}}$. If $\alpha = 0$, then $\boldsymbol{Z} = 0$, and we get $\|\boldsymbol{Y}\|_{\mathrm{F}2}^{\dagger} = \|\boldsymbol{Y}\|_{\mathrm{F}}$.

C VISUALIZATION OF DIFFERENT MATRIX NORMS

C.1 DUALS TO F* AND F2 NORMS

It follows from lemma 9 that the norm ball in $\|\cdot\|_{\mathrm{F}*}^{\dagger}$ is the Minkowski sum of the norm ball in $\alpha\|\cdot\|_{\mathrm{nuc}}$ and $(1-\alpha)\|\cdot\|_{\mathrm{F}}$ and the norm ball in $\|\cdot\|_{\mathrm{F2}}^{\dagger}$ is the Minkowski sum of the norm ball in $\alpha\|\cdot\|_{\mathrm{op}}$ and $(1-\alpha)\|\cdot\|_{\mathrm{F}}$.

In Fig. fig. 2 we plot these norms. On x-axis and y-axis, there are singular values σ_1, σ_2 respectively of a matrix from $\mathbb{R}^{m\times n}$ with $\min\{m,n\}=2$.



(a) lmo balls for F-Muon for different α

(b) lmo ball for F-Neon for different α

Figure 2: Balls in the duals to F* and F2 norms for different α

C.2 THE KY FAN NORM AND ITS DUAL

1-balls in l_∞, l_1 and l_2 norms are well-known from textbooks. But what about the Ky Fan k-norm? How can it be represented?

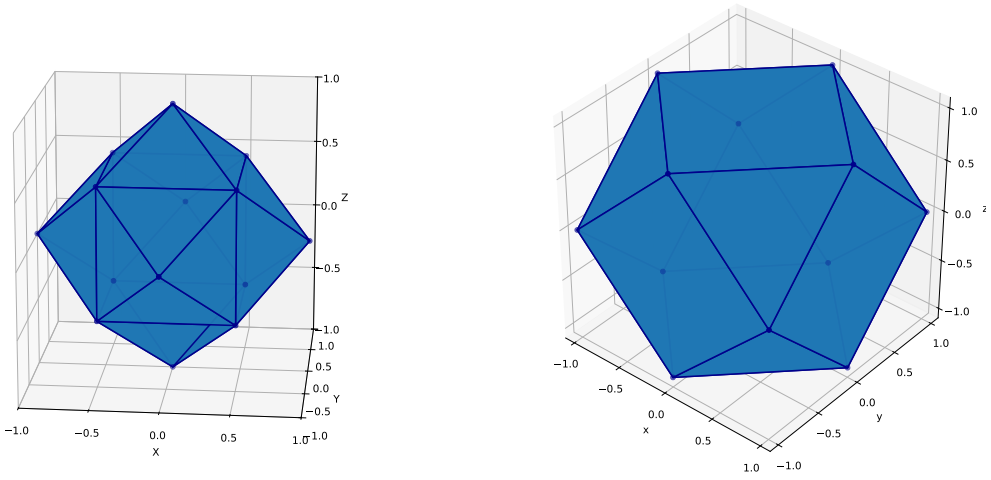
To showcase the complex structure of the Ky Fan k-norm and its dual, we suggest the illustrations fig. 3 with the ball in the Ky Fan 2-norm in fig. 3a and its dual in fig. 3b. On x-, y-, and z-axes, there are singular values σ_1, σ_2, and σ_3 respectively of a matrix from $\mathbb{R}^{m\times n}$ with $\min\{m,n\}=3$. In this particular case, we do not sort the singular values. In the proposed representation, we actually plot balls in the Top-2 norm $\max\{|x|+|y|, |x|+|z|, |y|+|z|\}$ and its dual norm $\max\{\max(|(|x), |y|, |z|), \frac{1}{2}(|x|+|y|+|z|)\}$. The resulting balls are much more complex than balls in l_∞, l_1 and l_2 norms.

In fact, those balls can be described easier if we use the results from Yu (2012). The Ky Fan 2-norm ball is an intersection of three l_1 balls in (x,y), (x,z), and (y,z) spaces. The 1-ball in the dual Ky Fan 2-norm is an intersection of 1-ball the in l_∞ norm and $\frac{1}{2}$-ball in the l_1 norm.

