Optimization Benchmark for Diffusion Models on Dynamical Systems

Fabian Schaipp[1]

[1]Inria, ENS, PSL Research University, Paris

EurIPS

TLDR

New optimizers like `Muon` and `SOAP` can outperform `AdamW` for training diffusion models.

Research Questions

Question 1: What are good optimizers for training diffusion models?

Question 2: Is diffusion training (for scientific applications) fundamentally different from LLM pretraining from an optimization perspective?

Background

- Recent large-scale optimization benchmarks **do not include diffusion** training [7, 4].
- Most recent optimization findings are only evaluated on image and text applications.
- Several new optimizers (e.g. `Shampoo`[2], `Muon`[3], `SOAP`[8]) have emerged for LLM training that outperform `AdamW`.

`Shampoo`, `Muon` and `SOAP`

Let $W_t \in \mathbb{R}^{m\times n}$ be the current weight **matrix**, and $G_t \in \mathbb{R}^{m\times n}$ its gradient.

$$\begin{aligned} L_{t+1} &= \beta_2 L_t + (1-\beta_2) G_t G_t^T, \\ R_{t+1} &= \beta_2 R_t + (1-\beta_2) G_t^T G_t, \\ W_{t+1} &= W_t - \eta_t L^{-1/4} G_t R^{-1/4}. \end{aligned} \tag{Shampoo}$$

We can rewrite `Shampoo` with $\beta_2 = 0$ as

$$W_{t+1} = W_t - \eta_t \texttt{Ortho}(G_t). \tag{Muon}$$

Here, for matrix $A \in \mathbb{R}^{m\times n}$ with SVD $A = U\Sigma V^T$, we define $\texttt{Ortho}(A) := UV^T$. In practice, `Muon` also uses momentum and weight decay, and uses `Adam` on 1D parameters.

`SOAP` is running `Adam` in the (approx.) eigenbasis of (L_t, R_t) of `Shampoo`:

$$\begin{aligned} Q_L,\ Q_R &= \texttt{QR}(L_t),\ \texttt{QR}(R_t) \\ \hat{G}_t &= Q_L^T G_t Q_R \to \texttt{Adam with } \hat{G}_t. \end{aligned} \tag{SOAP}$$

Takeaway

`Shampoo` and its follow-up methods `Muon` and `SOAP` use dense preconditioning (unlike `Adam`'s diagonal preconditioning); therefore, their computation **time per step is bigger**.

Benchmark problem

- We train a diffusion model that denoises trajectories of dynamical systems.
- Such models can be used for score-based data assimilation (see [6]).
- The training data are snapshots of the 2D velocity field governed by Navier-Stokes with Kolmogorov flow. The model is a standard `U-Net` model with 23M parameters.

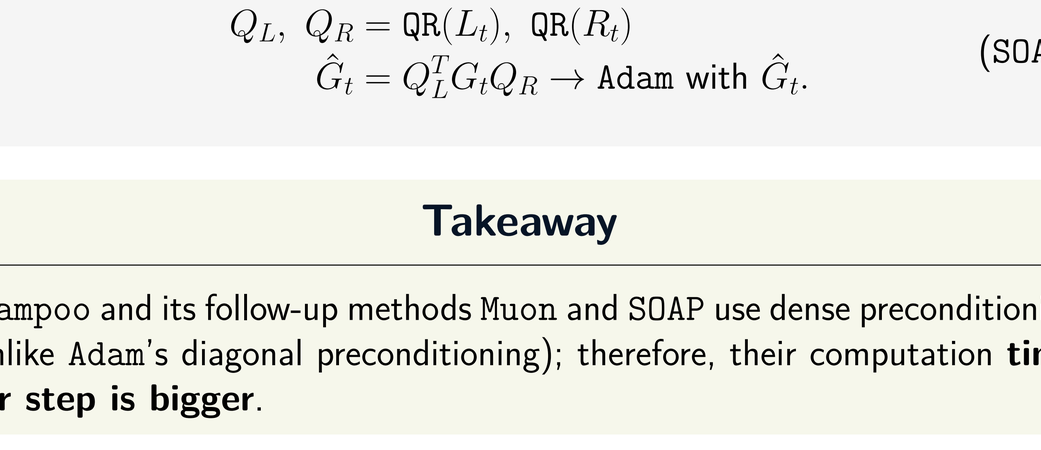

Figure 1. Generated example trajectories (5 snapshots).

Main result

- Learning-rate and weight decay tuned for each method individually
- Linear-decay schedule (except for `ScheduleFree` [1]), over 1024 epochs

