Zeroth-Order Optimization at the Edge of Stability

Minhak Song[1,2] Liang Zhang[3,4] Bingcong Li[3] Niao He[3] Michael Muehlebach[4] Sewoong Oh[5]

[1]KAIST [2]KRAFTON [3]ETH [4]MPI-IS [5]UW

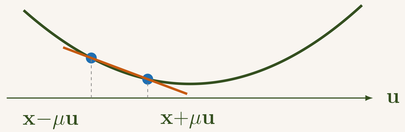
Zeroth-Order Optimization

Zeroth-order (ZO) methods optimize using only function evaluations, without computing gradients.

Used when gradients are unavailable or expensive (black-box learning; memory-efficient LLM fine-tuning). The gradient is estimated along a random direction $\mathbf{u}$ from two function evaluations:

$$\widehat{\nabla} f(\mathbf{x}) = \frac{f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x}-\mu\mathbf{u})}{2\mu}\mathbf{u}, \qquad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

ZO-GD: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\,\widehat{\nabla} f(\mathbf{x}_t)$; the momentum (ZO-GDM) and adaptive (ZO-Adam) variants apply the same estimator.



A finite-difference slope along $\mathbf{u}$: *unbiased* on a quadratic and *forward-pass only* (no backpropagation), so ZO is memory-efficient.

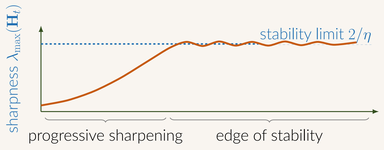
The Edge of Stability

In deep learning, gradient descent trains at the boundary between convergence and divergence under quadratic approximation of the loss (Cohen et al., 2021).

Stability threshold. On a quadratic objective $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{H}\mathbf{x}$ with curvature (*sharpness*) $s := \lambda_{\max}(\mathbf{H})$, gradient descent with step size η converges if and only if $\eta < 2/s$.

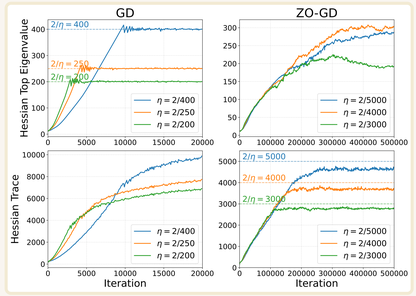
$\eta < 2/s$: converges $\qquad$ $\eta > 2/s$: diverges

Progressive sharpening. The sharpness $\lambda_{\max}(\mathbf{H}_t)$ rises during training until it reaches $2/\eta$, then stabilizes there — the ***edge of stability***.



First-Order vs. Zeroth-Order Stability

First-order stability depends only on the top eigenvalue; zeroth-order stability depends on the entire Hessian spectrum.



GD and ZO-GD on a CNN (CIFAR-10), varying η.

- For GD, the sharpness $\lambda_{\max}(\mathbf{H}_t)$ stabilizes at $2/\eta$.
- For ZO-GD it does not; instead the Hessian trace $\mathrm{Tr}(\mathbf{H}_t) = \sum_i \lambda_i$ stabilizes slightly below $2/\eta$.
- Random directions $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ make ZO stability is a property of the *full spectrum*.

Q. Do ZO methods operate at the edge of stability?

A. Yes. ZO stability is governed by the Hessian **trace** ($\mathrm{Tr}(\mathbf{H}) \approx 2/\eta$), rather than the top eigenvalue ($\lambda_{\max}(\mathbf{H}) \approx 2/\eta$) of FO methods.

Mean-Square Linear Stability

We derive exact mean-square thresholds for all three ZO methods, and compare them directly with first-order (FO) stability.

Theorem (informal)

With $\mathbf{M}$ the (preconditioned) Hessian, FO methods are stable iff $\eta < \dfrac{c_1}{\lambda_{\max}(\mathbf{M})}$, while the ZO mean-square threshold ($\eta < \eta^\star_{\mathrm{ms}} \iff \sup_t \mathbb{E}\|\mathbf{x}_t^2\| < \infty$) satisfies

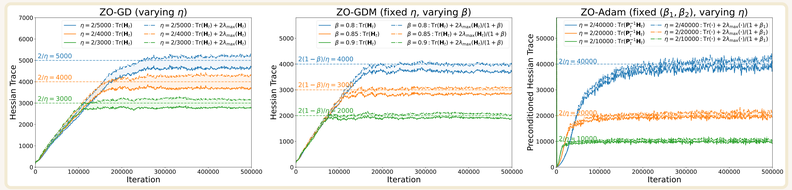
$$\frac{c_2}{\mathrm{Tr}(\mathbf{M}) + a\,\lambda_{\max}(\mathbf{M})} \le \eta^\star_{\mathrm{ms}} \le \frac{c_2}{\mathrm{Tr}(\mathbf{M})}.$$

		FO	ZO	
	$\mathbf{M}$	c_1	c_2	a
GD	$\mathbf{H}$	2	2	2
GDM	$\mathbf{H}$	$2(1+\beta)$	$2(1-\beta)$	$\frac{2}{1+\beta}$
Adam	$\mathbf{P}^{-1}\mathbf{H}$	$\frac{2(1+\beta_1)}{1-\beta_1}$	2	$\frac{2}{1+\beta_1}$

FO stability scales with $1/\lambda_{\max}$, ZO with $1/\mathrm{Tr}$; momentum reverses sign ($1+\beta$ vs $1-\beta$).

ZO Training at the Mean-Square Edge of Stability

Across optimizers and architectures, ZO training stabilizes at the theoretically predicted trace-based boundary — the *mean-square edge of stability*.



Full-batch ZO-GD, ZO-GDM, and ZO-Adam on a CNN.

Solid: lower band; dash-dot: upper band; dashed: predicted threshold.

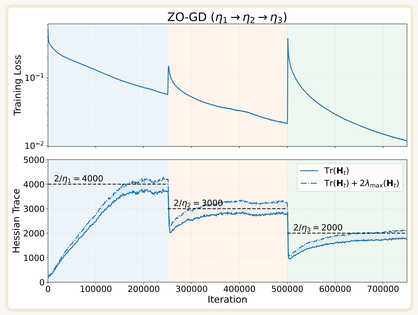
- After progressive sharpening, the curvature adapts so that the threshold remains within the predicted band throughout training.
- The (preconditioned) Hessian trace is the dominant stability signal for all three optimizers.
- The same behavior holds for ResNet-20 and ViT, and for LSTM and Mamba sequence models.

Implicit regularization of the Hessian trace.

Large step sizes bias ZO training toward solutions with **small trace**, whereas FO methods regularize only the top eigenvalue.

Catapult Dynamics

Increasing step size η mid-training induces a loss spike (the "catapult"), after which the trace re-equilibrates at the new threshold.



At each increase $\eta_1 \to \eta_2 \to \eta_3$, the step size temporarily exceeds the stability threshold, so the loss spikes; the trace $\mathrm{Tr}(\mathbf{H}_t)$ then drops and climbs back to the new $2/\eta$.