D UPDATES DERIVATIONS

Proof of lemma 2 follows from eq. (10) with $\alpha=1$. Indeed, $\|\cdot\|_{\mathrm{op}}^{\dagger} = 1\cdot\|\cdot\|_{\mathrm{nuc}} + 0\cdot\|\cdot\|_{\mathrm{F}}$.

Proof of lemma 4: Since $\|\cdot\|^{\dagger} = \|\cdot\|_{\mathrm{F}*}^{\dagger\dagger} = \|\cdot\|_{\mathrm{F}*}$, the goal is to reach $\|\boldsymbol{M}^k\|_{\mathrm{F}*} = \alpha\,\mathrm{tr}\,\boldsymbol{\Sigma} + (1-\alpha)\|\boldsymbol{M}^k\|_{\mathrm{F}}$.



(a) Ky Fan 2-norm

(b) Dual to Ky Fan 2-norm

Figure 3: Ky Fan 2-norm and its dual

Let us note that $\Delta = \alpha \boldsymbol{U}\boldsymbol{V}^\top + (1-\alpha)\frac{\boldsymbol{M}^k}{\|\boldsymbol{M}^k\|_{\mathrm{F}}}$ delivers this value. Indeed, by the trace property, $\langle \boldsymbol{M}^k, \Delta\rangle = \langle \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top, \alpha \boldsymbol{U}\boldsymbol{V}^\top + (1-\alpha)\frac{\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top}{\|\boldsymbol{M}^k\|_{\mathrm{F}}}\rangle = \alpha \operatorname{tr}\boldsymbol{\Sigma} + (1-\alpha)\|\boldsymbol{M}^k\|_{\mathrm{F}} = \|\boldsymbol{M}^k\|_{\mathrm{F}*}$, which completes the proof.

Proof of lemma 3 follows from eq. (12) with $\alpha = 1$. Indeed, $\|\cdot\|^\dagger_{\mathrm{nuc}} = 1 \cdot \|\cdot\|_{\mathrm{op}} + 0 \cdot \|\cdot\|_{\mathrm{F}}$.

Proof of lemma 5: Since $\|\cdot\|^\dagger = \|\cdot\|^{\dagger\dagger}_{\mathrm{F2}} = \|\cdot\|_{\mathrm{F2}}$, the goal is to reach $\|\boldsymbol{M}^k\|_{\mathrm{F2}} = \alpha\sigma_1 + (1-\alpha)\|\boldsymbol{M}^k\|_{\mathrm{F}}$.

Let us note that $\Delta = \alpha(u_1 v_1^\top) + (1-\alpha)\frac{\boldsymbol{M}^k}{\|\boldsymbol{M}^k\|_{\mathrm{F}}}$ delivers this value. Indeed, by the trace property and singular vectors orthogonality, $\langle \boldsymbol{M}^k, \Delta\rangle = \langle \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top, \alpha u_1 v_1^\top + (1-\alpha)\frac{\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top}{\|\boldsymbol{M}^k\|_{\mathrm{F}}}\rangle = \alpha \operatorname{tr}\operatorname{diag}(\sigma_1, 0, \ldots, 0) + (1-\alpha)\|\boldsymbol{M}^k\|_{\mathrm{F}} = \|\boldsymbol{M}^k\|_{\mathrm{F2}}$, which completes the proof.

Proof of lemma 6: Since $\|\cdot\|^\dagger = \|\cdot\|^{\dagger\dagger}_{\mathrm{KF-k}} = \|\cdot\|_{\mathrm{KF-k}}$, the goal to reach $\|\boldsymbol{M}^k\|_{\mathrm{KF-k}}$.

Let us note that $\Delta = \sum_{i=1}^k u_i v_i^\top$ delivers the value. Indeed, $\langle \boldsymbol{M}^k, \Delta\rangle = \langle \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top, \sum_{i=1}^k u_i v_i^\top\rangle = \sum_{i,j=1}^{r,k}\langle u_i\sigma_i v_i^\top, u_j v_j^\top\rangle = \sum_{i=1}^k \sigma_i = \|\boldsymbol{M}^k\|_{\mathrm{KF-k}}$, which completes the proof.

E TECHNICAL DETAILS OF THE EXPERIMENTS

We used RTX A4000 for CIFAR airbench.