THE NUCLEAR NORM 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 nuclear norm, its affine combinations with other norms, and their corresponding duals to develop a new family of Muon-like algorithms. We complement our theoretical analysis with an extensive empirical study of the algorithms across a wide range of tasks and settings.

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 and AdamW Kingma & Ba (2014), Loshchilov & Hutter (2017), recently proposed Muon has shown promising results on training very large models **??**. Its key difference from earlier algorthims 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). Several other attempts have been since made to construct new algorithms, mainly generalizations of Muon's paradigm, like Scion (Pethick et al. (2025) and Gluon Riabinin et al. (2025)).

Based upon recent theoretical advances that explain some theory behind Muon, Scion and Gluon (Bernstein (2025); Kovalev (2025); Pethick et al. (2025); 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) approach, which is equivalent to a factor to the trust region and the steepest descent under norm constraint approaches, 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.

Table that compares methods, updates, and sends to the proofs?

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 and Assumptions

Warning: the text was copied, while formulas were adapted. We need to rephrase the subsection! We consider the problem of minimizing a differentiable matrix function $F(\cdot)\colon \mathbb{R}^{m\times n} \to \mathbb{R}$

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

To make the theoretical analysis possible in the future, we make three reasonable non-restrictive assumptions. Idea: move gradient assumptions and smoothness to the L-smooth part. Here we need only custom norm and their relation to the Frobenius norm.

Non-Euclidean norm setting and Lipschitz continuous gradient. We assume that matrix space $\mathbb{R}^{m\times n}$ is equipped with a norm $\|\cdot\|\colon \mathbb{R}^{m\times n} \to \mathbb{R}_+$, which possibly does not coincide with the Frobenius, norm $\|\cdot\|_{\mathrm{F}}$. In addition, we assume that the gradient $\nabla f(\cdot)$ is Lipschitz continuous with respect to the norm $\|\cdot\|$, that is, the following inequality holds:

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

where $L > 0$ is the gradient Lipschitz constant, and $\|\cdot\|^{\dagger}\colon \mathbb{R}^{m\times n} \to \mathbb{R}_+$ is the dual norm associated with $\|\cdot\|$, i.e., $\|\boldsymbol{X}\|^{\dagger} = \sup_{\|\boldsymbol{X}'\|\le 1}\langle \boldsymbol{X}, \boldsymbol{X}'\rangle$ for all $\boldsymbol{X} \in \mathbb{R}^{m\times n}$.

2.2 Linear Minimization Oracle and Trust Region

Let us look at the problem from the perspective of linear minimization oracle (lmo) and unconstrained stochastic conditional gradient descent (uSCG) (Pethick et al. (2025)). 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:

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

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 simplify the update defined by the aforementioned equation:

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

In all the work, we define $\boldsymbol{U}$, $\boldsymbol{\Sigma}$, $\boldsymbol{V}^\top$ as components of the singular value decomposition of $\boldsymbol{M}^k$: $\boldsymbol{M}^k = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. We use common notations: $\boldsymbol{U} = [u_1, u_2, \ldots, u_r]$, $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$, and $\boldsymbol{V} = [v_1, v_2, \ldots, v_r]$.

$\|\boldsymbol{M}^k\|_{\mathrm{F}}$ and NSGD

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

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

It is an interesting observation, because in other works (Pethick et al. (2025)), $\|\cdot\|_{\mathrm{F}}$ was used to recover SGD. The difference is in how one states the problem.

$\|\boldsymbol{M}^k\|_{\mathrm{op}}$ **and Muon**

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

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

$\|\boldsymbol{M}^k\|_{\mathrm{nuc}}$ **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 Matrix side of updates, we will discuss how to compute an update efficiently.

$\|\boldsymbol{M}^k\|_{\mathrm{F}*}^\dagger$ **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 of $F*$-family.

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

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

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

$\|\boldsymbol{M}^k\|_{\mathrm{F2}}^\dagger$ **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 of F2-family.

Lemma 5. *When* $\|\cdot\| = \|\cdot\|_{\mathrm{F2}}^\dagger$, *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, which is curious. The implications are significant and discussed in the following sections.

3.1 ALGORITHMS FOR MATRICES $\leftrightarrow$ ALGORITHMS FOR VECTORS

Table that compares Muons and Neons to vector algorithms (NSGD).

