Where Does Warm-Up Come From?

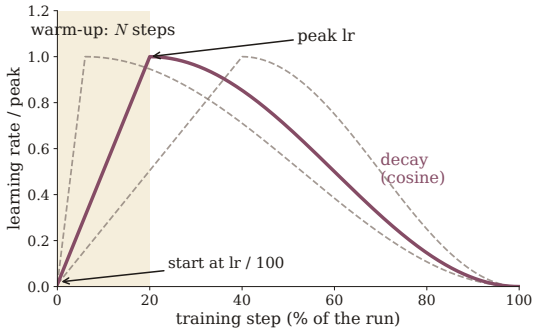
Adaptive Scheduling for Norm-Constrained Optimizers

Artem Riabinin Andrey Veprikov Arman Bolatov Martin Takáč Aleksandr Beznosikov

NeurIPS 2026

BRAIn Lab

What is warm-up?



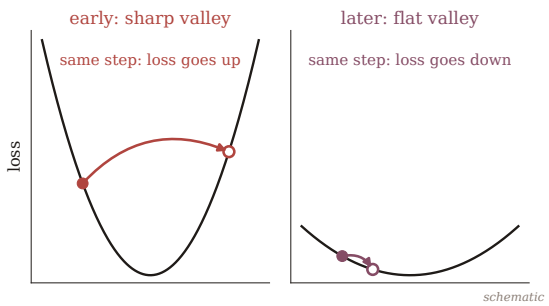
Linear warm-up, then cosine decay. Dashed: the same schedule with a shorter or a longer warm-up.

The recipe. Start with a small learning rate, raise it to the peak over the first N steps, then let it decay.

Everywhere. Standard since ResNets and Transformers, and almost every LLM run uses it.

One open knob: the length N. Which value is right?

Why start small?

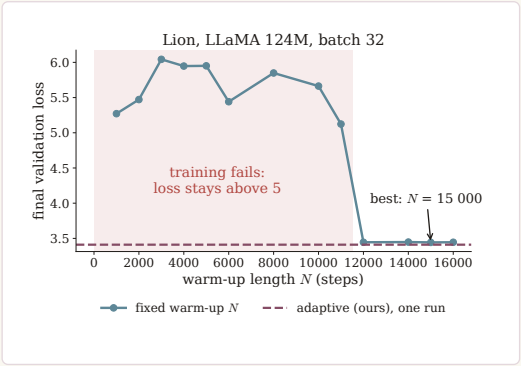


Schematic: the same learning rate on a sharp and on a flat valley.

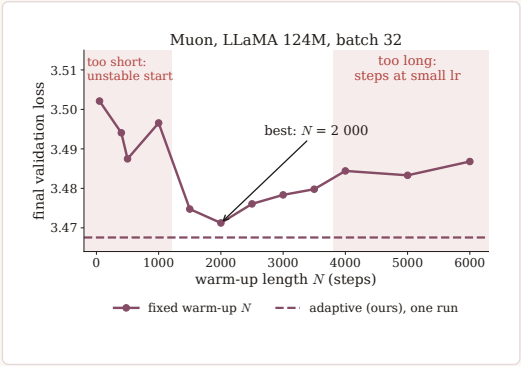
Early on the surface is sharp. The gradient changes fast, so a step that is fine later overshoots now and the loss jumps up.

Warm-up waits it out. Steps stay small until the surface flattens. Sharpness is known to drop during warm-up.

A wrong N costs a whole run



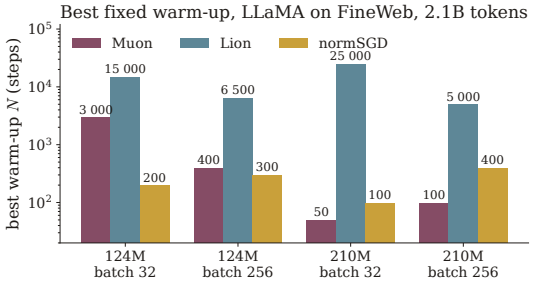
Lion, LLaMA 124M: below 12 000 warm-up steps training never recovers.



Muon, LLaMA 124M: too short and too long both hurt.

Every point is a full pretraining run that differs only in N. Dashed: our method, a single run.

And there is no rule for N



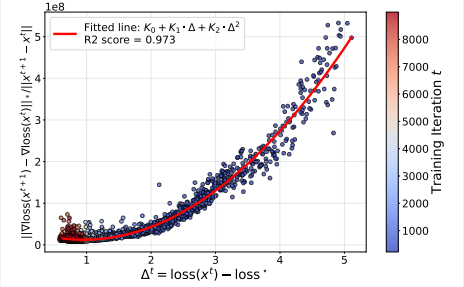
Best fixed warm-up per setting, LLaMA on FineWeb, 2.1B tokens.

50 to 25 000 steps. The best N moves with the optimizer, model size and batch size. A value tuned on a small proxy does not transfer.