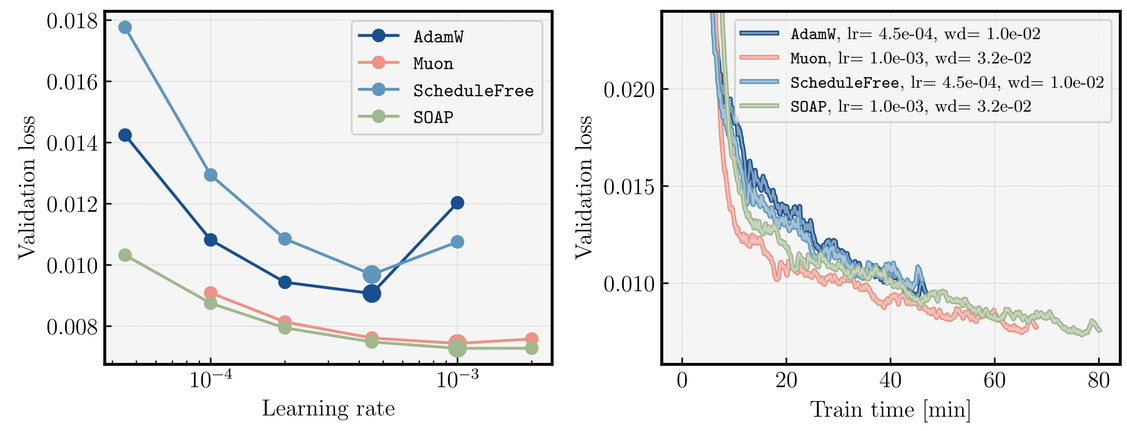

Figure 2. **(Left)** Final loss across learning rate. **(Right)** Best run per method.

Takeaways

- `Muon` and `SOAP` are very efficient for diffusion model training (18% lower final loss as `AdamW`, with $1.45\times$ and $1.7\times$ runtime).
- `AdamW` can not close this gap by simply training longer (Fig. 3 left).
- `Prodigy` [5] avoids learning-rate tuning (e.g. for preliminary experiments).
- We observe a mismatch between final loss and generative quality for `ScheduleFree` and (less pronounced) the `wsd` schedule.

.

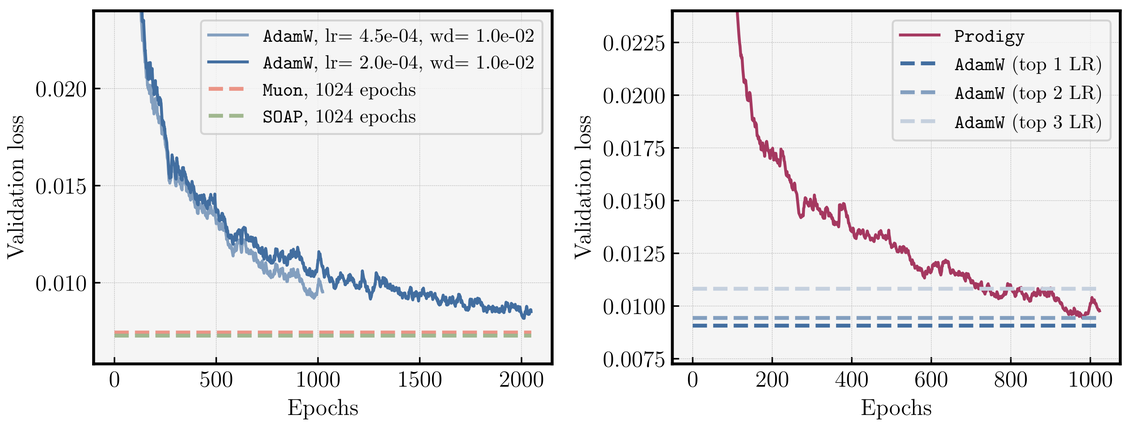

Figure 3. **(Left)** Longer training doesn't close the gap for `AdamW`. **(Right)** `Prodigy` avoids learning-rate tuning.

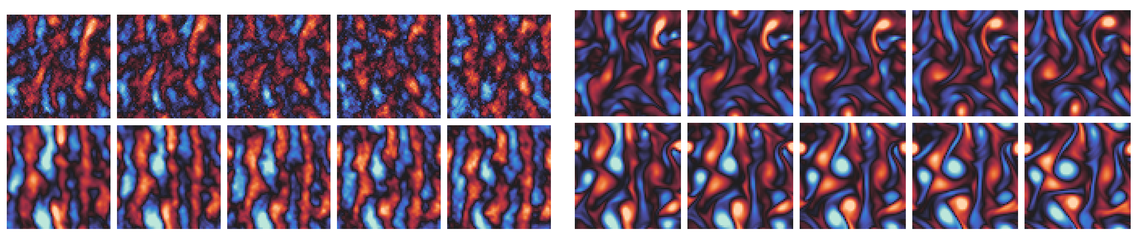

Figure 4. **(Left)** `ScheduleFree` (train loss 0.0099) and **(right)** `AdamW` (train loss 0.0102).

Links

Paper: `arxiv.org/abs/2510.19376`
Code: `github.com/fabian-sp/sda`

References

[1] Aaron Defazio, Xingyu Yang, Ahmed Khaled, Konstantin Mishchenko, Harsh Mehta, and Ashok Cutkosky. The road less scheduled. In *NeurIPS*, 2024.

[2] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *ICML*, 2018.

[3] Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024.

[4] Priya Kasimbeg, Frank Schneider, Runa Eschenhagen, Juhan Bae, Chandramouli Shama Sastry, Mark Saroufim, Boyuan Feng, Less Wright, Edward Z. Yang, Zachary Nado, Sourabh Medapati, Philipp Hennig, Michael Rabbat, and George E. Dahl. Accelerating neural network training: An analysis of the algoperf competition. In *ICLR*, 2025.

[5] Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. In *ICML*, 2024.

[6] François Rozet and Gilles Louppe. Score-based data assimilation. In *NeurIPS*, 2023.

[7] Andrei Semenov, Matteo Pagliardini, and Martin Jaggi. Benchmarking optimizers for large language model pretraining. 2025.

[8] Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham M. Kakade. SOAP: improving and stabilizing Shampoo using Adam for language modeling. In *ICLR*, 2025.