4 MATRIX SIDE OF UPDATES

4.1 THEORY

Not only formulas, but also citations! Solutions to efficiently find some parts of SVD:

- Basic SVD (complexity asymptotic!)
- Newton-Schulz (complexity asymptotic!)
- Our job: Power iterations, Randomized SVD and Lanczos

4.2 EXPERIMENTS

Nikolay's task: SVD, RSVD, Lanczos, Newton-Schulz. We use torch and cupy to test it. Options: all possible on torch, or additionally to compare all on cupy. The goal: to find how much do we lose due to cupy

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 Kovalev (2025).

Lemma 6. *To reach the precision* $\min_{k=1\ldots K}\|\nabla f(\boldsymbol{X}_k)\|^\dagger \leq \varepsilon$ *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:*

$$\eta = \mathcal{O}\left(\frac{\varepsilon}{L}\right), \qquad K = \mathcal{O}\left(\frac{L\Delta_0}{\varepsilon^2}\right). \tag{13}$$

In the stochastic case, from Corollary of 2 of Kovalev (2025), we directly get the following result that matches lower bounds, as was noted in Kovalev (2025):

Lemma 7. *To reach the precision* $\mathbb{E}\min_{k=1\ldots K}\|\nabla f(\boldsymbol{X}_k)\|^\dagger \leq \varepsilon$ *by equation 5 under the assumptions Assumption (A3), Assumption (A1), Assumption (A2), it is sufficient to choose the parameters as follows:*

$$\eta = \mathcal{O}\left(\min\left\{\frac{\varepsilon}{L}, \frac{\varepsilon^3}{\rho^2\sigma^2 L}\right\}\right), \qquad \alpha = \mathcal{O}\left(\min\left\{1, \frac{\varepsilon^2}{\rho^2\sigma^2}\right\}\right), \tag{14}$$

$$K = \mathcal{O}\left(\max\left\{\frac{\rho\sigma}{\varepsilon}, \frac{\rho^3\sigma^3}{\varepsilon^3}, \frac{L\Delta_0}{\varepsilon^2}, \frac{L\Delta_0\rho^2\sigma^2}{\varepsilon^4}\right\}\right). \tag{15}$$

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 RANDOMIZED LINEAR LEAST SQUARES

Since the provided by other authors Kovalev (2025); Riabinin et al. (2025) theoretical guarantees are almost norm-independent, we have to test them in practice.

To test the bounds from Kovalev (2025) in practice, we construct the following L-smooth problem:

$$F(\boldsymbol{X}) = \frac{1}{2}\langle(\boldsymbol{X} - \boldsymbol{S}), \boldsymbol{M}(\boldsymbol{X} - \boldsymbol{S})\boldsymbol{N}\rangle \tag{16}$$

where $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, $m = 10$, $n = 10$, $S \in \mathbb{R}^{m\times n}$, $\boldsymbol{M} \in \mathbb{S}_+^m$ and $\boldsymbol{N} \in \mathbb{S}_+^n$ are positive-semidefinite matrices. Spectra of $\boldsymbol{M}$ and $\boldsymbol{N}$ are uniformly distributed in (0, 1) interval.

It it easy to derive the gradient

$$\nabla F(\boldsymbol{X}) = \boldsymbol{M}(\boldsymbol{X} - \boldsymbol{S})\boldsymbol{N}, \tag{17}$$

Let us define γ as $\|\cdot\| \sim \gamma\|\cdot\|_{\mathrm{F}}$, which is asymptotics from the previous section. Then $\|\cdot\|^\dagger \sim \frac{1}{\gamma}\|\cdot\|_{\mathrm{F}}$, as $\|\cdot\|_{\mathrm{F}}^\dagger = \|\cdot\|_{\mathrm{F}}$. Hence, $\|\cdot\| \sim \gamma^2\|\cdot\|^\dagger$

Then $\|\nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{Y})\|^\dagger = \|\boldsymbol{M}(\boldsymbol{X} - \boldsymbol{Y})\boldsymbol{N}\|^\dagger \leq \|\boldsymbol{M}\|^\dagger\|\boldsymbol{N}\|^\dagger\|\boldsymbol{X} - \boldsymbol{Y}\|^\dagger \sim \|\boldsymbol{M}\|^\dagger\|\boldsymbol{N}\|^\dagger\gamma^2\|\boldsymbol{X} - \boldsymbol{Y}\|$, and $L \sim \gamma^2\|\boldsymbol{M}\|^\dagger\|\boldsymbol{N}\|^\dagger$.

