

# FROM MUON TO NEON: INTRODUCING NUCLEAR NORM TO LARGE MATRICES

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

In this paper, we develop a new algorithm for optimization of functions of weight matrices, which are typical for training large language models. Changing spectral norm, which was used to derive Muon, to nuclear norm, we pose a new optimization problem for an update matrix, solution of which defines a novel algorithm we name Neon. To make it feasible, we use Lanczos algorithm to find a required step. After providing theoretical guarantees of Neon convergence, we compare performances of Neon, Muon, and Adam on training multilayer perceptron, convolutional neural network and NanoGPT.

## 1 IDEA

The goal of the project is to make variations on Muon to speed it up. Recently, authors of Bernstein & Newhouse (2024) have proposed to derive update step for an optimizer as the solution of the certain optimization problem. This approach can be utilized to derive Muon Jordan et al. (2024), a novel algorithm for fast training of neural networks. Instead of using spectral norm, used to derive Muon, we utilize nuclear norm to produce a new optimization algorithm.

### 1.1 PROBLEM (PROJECT DESCRIPTION)

In this subsection, we provide a more detailed description of our idea and formulate it as a mathematical problem. The authors of Bernstein & Newhouse (2024) suggest obtaining the update step as a solution to the optimization problem:

$$\langle g, \delta w \rangle + \lambda \|\delta w\|^2 \rightarrow \min_{\delta w}, \quad (1)$$

where  $w$  is the weight vector,  $g$  is a gradient-like vector (e.g., obtained via momentum SGD), and  $\|\cdot\|$  represents a certain norm. Many popular optimizers, such as Adam (with exponential moving average disabled) and vanilla SGD, can be cast within this framework Bernstein & Newhouse (2024).

In large language models, most weights are structured as matrices, which offers additional opportunities for optimization. Let  $W$  be the weight matrix of a linear layer, and  $G$  be a gradient-like matrix. Then, the update step  $\delta W$  can be obtained as a solution to the optimization problem:

$$\langle G, \delta W \rangle + \lambda \|\delta W\|^2 \rightarrow \min_{\delta W}, \quad (2)$$

where  $\|\cdot\|$  denotes a certain matrix norm. By setting this norm to the RMS-to-RMS norm (a scaled version of the spectral norm), we recover the Muon optimizer Bernstein (2025); Bernstein & Newhouse (2024) with an update step defined by:

$$\delta W = -\frac{1}{\lambda} \sqrt{\frac{n}{m}} UV^T, \quad (3)$$

where  $m$  is the input dimension of the layer,  $n$  is the output dimension, and  $U$  and  $V$  are obtained from the singular value decomposition of the gradient matrix  $G = U\Sigma V$ .

Motivated by the recent achievements of the Muon optimizer (e.g., Liu et al. (2025)), we consider alternative choices of norms, specifically the nuclear norm  $\|\cdot\|_*$  and a custom  $F^*$  norm, given by

$$\|X\|_{F^*}^2 = \frac{\|X\|_F^2 + \|X\|_*^2}{2}, \quad (4)$$

where  $\|\cdot\|_F$  denotes the Frobenius norm.

Using the nuclear norm in equation 2 leads to a rank-one update of the weight matrices:

$$\delta W = -\frac{1}{2\lambda} u_1 \sigma_1 v_1^T, \quad (5)$$

where  $\sigma_1$  is the largest singular value, and  $u_1$  and  $v_1$  are the corresponding singular vectors. We expect one iteration of this method to be significantly faster than one iteration of Muon.

Another choice is the  $F^*$  norm. With this choice, equation 2 yields

$$\delta W = -\frac{1}{\lambda} U D V^T \quad (6)$$

with  $D = \text{diag}(d_i)$ , where  $d_i = [\sigma_i - \tau]_+$ , and  $\tau$  is given by

$$\sum_i [\sigma_i - \tau]_+ = \tau. \quad (7)$$

We anticipate that the method with this update step will perform well with large batch sizes.

