The Surprising Agreement Between Convex Optimization Theory and Learning-Rate Scheduling for Large Model Training

Fabian Schaipp [1] Alexander Hägele [2] Adrien Taylor [1] Umut Şimşekli [1] Francis Bach [1]

[1]Inria, ENS, PSL Research University, Paris [2]EPFL, Lausanne

TLDR

The behaviour of learning-rate schedules in LLM training can be described with *convex* optimization theory (surprisingly!).

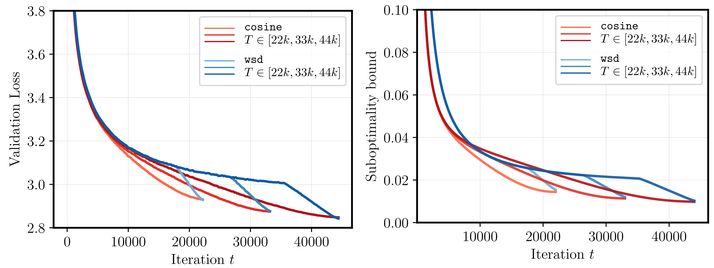


Figure 1. Real loss curves **(left)** and theoretical bound **(right)**

History/Background

- For optimal performance, the `cosine` schedule needs to match total number of iterations [3].
- More practical alternative: constant + cooldown schedule (called `wsd`) [2]. Observe **sudden drop of the loss** during cooldown.

Research Questions

Question 1: What does optimization theory say about learning-rate schedules?

Question 2: What happens during cooldown?

Setup

Solve $\min_{x\in\mathbb{R}^d} f(x)$ with

$$x_{t+1} = x_t - \gamma\eta_t g_t, \qquad \texttt{(SGD)}$$

where $\gamma > 0$ is **base learning-rate** and $(\eta_t)_{t\in\mathbb{N}}$ is **schedule**.

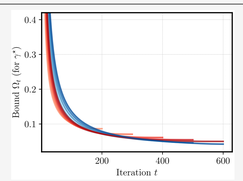
Takeaway 1

Standard bounds of the form

$$\min_{t=1,\dots,T} \mathbb{E}[f(x_t) - f(x_\star)] \leq \Omega_T$$

do not reflect the real-world performance.

⇝ **Need to look at last-iterate bound!**



Last-iterate bound

Bound by (Defazio et al., 2023): Let $D := \|x_1 - x^\star\|$. If f is convex and G-Lipschitz, then

$$\mathbb{E}[f(x_T) - f(x_\star)] \leq \frac{1}{2\gamma\sum_{t=1}^T \eta_t}\Big[D^2 + \gamma^2 G^2 \sum_{t=1}^T \eta_t^2\Big] + \frac{\gamma G^2}{2}\sum_{k=1}^{T-1}\Big(\frac{\eta_k}{(\sum_{t=k+1}^T \eta_t)(\sum_{t=k}^T \eta_t)}\sum_{t=k}^T \eta_t^2\Big) =: \Omega_T.$$

Compute optimal base learning-rate $\gamma^\star$ by hand ($\sim$ tuning). Plugging in `cosine` and `wsd` produces the figure from the beginning.

Takeaway 2

- Bound from convex, nonsmooth optimization **reproduces the empirical behaviour** of LR schedules. (we don't know why)
- Bound predicts that $\gamma^\star(\texttt{cosine}) \approx 2 \cdot \gamma^\star(\texttt{wsd})$. This matches in practice.
- Cooldown in `wsd` achieves an improvement of $\log(T)$.
- Bound predicts the optimal cosine cycle length reported in [3].

