Neon: Nuclear Norm to Beat Muop

Alexey Kravatskiy
kravtskii.aiu@phystech.edu

Ivan Kozyrev
kozyrev.in@phystech.edu

Nikolay Kozlov
kozlov.na@phystech.edu

Alexander Vinogradov
vinogradov.am@phystech.edu

April 22, 2025

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 method we name Neon. After providing theoretical guarantees of Neon convergence, we compare performances of Neon, Muon, and Adam on training multilayer perceptron and BERT vectorizer.

1 Idea

The goal of the project is to make variations on Muon to speed it up. Recently, authors of [1] have proposed to look at different optimizers as the solution of the optimization problem for an update. This approach can be utilized to derive Muon [2], a novel algorithm for fast training of neural networks. Instead of using spectral norm in the problem, we use nuclear norm to produce a new problem. Then, we add momentum to update. After finalizing the algorithm, we test it on MLP, and, if results are satisfactory, on transformers (probably something from huggingface). The result will be a fast algorithm, which we will convert into a new optimizer class for PyTorch, as was done with Muon.

References

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

[2] 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.