In this article we show how one can quickly compute weight updates defined by equation 5 or equation 6. Then we finalize the method by adding momentum and test their performance against those of Muon at training multilayer perceptron and transformer. The results will be fast algorithm, which we will convert into a new optimizer classes for PyTorch, as was done with Muon.

## 2 DERIVATION OF UPDATE RULES

**Lemma 1.** *Let  $G \in \mathbb{R}^{m \times n}$  and  $\lambda > 0$ . Then, the following optimization problem*

$$f(\delta W) = \langle G, \delta W \rangle + \lambda \|\delta W\|_*^2 \rightarrow \min_{\delta W}$$

*has solution*

$$\delta W = -\frac{1}{2\lambda} u_1 \sigma_1 v_1^T,$$

*where  $\sigma_1$  is the largest singular value of  $G$ , and  $u_1$  and  $v_1$  are the corresponding singular vectors.*

*Proof.* Let us denote  $r = \min\{m, n\}$ . Then by Von Neumann's trace inequality,

$$|\langle G, \delta W \rangle| \leq \sum_{i=1}^r \sigma_i(G) \sigma_i(\delta W) \Rightarrow \langle G, \delta W \rangle \geq -\sum_{i=1}^r \sigma_i(G) \sigma_i(\delta W).$$

Thus, expressing nuclear norm through singular values, we can write down

$$f(\delta W) \geq -\sum_{i=1}^r \sigma_i(G) \sigma_i(\delta W) + \lambda \left( \sum_{i=1}^r \sigma_i(\delta W) \right)^2 \geq \min_{d_1, \dots, d_r \geq 0} -\sum_{i=1}^r \sigma_i(G) d_i + \lambda \left( \sum_{i=1}^r d_i \right)^2.$$

By Karush-Kuhn-Tucker theorem, necessary conditions of minimum are

$$\left( d_i \geq 0 \text{ and } -\sigma_i(G) + 2\lambda \sum_{j=1}^r d_j = 0 \right) \text{ or } \left( d_i = 0 \text{ and } -\sigma_i(G) + 2\lambda \sum_{j=1}^r d_j \geq 0 \right) \quad i = 1, \dots, r.$$

These conditions simplify to

$$\sum_{i \in S} d_i = \sigma_1(G), \quad \begin{cases} d_i \geq 0 & \text{if } \sigma_i(G) = \sigma_1(G), \\ d_i = 0 & \text{otherwise.} \end{cases}$$

All points satisfying those conditions deliver minimum, and

$$f(\delta W) \geq -\frac{\sigma_1^2(G)}{4\lambda}.$$

Now let

$$\delta W^* = -\frac{1}{2\lambda} u_1 \sigma_1(G) v_1^T.$$

Inserting it to  $f(\delta W)$  gives

$$f(\delta W^*) = -\frac{\sigma_1(G)^2}{2\lambda} + \frac{\sigma_1(G)^2}{4\lambda} = -\frac{\sigma_1(G)^2}{4\lambda}$$

This matches the derived lower bound. Thus,  $\delta W^*$  minimizes  $f(\delta W)$ .  $\square$

**Lemma 2.** Let  $G \in \mathbb{R}^{m \times n}$ ,  $r = \min\{m, n\}$  and  $\lambda > 0$ . Then, the following optimization problem

$$f(\delta W) = \langle G, \delta W \rangle + \lambda \|\delta W\|_{F^*}^2 \rightarrow \min_{\delta W},$$

where  $\|\cdot\|_{F^*}$  is defined in equation 4 has solution

$$\delta W = -\frac{1}{\lambda} U D V^T \quad (8)$$

with  $D = \text{diag}(d_i)$ , where  $d_i = [\sigma_i - \tau]_+$ , and  $\tau$  is given by

