Why Do We Need Warm-up?
A Theoretical Perspective

Foivos Alimisis* Rustem Islamov* Aurelien Lucchi
University of Basel, Switzerland

University of Basel

ICML International Conference On Machine Learning

Main Claim

Learning rate warm-up is what you get when the learning rate adapts to an early-training curvature reduction.

Motivation: the warm-up puzzle

A common schedule first *increases* the learning rate and later decays it (especially in LLMs and vision transformers).

Question: Is there an intrinsic property of neural-network loss landscapes that justifies LR warm-up?

Core issue. A gradient-dependent condition such as (L_0, L_1)-smoothness can prescribe smaller steps when the gradient grows early in training, which is the opposite of practical warm-up behavior. Our work replaces gradient magnitude with loss suboptimality.

Contributions

- **New condition:** We introduce (H_0, H_1)-smoothness: curvature is bounded by an affine function of loss suboptimality.
- **Landscape evidence:** We show theoretically and empirically that this is representative of the early sharpness-reduction phase.
- **Warm-up theory:** We derive faster convergence guarantees for adaptive warm-up schedules than for fixed-step GD under structured non-convexity.
- **Practical schedule:** We test a one-parameter loss-driven warm-up matching tuned linear warm-up on language and vision tasks.

The (H_0, H_1)-smoothness condition

Definition. A function $f : \mathbb{R}^d \to \mathbb{R}$ with minimum $f^\star > -\infty$ is (H_0, H_1)-smooth when

$$\|\nabla^2 f(w)\|_2 \le H_0 + H_1(f(w) - f^\star) \qquad \forall w \in \mathbb{R}^d.$$

Loss decreases ⇒ curvature bound decreases ⇒ stable LR should increase.

- Closed under finite sums and affine maps.
- Contains classical (L_0, L_1)-smoothness under bounded-below assumptions.

Theoretical justification of (H_0, H_1)

For the simple $1 \times 1 \times 1$ linear network

$$f(u, v) = (y - uvx)^2,$$

the Hessian spectral norm has two phases along gradient flow from near-zero initialization:

$$\frac{d}{dt}\|H(t)\|_2 < 0 \text{ when } \frac{x^2(u(0) - v(0))^2}{4} \le (y - 2u(t)v(t)x)^2,$$
$$\frac{d}{dt}\|H(t)\|_2 > 0 \text{ when } y < 2u(t)v(t)x.$$

Thus, **initial phase is sharpness reduction.**

We also tackle more realistic deep model architectures. A summary can be seen below:

Architecture, loss	Result in the paper
Deep linear net, MSE	Under balancedness, early-training and global (H_0, H_1) bounds.
Nonlinear feedforward net	Extension to leaky-ReLU networks under weak balancedness and mild spectral assumptions.
Two-layer net, CE	L2 regularization gives local and global (H_0, H_1) bounds.
One-layer transformer	In-context loss satisfies the condition under distributional regularity and L2 regularization.

Empirical justification of (H_0, H_1)

The initial phase of training is usually a phase of sharpness reduction. It precedes the well-studied progressive sharpenning and edge-of-stability phases. This experiment uses a smoothness proxy.

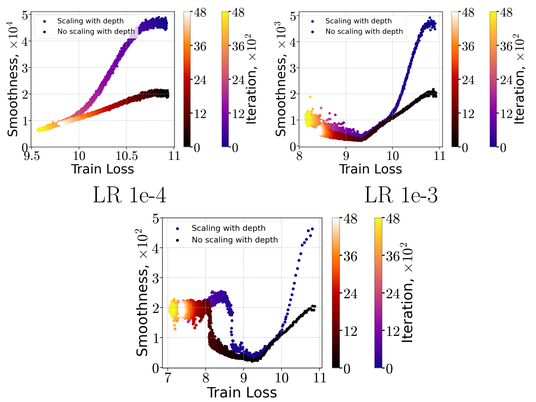


Figure 1:Training of 70M model on FineWeb dataset with SGD varying fixed learning rate and initialization scheme. Colored points correspond to depth-scaled initialization, while black points correspond to fixed-variance initialization. Color indicates training progress.

