A Simple and Efficient Measure of Loss Landscape Curvature

Hee-Sung Kim Sungyoon Lee
Department of Computer Science, Hanyang University
HiLD 2026: 4th Workshop on High-dimensional Learning Dynamics

Motivation & Contributions

The **Edge of Stability** tracks training through a **sharpness measure** reaching an optimizer-dependent threshold — yet measuring the Hessian eigenvalue $\lambda_{\max}^{H}$ is **prohibitive at scale** and **incompatible** with modern training kernels (e.g. FlashAttention). We revisit **directional curvature** λ_{dir} along the optimizer's update direction as a scalable alternative.

(i) Optimizer-independent $2/\eta$ boundary. Absorbing momentum and preconditioning into the update direction $\Delta\theta$ — rather than the threshold — keeps the one-step descent boundary at $2/\eta$ for GD, momentum, *and* Adam.

(ii) Forward-only estimators. A finite-difference (λ_{FD}) and a KL-divergence (λ_{KL}) estimator recover λ_{dir} with *no* Hessian-vector products — just 1–2 extra forward passes per step.

The Problem: Hessian Sharpness Is Expensive

- Under full-batch GD, $\lambda_{\max}^{H}$ rises to $2/\eta$, then hovers there — the **Edge of Stability**.
- Power iteration / Lanczos need many **Hessian-vector products**: costly at LLM scale.
- Double backpropagation is **incompatible** with modern kernels.

⇒ Measure curvature only *along the direction $\Delta\theta$ the optimizer actually takes.*

Directional Curvature from One-Step Descent

For an update $\theta_{\text{new}} = \theta - \eta\,\Delta\theta$ with effective direction $\Delta\theta$ (momentum / preconditioning included):

$$\Delta L = -\eta\, g^{\top}\Delta\theta + \frac{1}{2}\eta^{2}\,\Delta\theta^{\top} H \Delta\theta + O(\eta^{3}).$$

The boundary $\Delta L = 0$ marks the smallest η at which one-step descent ceases:

Directional Curvature

$$\lambda_{\text{dir}} = \frac{\Delta\theta^{\top} H\,\Delta\theta}{g^{\top}\Delta\theta} \approx \frac{2}{\eta}$$

Numerator = curvature cost along $\Delta\theta$;
denominator = first-order descent gain.

One Boundary for Every Optimizer

Momentum and preconditioning enter $\Delta\theta$**, not the threshold**:

- **Full-batch GD** ($\Delta\theta = g$) recovers the standard directional sharpness $\lambda_{\text{dir}}^{\text{SGD}} = \dfrac{g^{\top} H g}{g^{\top} g}$.
- **Momentum / Adam:** $\Delta\theta$ is no longer aligned with g, yet the same Taylor expansion keeps the boundary at $2/\eta$.

Stability threshold	SGD	Heavy-ball mom.
Hessian sharpness $\lambda_{\max}^{H}$	$2/\eta$	$(2+2\beta)/\eta$
Directional λ_{dir}	$\mathbf{2/\eta}$	$\mathbf{2/\eta}$

Key idea: shift the optimizer dependence *from the threshold to the measured quantity*. This one-step descent view is complementary to — not in conflict with — long-horizon iterate-stability thresholds.

A stochastic extension in expectation keeps the same $2/\eta$ boundary and subsumes the Interaction-Aware Sharpness of Lee & Jang (2023).

Conclusion & Outlook

- λ_{dir} is a **scalable replacement** for Hessian sharpness with a fixed $2/\eta$ boundary: the optimizer state enters through $\Delta\theta$, not the threshold.
- Two **forward-only** estimators; λ_{KL} is the preferred default for cross-entropy — cleaner signal at half the cost.
- **Next:** validate λ_{KL} at LLM scale, where HVP-based monitoring breaks down, and plug it into curvature-aware learning-rate tuners.

Forward-Pass Estimators (No HVPs)

We probe the second-order Taylor coefficient along $\Delta\theta$ with forward evaluations only.

Finite-Difference λ_{FD} — 2 forward passes

$$\lambda_{\text{FD}} \triangleq \frac{L(\theta+\epsilon\Delta\theta) - 2L(\theta) + L(\theta-\epsilon\Delta\theta)}{\epsilon^{2}\, g^{\top}\Delta\theta} = \lambda_{\text{dir}} + O(\epsilon^{2})$$

The symmetric probe cancels odd-order terms; ϵ trades Taylor error against floating-point cancellation.

