

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

**Alexey Kravatskiy**

*Moscow Institute of Physics and Technology (MIPT)*

KRAVTSKII.AI@PHYSTECH.EDU

**Ivan Kozyrev**

*Moscow Institute of Physics and Technology (MIPT)*

*Marchuk Institute of Numerical Mathematics*

KOZYREV.IN@PHYSTECH.EDU

**Nikolai Kozlov**

*Moscow Institute of Physics and Technology (MIPT)*

KOZLOV.NA@PHYSTECH.EDU

**Alexander Vinogradov**

*Moscow Institute of Physics and Technology (MIPT)*

VINOGRADOV.AM@PHYSTECH.EDU

**Daniil Merkulov**

*Moscow Institute of Physics and Technology (MIPT)*

*Skoltech, HSE, AI4Science*

DANIIL.MERKULOV@PHYSTECH.EDU

**Ivan Oseledets**

*AIRI, Skoltech*

*Marchuk Institute of Numerical Mathematics*

I.OSELEDETS@SKOLTECH.RU

## Abstract

In this article, we explore the use of various 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 duals to the Ky Fan  $k$ -norms to introduce a family of Muon-like algorithms we name *Fanions*, which happen to be similar to Dion. Then working with the duals of convex combinations of the Ky Fan  $k$ -norms and the Frobenius norm or the  $l_\infty$  norm, we construct the families of *F-Fanions* and *S-Fanions* respectively. Their most prominent members are *F-Muon* and *S-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 F-Muon and S-Muon are always on par with Muon, while on fine-tuning of NanoGPT and synthetic linear least squares they are even better than vanilla Muon optimizer.

## 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 and Ba, 2014) and AdamW (Loshchilov and Hutter, 2017), recently proposed Muon (Jordan et al., 2024b) has shown promising results on training very large models (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 a theoretical perspective, a key innovation of Muon was its principled derivation of the update rule, which emerged as the solution to an optimization problem constrained by the RMS-to-RMS norm (scaled version of the spectral norm) (Bernstein, 2025).

Motivated by the success of Muon, many generalizations and variations of it were proposed. Among the notable ones are Scion (Pethick et al., 2025b), Dion (Ahn et al., 2025) and Gluon (Riabini et al., 2025). Those works try to explain Muon’s efficiency and establish convergence bounds. One central question, however, remains unanswered:

*In deriving Muon’s update step, why should one constrain by the spectral or any other operator norm? How would alternative norms affect performance and computational cost?*

In this article, we tackle this question by actually showing that there are many viable non-operator norms. We leverage the family of norms dual to Ky Fan  $k$ -norms to derive a new family of **Fanions**, algorithms with low-rank updates. This approach theoretically explains the backbone of Dion’s update (Ahn et al., 2025) and generalizes the memory-motivated application of the nuclear norm to Sharpness-Aware Minimization (Pethick et al., 2025a). As it was done with Muon, we come up with an effective procedure for computing Fanions’ updates. Lanczos algorithm, which is described and compared with its competitors in Section ??, is the most operation-effective algorithm, which, however, for now lacks an effective GPU- and PyTorch-friendly implementation.

Working with duals to conic combinations of dual norms, we construct the families of **F-Fanions** and **S-Fanions**, which are hybrids of Muon and NormalizedSGD and SignSGD, respectively.

Then we compare the performances of the algorithm families on various model and real-world problems:

- Synthetic least squares experiment Section ??
- CIFAR-10 airbench (Keller, 2023)
- Pre-training NanoGPT on FineWeb dataset (Jordan et al., 2024a)
- Fine-tuning NanoGPT on Tiny Stories (Eldan and Li, 2023)

Our experiments reveal important insights into the role of matrix norms in optimization. First, we show on the example of Neon, the one-rank Fanion, that not every LMO-based algorithm is effective, despite the same asymptotics in the general bounds of Kovalev (2025) and Riabini et al. (2025). This suggests that existing theoretical guarantees should be reworked to explain empirical performance.