$$\sum_{i=1}^r [\sigma_i - \tau]_+ = \tau. \quad (9)$$

*Proof.* Analogously to the proof of Lemma 1, we can use Von Neumann's trace inequality to write down

$$\begin{aligned} f(\delta W) &\geq -\sum_{i=1}^r \sigma_i(G) \sigma_i(\delta W) + \frac{\lambda}{2} \left( \sum_{i=1}^r \sigma_i(\delta W) \right)^2 + \frac{\lambda}{2} \sum_{i=1}^r \sigma_i^2(\delta W), \\ f(\delta W) &\geq \frac{1}{\lambda} \min_{d_1, \dots, d_r \geq 0} -\sum_{i=1}^r \sigma_i(G) d_i + \frac{1}{2} \left( \sum_{i=1}^r d_i \right)^2 + \frac{1}{2} \left( \sum_{i=1}^r d_i^2 \right), \end{aligned} \quad (10)$$

By Karush-Kuhn-Tucker theorem, necessary conditions of minimum are

$$\left( d_i \geq 0 \text{ and } -\sigma_i(G) + \sum_{j=1}^r d_j + d_i = 0 \right) \text{ or } \left( d_i = 0 \text{ and } -\sigma_i(G) + \sum_{j=1}^r d_j + d_i \geq 0 \right) \quad i = 1, \dots, r.$$

Denoting  $\tau = \sum_{i=1}^r d_i$  gives  $d_i = [\sigma_i(G) - \tau]_+$ , where  $\tau$  satisfies

$$\sum_{i=1}^n [\sigma_i(G) - \tau]_+ = \tau. \quad (11)$$

Inserting found minimum point into equation 10 yields

$$f(\delta W) \geq -\sum_{i=1}^r d_i(\tau + d_i) + \frac{\tau^2}{2\lambda} + \frac{1}{2\lambda} \sum_{i=1}^r d_i^2 = -\frac{1}{2\lambda} \left( \tau^2 + \sum_{i=1}^r d_i^2 \right).$$

Now let

$$\delta W^* = -\frac{1}{\lambda} U D V^T \quad (12)$$

with  $D = \text{diag}(d_i)$ . Inserting it to  $f(\delta W)$  gives

$$f(\delta W^*) = -\sum_{i=1}^r d_i(\tau + d_i) + \frac{\tau^2}{2\lambda} + \frac{1}{2\lambda} \sum_{i=1}^r d_i^2 = -\frac{1}{2\lambda} \left( \tau^2 + \sum_{i=1}^r d_i^2 \right).$$

This matches the derived lower bound. Thus,  $\delta W^*$  minimizes  $f(\delta W)$ .  $\square$

### 3 QUALITY METRICS

1. The derivation is theoretically solid
2. The numerical procedure used to compute a step is grounded and has estimated time overhead (say, in FLOPS)
3. The code with Neon trains MLP and CNN (and NanoGPT, but it's a bonus) less than 3 times slower than Adam
4. Instruction of setting the parameters of the algorithm are presented and justified
5. The announced article has full structure (Abstract, Introduction, Theory, Experiments, Conclusion, Appendix)
6. If results are positive, it is written with NeurIPS template.