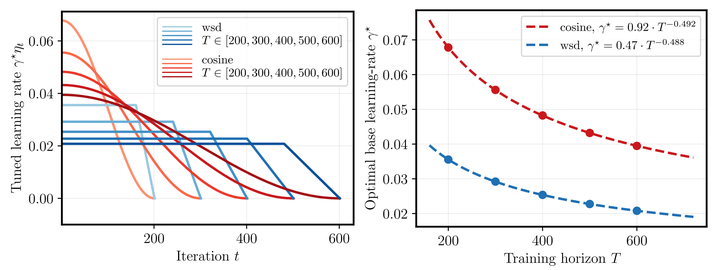


Figure 2. **(Left)** `cosine` and `wsd` schedule. **(Right)** Optimal base learning-rate $\gamma^\star$.

Applications

(1) Continual training: Extend training length from T_1 (short) to T_2 (long) steps without starting from scratch by reusing constant LR checkpoint. Problem: optimal base LR decreases with training length. We can (i) decrease schedule during $[T_1, T_2]$ or (ii) extend cooldown length for the long run.

⇝ For both approaches, use the bound as testbed for schedule design!

⇝ **Improvements equivalent to ~ 7% of additional steps** (see experiments).

(2) Learning-rate transfer: If we know optimal base LR $\gamma^\star$ for some schedule, can we infer the optimal LR for a different schedule? For example, transfer from cooldown length 20% to 100% (linear-decay).

Experiments

We train `Llama`-style transformer models (124M and 210M) on `SlimPajama-6B` dataset.

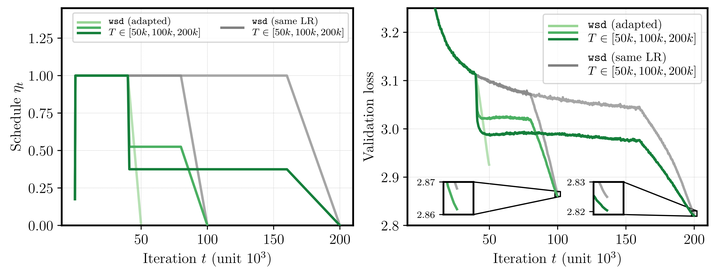


Figure 3. Continual training, approach (i). **(Left)** Schedule design, **(right)** loss

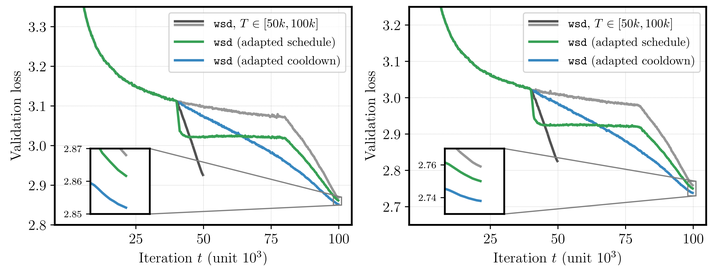


Figure 4. Continual training, approach (ii). **(Left)** 124M model, **(right)** 210M model

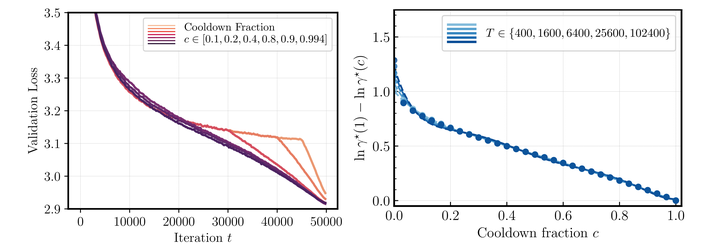


Figure 5. **(Left)** Different cooldown lengths, **(right)** learning-rate transfer

References

[1] Aaron Defazio, Ashok Cutkosky, Harsh Mehta, and Konstantin Mishchenko. Optimal linear decay learning rate schedules and further refinements. *arXiv:2310.07831*, 2023.

[2] Alex Hägele, Elie Bakouch, Atli Kosson, Loubna Ben allal, Leandro Von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations. In *NeurIPS*, 2024.

[3] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *NeurIPS*, 2022.

Link to paper