Most notably, our experiments on real-world tasks demonstrate that the choice of underlying matrix norm is remarkably flexible. On CIFAR-10 airbench, properly-tuned F-Muon and S-Muon achieve  $94.02 \pm 0.13\%$  and  $94.03 \pm 0.13\%$  accuracy, matching Muon’s  $94.01 \pm 0.13\%$  performance. On NanoGPT pre-training, F-Muon achieves 3.281 cross-entropy loss, while fully-tuned Muon achieves 3.279. Finally, S-Muon matches Muon on fine-tuning of NanoGPT on Tiny Stories, while F-Muon is far more resistant to the learning rate choice than Muon. These results show that Muon-like algorithms can maintain competitive performance even when the underlying norm constraint is significantly modified, answering affirmatively the central question posed above. Moreover, the tools from

Section ?? give the researchers an unheard-of flexibility in designing algorithms that do not have to be modifications of Muon.

## 2. Preliminaries: Linear Minimization Oracle Framework

Training a neural network is essentially an optimization of a function of several weight matrices and a few vectors. Let us start by considering the problem of minimizing a differentiable function of a *single* matrix:

$$F(\cdot): \mathbb{R}^{m \times n} \rightarrow \mathbb{R}, \quad F(\mathbf{X}) \rightarrow \min_{\mathbf{X} \in \mathbb{R}^{m \times n}} \quad (1)$$

We equip the matrix space  $\mathbb{R}^{m \times n}$  with a standard dot product  $\langle \cdot, \cdot \rangle \rightarrow \mathbb{R}$  and a norm  $\|\cdot\|: \mathbb{R}^{m \times n} \rightarrow \mathbb{R}_+$ , which does not have to coincide with the Frobenius norm  $\|\cdot\|_F$ . The dual norm  $\|\cdot\|^\dagger: \mathbb{R}^{m \times n} \rightarrow \mathbb{R}_+$  that is associated with  $\|\cdot\|$  is defined as

$$\|\mathbf{G}\|^\dagger = \sup_{\mathbf{D} \in \mathbb{R}^{m \times n}: \|\mathbf{D}\| \leq 1} \langle \mathbf{G}, \mathbf{D} \rangle. \quad (2)$$

Such problems can be solved with an iterative algorithm based on the Linear Minimization Oracle (LMO):

$$\text{LMO}(\mathbf{M}^k) \in \arg \min_{\mathbf{D} \in \mathcal{S}} \langle \mathbf{M}^k, \mathbf{D} \rangle, \quad (3)$$

where  $\mathbf{M}^k$  is a gradient (or a momentum) of  $F$  in  $\mathbf{X}^k$  and  $\mathcal{S} \subset \mathbb{R}^{m \times n}$  is some set. The update of the algorithm is defined as follows:

$$\mathbf{X}^{k+1} = \mathbf{X}^k + \gamma_k \text{LMO}(\mathbf{M}^k). \quad (4)$$

We are particularly interested in the case when  $\mathcal{S}$  is a unit ball in the norm  $\|\cdot\|$ :

$$\mathcal{S} = \mathcal{B}_{\|\cdot\|} = \{\mathbf{D} \in \mathbb{R}^{m \times n} \mid \|\mathbf{D}\| \leq 1\}.$$

In this case,

$$\arg \min_{\mathbf{D} \in \mathcal{S}} \langle \mathbf{M}^k, \mathbf{D} \rangle = -\{\mathbf{D} \in \mathcal{B}_1 \mid \langle \mathbf{M}^k, \mathbf{D} \rangle = \|\mathbf{M}^k\|^\dagger\},$$

and update for  $\mathbf{X}^k$  in Equation (4) simplifies to

$$\mathbf{X}^{k+1} = \mathbf{X}^k - \gamma_k \{\mathbf{D} \in \mathcal{B}_1 \mid \langle \mathbf{M}^k, \mathbf{D} \rangle = \|\mathbf{M}^k\|^\dagger\}. \quad (5)$$

Using this formula it is easy to compute updates for algorithms induced by various norms  $\|\cdot\|$ :

**Frobenius norm and Normalized SGD** When the norm  $\|\cdot\|$  is the Frobenius norm  $\|\cdot\|_F$ , Equation (5) turns into

$$\mathbf{X}^{k+1} = \mathbf{X}^k - \eta_k \frac{\mathbf{M}^k}{\|\mathbf{M}^k\|_F}, \quad (6)$$

which recovers Normalized SGD (NSGD).