### 4 PRELIMINARY PLAN

#### Week April 28 - May 4

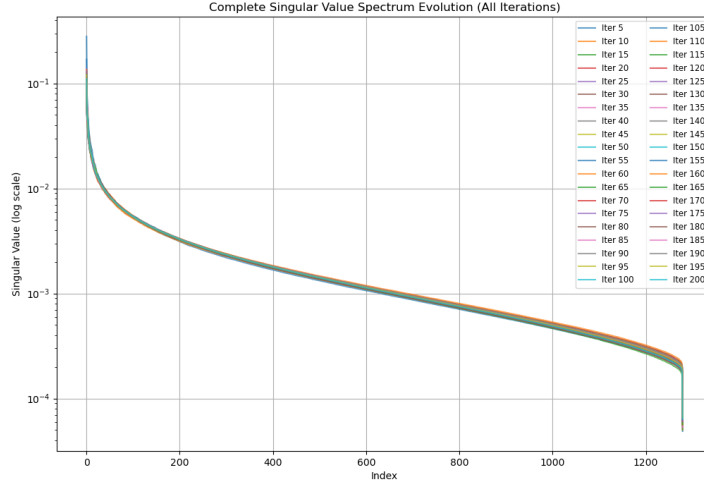
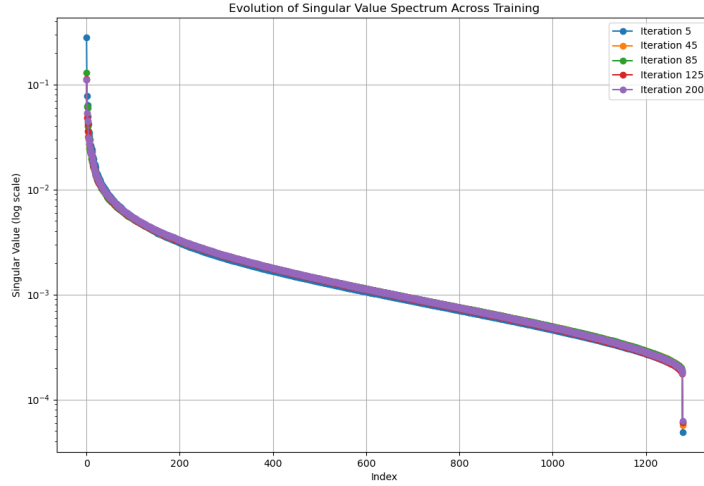
- For Alexey: solve how to tune the algorithms for MLP and CNN, try formulating theory (and an appropriate model of the problem) why Muon and Neon are so successful, and create the drafts of the proofs. Register at NeurIPS site.
- For Ivan: write the theory for an update from the algebra point of view (as for an article)
- For Nikolay: write the theory for computing an update, and implement the method, if required
- For Alexander: reproduce results of Jordan on NanoGPT and ResNet (CIFAR-10), learn to train both models with Neon.

#### Week May 5 - May 11

- For Alexey: finalize the proofs. Verify them via small experiments on MLP and CNN. Write with Alexander Experiments for the article.
- For Ivan: join Nikolay to finalize algebra part of the article. Estimate FLOPS, memory and other overheads (produce O bounds)
- For Nikolay: write a draft of the poster (before May 6), and work with Ivan
- For Alexander: aggressively test algorithms, prove that Neon outperforms competitors and prepare the results for the article.
- For everybody: write and edit the article
- May 11: submit an abstract to NeurIPS.
- May 12-14: the article is being polished.
- May 15: the article must be sent.

### 5 PROTOTYPING PHASE REPORT

1. Update rule is derived, see idea
2. Update rule methods are tested: power iteration vs Lanczos (see 1)
3. Recorded the distribution of singular values of gradients during NanoGPT training (see Figures 1, 1).
4. NanoGPT is tested on Muon and Adam. For now, Neon (rank-1 version) does not converge (see Figures 3 4)  
The pictures show the best results achieved so far. The experiments were conducted with two 4090 24GB GPUs for nanotgpt-large on the tiny stories dataset.
5. Neon (rank-1 version), Muon, AdamW and SGD are compared on MLP and CNN (see Figures 5, 6, 7, and 8). All methods work correctly, but again there is the problem with which one is the fastest (for now, it's SGD).

Figure 1: Singular values of  $50257 \times 1280$  layer via 200 iterationsFigure 2: Singular values of of  $50257 \times 1280$  layer for 5-th,45-th, 65-th,175-th and 200-th iteration

## REFERENCES

- 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.
- 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/>.
- 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.