Adaptive warm-up derived from the (H_0, H_1) condition

The theoretical learning rate derived by (H_0, H_1)-smoothness

$$\eta_k := \frac{1}{10H_0 + 20H_1(f(w_k) - f^\star)}.$$

Since $f(w_k) - f^\star =: \Delta_k$ decreases during the early phase, η_k increases automatically.

Fixed step-size
must be stable for worst early curvature
$\eta \lesssim \frac{1}{H_0 + H_1\Delta_0}$

→

Adaptive warm-up
tracks decreasing loss and curvature bound
$\eta_k = \frac{1}{10H_0 + 20H_1\Delta_k}$

Iteration complexity analysis

Upper bounds for realistic function classes and a warm-up LR η_k

Aiming condition: If f is (H_0, H_1)-smooth and satisfies the Aiming condition with constant θ around the minimizer set $\mathcal{S}$, then adaptive GD reaches $f(w_K) - f^\star \le \varepsilon$ after at most

$$\frac{40H_0\,\mathrm{dist}(w_0, \mathcal{S})^2}{\theta^2\varepsilon} + \frac{40H_1\,\mathrm{dist}(w_0, \mathcal{S})^2}{\theta^2} \quad \text{iterations.}$$

PL condition: If f also satisfies μ-PL, adaptive GD reaches ε after at most

$$\frac{40H_1}{\mu}(f(w_0) - f^\star) + \frac{20H_0}{\mu}\log\frac{H_0}{2H_1\varepsilon}.$$

Lower bounds for the same function classes and a fixed LR η are larger for sufficiently large initial loss suboptimality.

Why fixed steps are slower

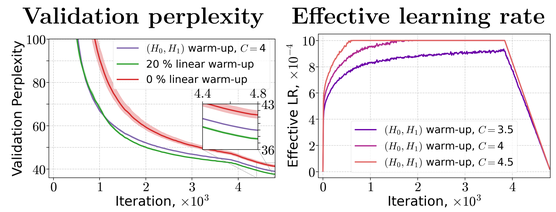
Fixed-step GD must choose η small enough for the largest early curvature. The lower bounds contain the initialization factor

$$\frac{H_1(f(w_0) - f^\star)}{\log(f(w_0) - f^\star) + 1},$$

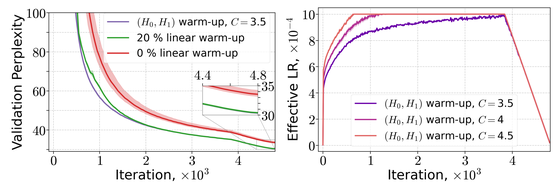
so a high initial loss forces conservative steps even after curvature has dropped.

Warm-up schedule experiments

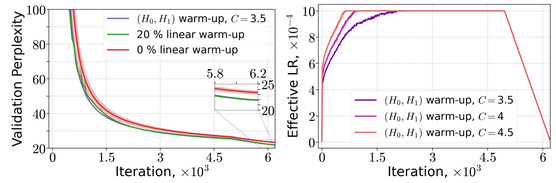
We test the practical warm-up schedule produced by the (H_0, H_1) condition against a classic 20% linear warm-up strategy, and no warm-up. The (H_0, H_1)-derived warm-up performs on par with the linear warm-up schedule, which reveals that our theory is likely realistic.



70M on FineWeb



160M on FineWeb



410M on FineWeb

Left column: validation perplexity for 70M, 160M, and 410M language models on FineWeb under no warm-up, tuned linear warm-up, and (H_0, H_1) warm-up.
Right column: effective LR induced by (H_0, H_1) warm-up for peak LR 10^{-3} and different values of C. All runs use a final 20% linear decay to 10^{-5}.

Practical surrogate used in experiments

The exact constants $H_0, H_1, f^\star$ are unknown in training. The implementation therefore uses the one-parameter schedule

$$\eta_k^{\text{eff}} = \frac{\eta_k^{\text{base}}}{\max\{1, f_{S_k}(w_k)/C\}},$$

where C controls when the loss-dependent denominator becomes active and thus sets the effective warm-up length.

* First two authors contributed equally