**Spectral norm and Muon** When the norm is the spectral norm  $\|\cdot\|_2$ , we get

$$\mathbf{X}^{k+1} = \mathbf{X}^k - \eta_k \mathbf{U} \mathbf{V}^\top, \quad (7)$$

which is Muon without the  $\sqrt{m/n}$  factor. Here,  $\mathbf{M}^k = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top$  is the Singular Value Decomposition (SVD) of  $\mathbf{M}^k$  ( $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r]$ ,  $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_r)$ , and  $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r]$ ). Muon can be recovered by taking the RMS-to-RMS operator norm  $\sqrt{\frac{n}{m}} \|\cdot\|_2$ .

**Chebyshev norm and SignSGD** When the norm is the Chebyshev norm  $\|\cdot\|_C$ , we get

$$\mathbf{X}^{k+1} = \mathbf{X}^k - \eta_k \text{sign}(\mathbf{M}^k), \quad (8)$$

which recovers SignSGD (Bernstein et al., 2018). Here,  $\text{sign}(\mathbf{M}^k)$  denotes the element-wise sign function. SignSGD is particularly notable for its communication efficiency in distributed training, as it compresses gradients to 1-bit per parameter.

### 3. Duals to Ky Fan Norms Instead of the Spectral Norm

#### 3.1. $\|\mathbf{M}^k\|_*$ and Neon

After considering  $\|\mathbf{M}^k\|_F$  and  $\|\mathbf{M}^k\|_2$ , it is natural to look at the nuclear norm  $\|\mathbf{M}^k\|_*$ .

**Lemma 1** When  $\|\cdot\| = \|\cdot\|_*$ , Equation (5) turns into:

$$\mathbf{X}^{k+1} = \mathbf{X}^k - \eta_k \mathbf{u}_1 \mathbf{v}_1^\top \quad (9)$$

**Proof** Since  $\|\cdot\|^\dagger = \|\cdot\|_2$ , the goal is to reach  $\|\mathbf{M}^k\|_2 = \sigma_1$ .

Let us note that  $\Delta = \mathbf{u}_1 \mathbf{v}_1^\top$  delivers this value. Indeed,  $\|\Delta\|_* = 1$  and by the trace property and orthogonality of the singular vectors,

$$\langle \mathbf{M}^k, \Delta \rangle = \langle \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top, \mathbf{u}_1 \mathbf{v}_1^\top \rangle = \text{tr} \text{diag}(\sigma_1, 0, \dots, 0) = \|\mathbf{M}^k\|_2,$$

which completes the proof. ■

We name the derived algorithm *Neon*. In Section ??, we will discuss how to compute the Neon's update.

#### 3.2. $\|\mathbf{M}^k\|_{\text{KF-}k}^\dagger$ and Muon, Neon, and Centralized Dion without error feedback

Neon's and Muon's updates seem to be complete opposites: one has rank one, while the other is full-rank. It would be interesting to derive algorithms with updates of intermediate ranks.