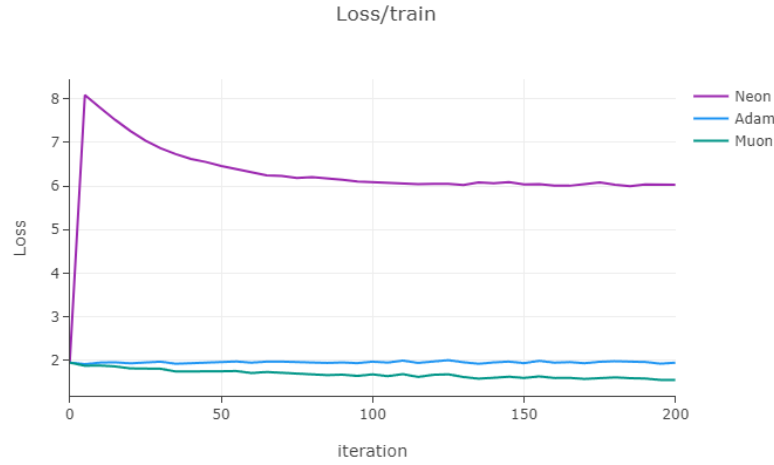


Figure 3: Train loss

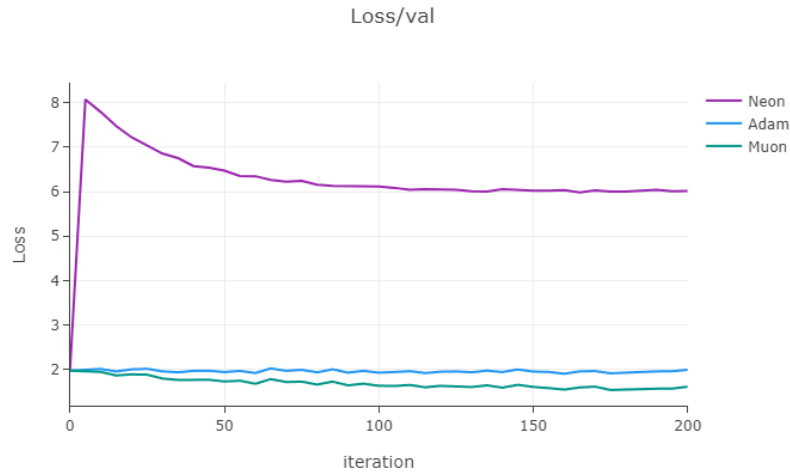


Figure 4: Validation loss

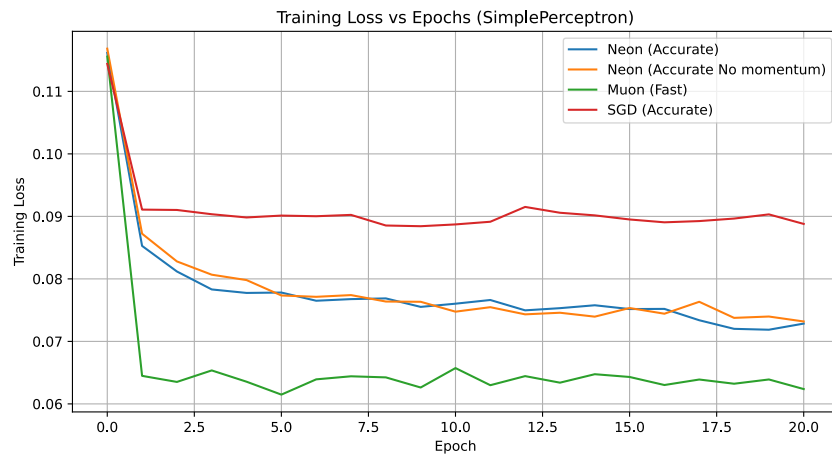


Figure 5: MLP: self.linear1 = nn.Linear(32\*32\*3, 512), self.linear2 = nn.Linear(512, 10), self.activ = nn.GELU()