Now for the known norms:

Frobenius (NSGD): $L = \|M\|_{\mathrm{F}}\|N\|_{\mathrm{F}}$

Spectral (Muon): $\gamma \sim \frac{2}{\sqrt{n}} \implies L \sim \frac{2}{n}\|M\|_{\mathrm{nuc}}\|N\|_{\mathrm{nuc}}$

Nuclear (Neon): $\gamma \sim \sqrt{n} \implies L \sim n\|M\|_{\mathrm{op}}\|N\|_{\mathrm{op}}$

We take learning rate η and iteration number K from eq. (13).

We are interested to get $\varepsilon_2 =$ 1e-1 precision in $\|\cdot\|_{\mathrm{op}}$. Hence, we target $\varepsilon = \varepsilon_2$ for the nuclear norm (Neon), $\varepsilon = \frac{n}{2}$ for the spectral norm (Muon), and $\varepsilon = \frac{\sqrt{n}}{2}$ for the Frobenius norm (NSGD).

An independent experiment To make the problem more real-world, we do not use known information about smoothness. We run NormalizedSGD, Muon, F-Muon with $\alpha = 1/2$, Neon, and F-Neon with $\alpha = 1/2$ for 100 000 iterations with learning rate = 1e-3.

6.2 LOGISTIC REGRESSION

6.3 BENCHMARKS

CNN benchmark NanoGPT benchmark

7 RELATED WORK

What to include: Muon and its upgrades (say that the same can be done with Neon), alternatives to Newton-Schulz, nuSAM and so on.

8 CONCLUSION

- Future work: (L0, L1)-smoothness, probably extrapolation (to make dependency on eps milder, but we don't to spend memory here so never mind), hyperparameter-free (Bernstein2023)

REFERENCES

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

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

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 2013.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 2016.

Eduard Gorbunov, Dmitry Kovalev, and Peter Richtárik. Linearly converging error compensated sgd. *Advances in Neural Information Processing Systems*, 2020.

Robert M Gower et al. Sgd: General analysis and improved rates. *International Conference on Machine Learning*, 2019.

Samuel Horváth, Dmitry Kovalev, and Peter Richtárik. Stochastic recursive momentum for nonconvex optimization. *arXiv preprint arXiv:2302.07731*, 2023.

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.

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

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.

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.

Ruoyu Sun et al. Momentum methods for stochastic optimization: A survey and new results. *arXiv preprint arXiv:2302.06675*, 2023.

9 Appendix

9.1 Formal assumptions

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:

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

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., $\|\boldsymbol{X}\|_{\mathrm{F}} = \sqrt{\langle \boldsymbol{X}, \boldsymbol{X}\rangle} = \sqrt{\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X})}$. These assumptions have been widely adopted for the analysis of many stochastic gradient optimization algorithms (Ghadimi & Lan, 2013; Ghadimi et al., 2016; Cutkosky et al., 2020; Sun et al., 2023; Horváth et al., 2023; Gorbunov et al., 2020).

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 (A3) 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:

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

where $\rho > 0$ is a positive constant. Note that such a constant always exists due to the norm equivalence theorem, which always holds in the finite-dimensional space $\mathbb{R}^{m\times n}$. We recount ρ for different norms $\|\cdot\|$ in the appendix.

9.2 NORMS $\|\cdot\|^{\dagger}_{\mathrm{F}*}$ AND $\|\cdot\|^{\dagger}_{\mathrm{F2}}$

Here we provide the derivation of these norms and plot them in the case of 2×2 matrices.

9.3 UPDATES DERIVATIONS

Derivation of eq. (7) follows from eq. (10) with $\alpha = 1$. Indeed, $\|\cdot\|^{\dagger}_{\mathrm{op}} = 1 \cdot \|\cdot\|_{\mathrm{nuc}} + 0 \cdot \|\cdot\|_{\mathrm{F}}$.

Derivation of eq. (10): Since $\|\cdot\|^{\dagger} = \|\cdot\|^{\dagger\dagger}_{\mathrm{F}*} = \|\cdot\|_{\mathrm{F}*}$, the goal is to reach $\|\boldsymbol{M}^k\|_{\mathrm{F}*} = \alpha \operatorname{tr} \boldsymbol{\Sigma} + (1 - \alpha)\|\boldsymbol{M}^k\|_{\mathrm{F}}$.

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.

Derivation of eq. (8) follows from eq. (12) with $\alpha = 1$. Indeed, $\|\cdot\|^{\dagger}_{\mathrm{nuc}} = 1 \cdot \|\cdot\|_{\mathrm{op}} + 0 \cdot \|\cdot\|_{\mathrm{F}}$.

Derivation of eq. (12): 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.

9.4 TECHNICAL DETAILS OF THE EXPERIMENTS

Note: nuSAM and their update. They had no Lanczos