
ETH Zürich - Deep Learning class 2025

Anna Mihalkova ^{* 1} Paul Gerry ^{* 1} Ilias Mc Auliffe ^{* 1}

Abstract

Muon orthogonalizes momentum at every step, which is computationally expensive and might be unnecessary. An approach would be to orthogonalize less frequently with an adaptive schedule. This is because early in training gradients tend to be noisier and require frequent orthogonalization, but later they are typically smaller and more stable. An adaptive approach of using heuristics for skipping orthogonalisation occasionally could reduce Muon’s computational cost while preserving its benefits. We would do a comparative analysis between our approach, classic Muon, and other optimizers based on sharpness and rank of the updates, performance, time, and disentanglement of the representations.

1. Background

Unlike traditional convex optimization, where the objective function has a single global minimum, deep learning optimization is non-convex. Several optimizers have been proposed to tackle the non-convex loss landscape, such as RMSprop, AdaGrad, Adam (Kingma & Ba, 2017). Finding a more efficient optimizer, in terms of compute, memory, and accuracy, is still an open problem.

A novel optimizer proposed recently is Muon (Jordan et al., 2024b; Bernstein, 2025). Instead of treating all model parameters as vectors and updating them per-coordinate, Muon uses orthogonalization to project the momentum into a direction with nice spectral properties and then scales it according to the layer dimensions. The idea is to get updates that are more isotropic, meaning uniform in all singular directions. Specifically, while optimizing a loss function $\mathcal{L}(\mathbf{W})$ that depends on a weight matrix $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, at iteration t

^{*}Equal contribution ¹Department of Computer Science, ETH Zürich. Correspondence to: Anna Mihalkova <amihalkova@student.ethz.ch>, Paul Gerry <pgerry@student.ethz.ch>, Ilias Mc Auliffe <imcauliffe@student.ethz.ch>.

we follow an update:

$$\mathbf{M}_t = \mu \mathbf{M}_{t-1} + \nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}_t), \quad (1)$$

$$\mathbf{P}_t = \text{orthogonalize}(\mathbf{M}_t), \quad (2)$$

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta_t \sqrt{d_{\text{out}}/d_{\text{in}}} \mathbf{P}_t, \quad (3)$$

The Muon optimizer has already been proven useful in practice (Liu et al., 2025), breaking the speed record by reaching 94% accuracy in 2.59 s (Jordan et al., 2024b) in the CIFAR-10 dataset (Krizhevsky et al., 2009).

2. Project Idea

Building on Muon’s success, we seek to preserve its performance while lowering the cost of orthogonalization. We intend to explore the following three Research Questions.

RQ1. How much does Muon performance degrade when the orthogonalization steps are skipped? We will use the available implementation (Jordan et al., 2024a) of the CNN trained with Muon that broke the record on the CIFAR-10 dataset (Krizhevsky et al., 2009). We will do an ablation study, where we will skip the orthogonalization step in certain parts of training.

RQ2. What alternatives to orthogonalization in Muon can reduce computational cost while preserving performance? The original implementation of Muon uses Newton-Schulz to orthogonalize the momentum matrix in equation (2), which is computationally expensive at $6nm^2$ FLOPS (Jordan et al., 2024b) for a $\mathbb{R}^{n \times m}$ weight matrix. Other works (Boreiko et al., 2025; Amsel et al., 2025) have adapted the orthogonalization in Muon via parallelisation (speed enhancement), learning rate and orthogonalisation hyperparameters (performance enhancement). We intend to check whether these techniques can also improve our approach.

RQ3. How are the Hessian sharpness, rank and disentanglement affected by adaptive orthogonalization? Sharpness (Cohen et al., 2022), (Cohen et al., 2025) and rank (MacKay, 1991) are indicators that measure generalisation properties of the optimisation scheme and efficiency in the utilization of the available parameters. Representation disentanglement can also be tracked via the flatness of the Hessian and mutual information (Achille & Soatto, 2018), (Carboneau et al., 2022). These metrics can give us insight into the quality of our approach, compared to other ones.

References

- Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations, 2018. URL <https://arxiv.org/abs/1706.01350>.
- Amsel, N., Persson, D., Musco, C., and Gower, R. M. The polar express: Optimal matrix sign methods and their application to the muon algorithm, 2025. URL <https://arxiv.org/abs/2505.16932>.
- Bernstein, J. Deriving muon, 2025. URL <https://jeremybernste.in/writing/deriving-muon>.
- Boreiko, V., Bu, Z., and Zha, S. Towards understanding of orthogonalization in muon. In *High-dimensional Learning Dynamics 2025*, 2025. URL <https://openreview.net/forum?id=ppmyFtr9EW>.
- Carboneau, M.-A., Zaidi, J., Boillard, J., and Gagnon, G. Measuring disentanglement: A review of metrics, 2022. URL <https://arxiv.org/abs/2012.09276>.
- Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient descent on neural networks typically occurs at the edge of stability, 2022. URL <https://arxiv.org/abs/2103.00065>.
- Cohen, J. M., Damian, A., Talwalkar, A., Kolter, J. Z., and Lee, J. D. Understanding optimization in deep learning with central flows, 2025. URL <https://arxiv.org/abs/2410.24206>.
- Jordan, K., Bernstein, J., Rappazzo, B., @fern-bear.bsky.social, Vlado, B., Jiacheng, Y., Cesista, F., Koszarsky, B., and @Grad62304977. modded-nanogpt: Speedrunning the nanogpt baseline, 2024a. URL <https://github.com/KellerJordan/modded-nanogpt>.
- Jordan, K., Jin, Y., Boza, V., You, J., Cesista, F., Newhouse, L., and Bernstein, J. Muon: An optimizer for hidden layers in neural networks, 2024b. URL <https://kellerjordan.github.io/posts/muon/>.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017. URL <https://arxiv.org/abs/1412.6980>.
- Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.
- Liu, J., Su, J., Yao, X., Jiang, Z., Lai, G., Du, Y., Qin, Y., Xu, W., Lu, E., Yan, J., Chen, Y., Zheng, H., Liu, Y., Liu, S., Yin, B., He, W., Zhu, H., Wang, Y., Wang, J., Dong, M., Zhang, Z., Kang, Y., Zhang, H., Xu, X., Zhang, Y., Wu, Y., Zhou, X., and Yang, Z. Muon is scalable for llm training, 2025. URL <https://arxiv.org/abs/2502.16982>.
- MacKay, D. Bayesian model comparison and back-prop nets. In Moody, J., Hanson, S., and Lippmann, R. (eds.), *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991. URL https://proceedings.neurips.cc/paper_files/paper/1991/file/c3c59e5f8b3e9753913f4d435b53c308-Paper.pdf.