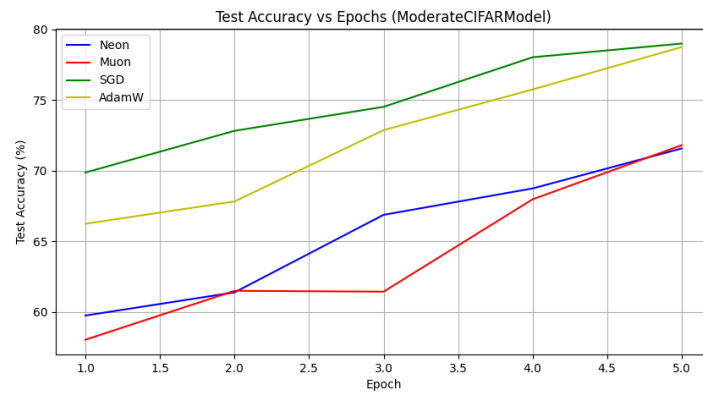


Figure 6: CNN: 2 convolutional blocks, 2 fully connected layers, activation + dropout

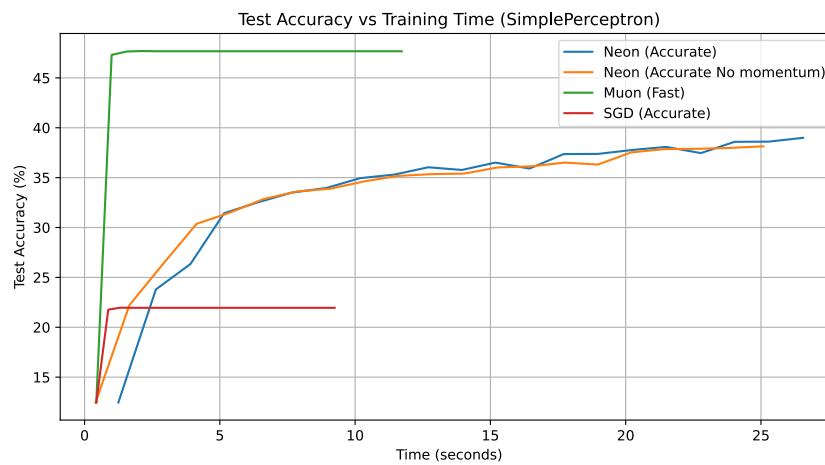


Figure 7: MLP: wallclock time measurements

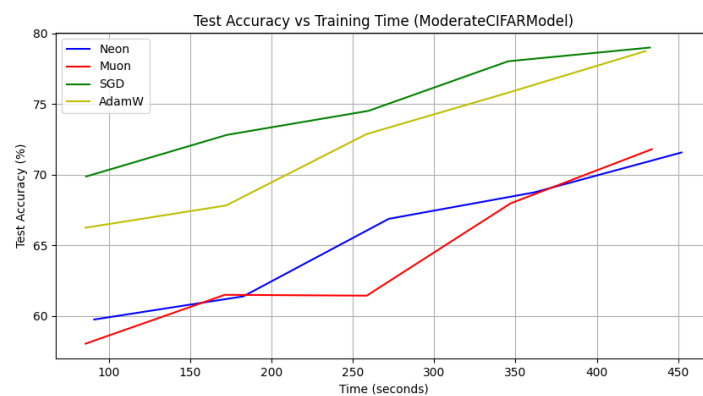


Figure 8: CNN: wallclock time measurements

## A APPENDIX

You may include other additional sections here.

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 1: k-rank updated comparison

Comparison of different numerical methods to calculate k-rank update on  $5000 \times 5000$  matrix of real numbers, rtol is an error in Frobenius norm relative to the k-rank approximation of truncated svd. 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 Lanczos method gives good and fast approximation for a matrix, but not so good approximation for singular values.