Cost. These sweeps took 225 runs, about 1.2k H200 GPU-hours.

Our questions. Can warm-up be derived instead of guessed? Can training set N itself?

Clue: curvature falls as the loss falls



Lion on LLaMA: measured curvature $\mathcal{K}^t$ against the loss gap Δ^t.

Optimizers. Muon, Lion and normSGD step along a unit direction, $x^{t+1} = x^t + \eta^t \,\mathrm{LMO}(g^t)$, so the learning rate is the step length.

What we measured. How fast the gradient changes per unit step, against the gap $\Delta = f(x) - f^\star$ to the target loss. It is a parabola: large while the loss is high, small later.

Not caused by warm-up: the same shape with no warm-up at all.

Safe step = gap / curvature: warm-up appears by itself

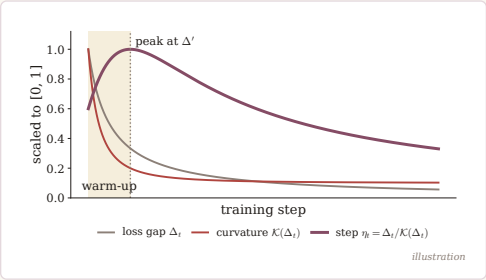


Illustration: the gap and the curvature fall, their ratio rises, then falls.

Assumption. Curvature grows with the gap,

$$\mathcal{K}(\Delta) = K_0 + K_1\Delta + K_\rho\Delta^\rho.$$

Theorem (informal). The step $\eta^t = \Delta^t/(D\,\mathcal{K}(\Delta^t))$ decreases the loss at every step and converges as $\mathcal{O}(1/T)$.

Rises, then falls. The gap only shrinks, so the step first grows (warm-up) and then decays. Nobody put warm-up in: the proof selects it.

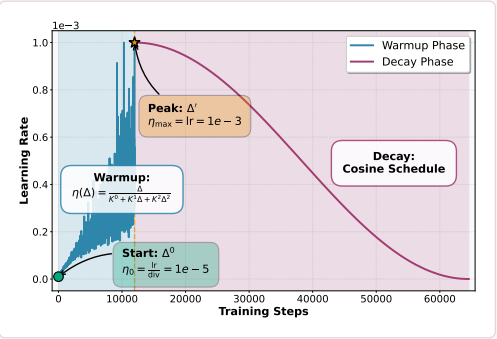
A scheduler that finds N by itself

The loss is known at every step, so is the gap Δ^t. The learning rate is

$$\eta(\Delta) = \frac{\Delta}{K_0 + K_1\Delta + K_2\Delta^2}.$$

Inputs you already set: peak lr, start lr, a target loss $f^\star$. They fix the constants.

No N anywhere. Warm-up ends when the loss is low enough, then the usual decay takes over. Any $f^\star \in [3.0, 4.0]$ beats the tuned warm-up.



The schedule our method produced for Lion. The loss chose the warm-up length.

Plugging it in: one line

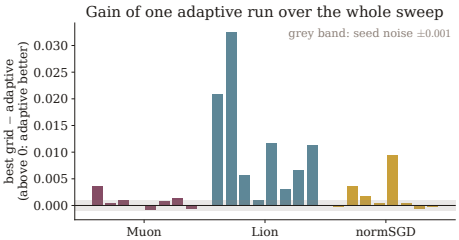
```
sched = LipschitzScheduler(opt, max_lr=lr, min_lr=lr/100,
                           loss_star=3.3, max_steps=T)
for batch in data:
    loss = model(batch); loss.backward()
    opt.step(); opt.zero_grad()
-   sched.step()
+   sched.step(loss.item())      # gap = loss - f*
```

Simplified from `src/optim/lipschitz_scheduler.py` in our repository.

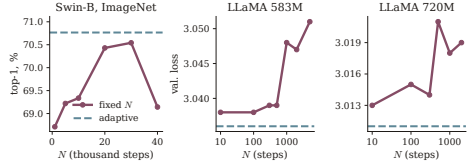
Inside. Once: fit K_0, K_1, K_2, Δ' to lr, div and the target shape. Then $\eta(\Delta)$ while $\Delta \geq \Delta'$, cosine decay after.

What you drop: `warmup_steps`. What you add: a rough target loss $f^\star$.

One run instead of a sweep



All 24 settings: 124M and 210M, batch 32 and 256, 1B and 2.1B tokens.



Swin-B on ImageNet, LLaMA 583M and 720M. Dashed: one adaptive run.

> **Never worse than the best grid point by more than seed noise,** clearly better for Lion, and the same at larger scale and on vision.

Takeaways

- Warm-up is needed because curvature is high early, and it falls as the loss falls.
- Step = gap / curvature gives warm-up and decay straight from the proof.
- Our scheduler lets the loss pick the warm-up length: one run instead of a sweep.

Paper

Code

Thank you