##### 3.2.1. SCHATTEN NORMS

Cesista (2025) considered Schatten- $p$  norms:

$$\|\mathbf{M}^k\|_{S_p} = \left( \sum_{i=1}^{\min(m,n)} \sigma_i^p \right)^{1/p},$$

which produce the update

$$\mathbf{X}^{k+1} = \mathbf{X}^k - \eta_k \mathbf{U} \frac{\text{diag}(\sigma_1^{q-1}, \dots, \sigma_{\min(m,n)}^{q-1})}{\left(\sum_{i=1}^{\min(m,n)} \sigma_i^q\right)^{\frac{q-1}{q}}} \mathbf{V}^\top$$

for  $p$  and  $q$  such that:  $\frac{1}{p} + \frac{1}{q} = 1$ . This formula recovers Neon when  $p \rightarrow 1$  provided that  $\sigma_1 > \sigma_2$ , which is true on real data; NSGD when  $p = 2$ , and Muon when  $p \rightarrow \infty$ .

However, Schatten norms do not fill the gap rank: indeed, when  $p > 1$ , the rank of the update is full, while when  $p = 1$  ( $p \rightarrow 1$ ) it is one. Moreover, it is not clear how to calculate the update for  $p \neq 1, 2, \infty$ : it seems one has to know all  $\sigma_i$  to compute the update, which makes the problem as hard as the full SVD.

### 3.2.2. KY FAN NORMS

There is another family of matrix norms, which might help us: Ky Fan norms. For  $k \in \{1, \dots, \min(m, n)\}$ , the Ky Fan  $k$ -norm  $\|\cdot\|_{\text{KF-}k}$  is  $\sum_{i=1}^k \sigma_i$ , i.e. the sum of the  $k$  largest singular values of the matrix. Special cases of the Ky Fan  $k$ -norm are the Ky Fan 1-norm, which is the spectral norm, and the Ky Fan  $\min\{m, n\}$ -norm, which is the nuclear norm.

Let us derive the update for an arbitrary  $k$ . Since  $\|\cdot\|_{\text{KF-}k}^\dagger = \max\{\frac{1}{k}\|\cdot\|_*, \|\cdot\|_2\}$  (see [Bhatia \(2013\)](#), p. 96), the goal is to reach  $\max\{\frac{1}{k} \sum_{i=1}^{\min(m,n)} \sigma_i, \sigma_1\}$ .  $\Delta = \mathbf{u}_1 \mathbf{v}_1^\top$  from Neon delivers  $\sigma_1$ , while  $\Delta = \frac{1}{k} \sum_{i=1}^{\min(m,n)} \mathbf{u}_i \mathbf{v}_i^\top$  from Muon (with an extra factor of  $1/k$ ) delivers  $\frac{1}{k} \sum_{i=1}^{\min(m,n)} \sigma_i$ . Thus, the update is either

$$\mathbf{X}^{t+1} = \mathbf{X}^t - \eta_t \mathbf{u}_1 \mathbf{v}_1^\top \text{ or } \mathbf{X}^{t+1} = \mathbf{X}^t - \frac{\eta_t}{k} \mathbf{U} \mathbf{V}^\top,$$

depending on which one better minimizes the linear function  $L(\mathbf{X}) \equiv F(\mathbf{X}^t) + \mathbf{M}^t(\mathbf{X} - \mathbf{X}^t)$ .

The rank gap is not closed by the Ky Fan norms because the obtained update is either one-rank or full-rank.

### 3.2.3. DUALS TO KY FAN NORMS

While Schatten norms are closed under dualization and  $\|\mathbf{M}^k\|_{S_p}^\dagger = \|\mathbf{M}^k\|_{S_q}$ , for Ky Fan norms it is not generally the case.  $k = 1$  and  $k = \min(m, n)$  are exceptional: for  $k = 1$  the dual norm is  $\|\cdot\|_*$  with  $k = \min(m, n)$  and for  $k = \min(m, n)$  the dual norm is  $\|\cdot\|_2$  with  $k = 1$ . For each other  $k$ ,  $\|\cdot\|_{\text{KF-}k} \neq \|\cdot\|_{\text{KF-}k'}$  for any  $k'$ :  $k' = \min(m, n)$  and  $k' = 1$  correspond to the already discussed cases, while for other  $k'$  one can change  $\|\mathbf{M}^k\|_{\text{KF-}k'}$  by moving a small value between  $\sigma_{k'}$  and  $\sigma_{k'+1}$ . During this change,  $\|\mathbf{M}^k\|_2$  and  $\|\mathbf{M}^k\|_*$  will remain constant, hence  $\|\mathbf{M}^k\|_{\text{KF-}k}$  will not change as well.