KL-Divergence λ_{KL} — 1 forward pass

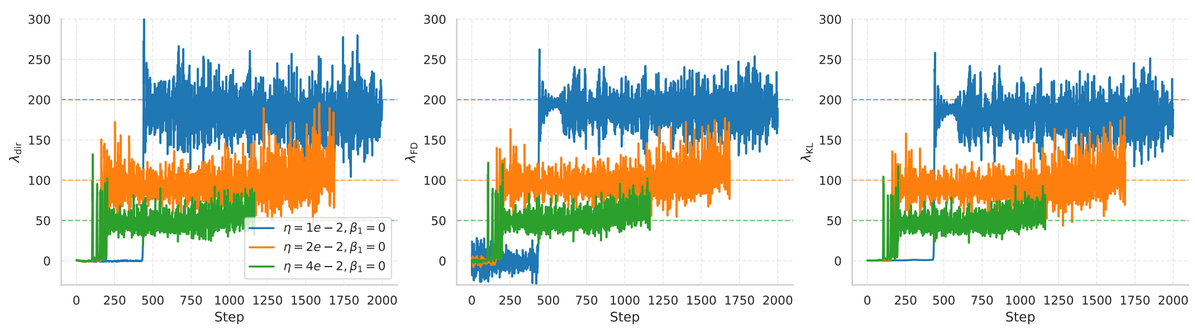
$$\lambda_{\text{KL}} \triangleq \frac{2\,\mathbb{E}_{x}\big[\,\mathrm{D}_{\text{KL}}\big(p_{\theta}\,\|\,p_{\theta+\epsilon\Delta\theta}\big)\big]}{\epsilon^{2}\, g^{\top}\Delta\theta} = \frac{\Delta\theta^{\top} F\,\Delta\theta}{g^{\top}\Delta\theta} + O(\epsilon)$$

Under cross-entropy, KL curvature is the **Fisher information** F — the Gauss-Newton part of H: a Fisher analogue of λ_{dir} at *half* the cost.

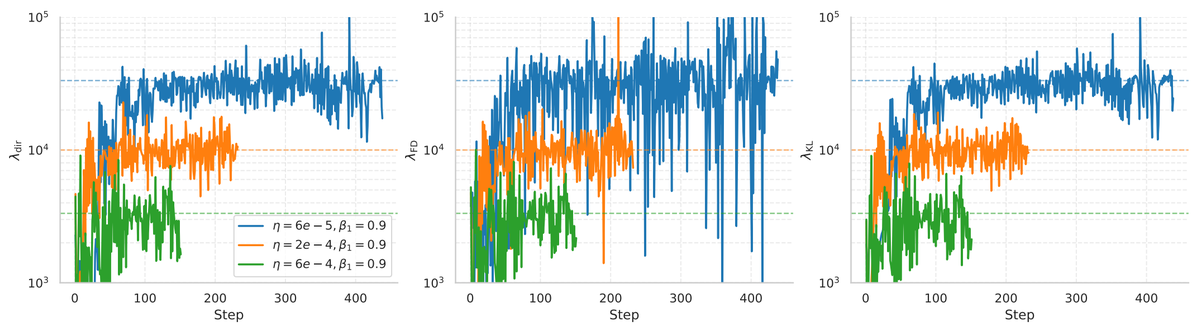
Why Fisher Tracks the Hessian

For softmax cross-entropy $H = F + R$, with the residual R weighted by errors $(f_i - y_i)$; near interpolation $R \to 0$, so $\lambda_{\text{KL}} \to \lambda_{\text{dir}}$. Moreover, $\Delta\theta$ aligns with the few outliers of the Fisher spectrum, so λ_{KL} captures the dominant curvature mode *throughout* training.

Results: Both Estimators Track $2/\eta$



(a) ResNet-9 (no BN), full-batch CIFAR-10, **GD** (linear scale). Dashed lines mark $2/\eta \in \{50, 100, 200\}$.



(b) Same model, **AdamW** (β_1=0.9, β_2=0.999, log scale) — same $2/\eta$ comparison, *no* optimizer-specific adjustment.

- All three **rise, then oscillate around** $2/\eta$ — the progressive-sharpening / EoS signature.
- λ_{KL} tracks the exact λ_{dir} (HVP reference) **cleanly**; λ_{FD} is noisier at twice the cost.