**Lemma 2** When  $\|\cdot\| = \|\mathbf{M}^t\|_{\text{KF-}k}^\dagger$ , Equation (5) turns into:

$$\mathbf{X}^{t+1} = \mathbf{X}^t - \eta \sum_{i=1}^k \mathbf{u}_i \mathbf{v}_i^\top \quad (10)$$

**Proof** Since  $\|\cdot\|^\dagger = \|\cdot\|_{\text{KF-}k}^{\dagger\dagger} = \|\cdot\|_{\text{KF-}k}$ , the goal is to reach  $\|\mathbf{M}^t\|_{\text{KF-}k}$ .

Let us note that  $\Delta = \sum_{i=1}^k \mathbf{u}_i \mathbf{v}_i^\top$  delivers the value. Indeed,

$$\langle \mathbf{M}^t, \Delta \rangle = \langle \mathbf{U} \Sigma \mathbf{V}^\top, \sum_{i=1}^k \mathbf{u}_i \mathbf{v}_i^\top \rangle = \sum_{i,j=1}^{r,k} \langle \mathbf{u}_i \sigma_i \mathbf{v}_i^\top, \mathbf{u}_j \mathbf{v}_j^\top \rangle = \sum_{i=1}^k \sigma_i = \|\mathbf{M}^t\|_{\text{KF-}k},$$

which completes the proof. ■

These updates constitute the family of *Fanions*, LMO-based algorithms under the  $\|\mathbf{M}^t\|_{\text{KF-}k}^\dagger$  norms. The algorithm for a particular  $k$  we will call *Fanion- $k$* . Thus, Muon is a Fanion- $\min\{m, n\}$ , while Neon is a Fanion-1. Moreover, the unsharded version of the rank- $r$  Dion (Algorithm 1 from [Ahn et al. \(2025\)](#)) without the error feedback and without scaling of the update is actually a Fanion- $r$  (see [Pethick \(2025\)](#), where Dion is written down in a notation more similar to ours).

## References

- Kwangjun Ahn, Byron Xu, Natalie Abreu, Ying Fan, Gagik Magakyan, Pratyusha Sharma, Zheng Zhan, and John Langford. Dion: Distributed orthonormalized updates, 2025. URL <https://arxiv.org/abs/2504.05295>.
- Jeremy Bernstein. Deriving muon, 2025. URL <https://jeremybernste.in/writing/deriving-muon>.
- Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International conference on machine learning*, pages 560–569. PMLR, 2018.
- Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.
- Franz Louis Cesista. Steepest Descent under Schatten- $p$  Norms, February 2025. URL <https://leloykun.github.io/ponder/steepest-descent-schatten-p/>.
- Ronen Eldan and Yuanzhi Li. Tinstories: How small can language models be and still speak coherent english?, 2023. URL <https://arxiv.org/abs/2305.07759>.
- Keller Jordan, Jeremy Bernstein, Brendan Rappazzo, @fernbear.bsky.social, Boza Vlado, You Jiacheng, Franz Cesista, Braden Koszarsky, and @Grad62304977. modded-nanogpt: Speedrunning the nanogpt baseline, 2024a. URL <https://github.com/KellerJordan/modded-nanogpt>.
- Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024b. 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. Understanding dion, 2025. URL <https://pethick.dk/posts/2025-08-18-understanding-dion/>.
- Thomas Pethick, Parameswaran Raman, Lenon Minorics, Mingyi Hong, Shoham Sabach, and Volkan Cevher.  $\nu$ SAM: 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.
- Artem Riabinin, Egor Shulgin, Kaja Grunkowska, 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.