Enhancing LLM Training via Spectral Clipping

Xiaowen Jiang[1] Andrei Semenov[2] Sebastian U. Stich[1]

[1]CISPA Helmholtz Center for Information Security [2]EPFL

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).

Background

Let $\mathbf{X}_0 \in \mathbb{R}^{m\times n}$. The standard update rule with decoupled weight decay, for $k = 0, 1, \ldots$:

$$\mathbf{X}_{k+1} \;=\; (1-\lambda\eta_k)\,\mathbf{X}_k \;-\; \eta_k\,\mathbf{U}_k.$$

Here $\lambda \geq 0$ is the weight-decay factor, $\eta_k > 0$ the learning rate, and $\mathbf{U}_k \in \mathbb{R}^{m\times n}$ is the update matrix produced by the optimizer.

Common optimizers, e.g., AdamW, Signum, AdEMAMix, MARS, Muon, differ in how $\mathbf{U}_k$ is computed from the stochastic gradients.

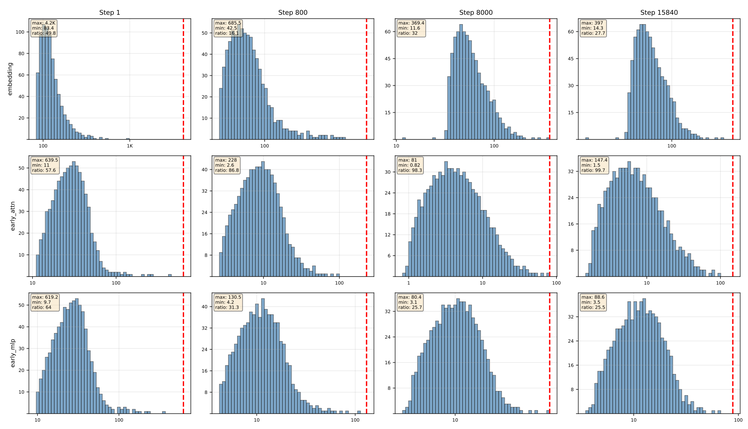
Phenomenon I — Excessive update spectral norms

Spike in $\|\mathbf{U}_i\|_2$ may propagate to the weights:

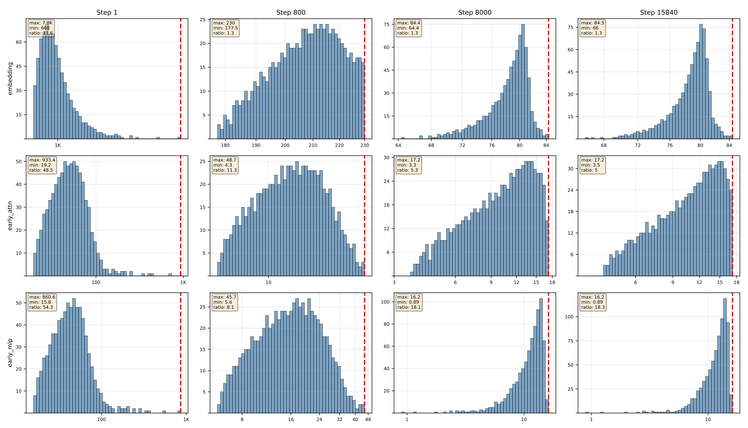
$$\|\mathbf{X}_k\|_2 \leq (1-\lambda\eta)^k\|\mathbf{X}_0\|_2 + \tfrac{1-(1-\lambda\eta)^k}{\lambda}\max_{i<k}\|\mathbf{U}_i\|_2.$$

For **sign-based** updates, (e.g., $\mathbf{U}_k = \operatorname{sign}(\mathbf{M}_k)$, $\mathbf{M}_k$ is the momentum, $\sqrt{\max(m,n)} \leq \|\mathbf{U}_k\|_2 \leq \sqrt{mn}$), $\|\mathbf{U}_i\|_2$ can be large.

Risk: Training instability, large final weight norms, leading to small learning rates, extended warm-up, slower convergence and worse generalization.



AdamW — update spectrum distribution with no hard ceiling.

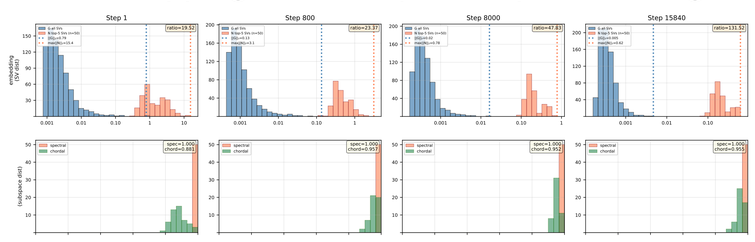


Spectra-AdamW — spectral norms strictly capped.

Phenomenon II — Sparse spectral spikes

Decompose the stochastic gradient $\mathbf{g} = \mathbf{G} + \mathbf{N}$ into signal $\mathbf{G}$ and noise $\mathbf{N}$. The noise often carries *low-rank spectral spikes*: a few singular values orders of magnitude larger than the signal, in directions *nearly orthogonal* to the signal subspace.

Risk: loss spikes / training instability; global / coordinate clipping cannot separate spike from signal.



Top row: Singular value (SV) distributions — blue = SVs of signal $\sigma(\mathbf{G})$, red = top-5 SVs of noise over 50 samples. Bottom row: spectral / chordal subspace distance between noise top-5 directions and signal directions ($\to 1$ = orthogonal). Confirmation of sparse spectral spikes.

Remedy: Spectral Clipping

Let $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$ (compact SVD), $q = \min(m,n)$, and $\operatorname{clip}_c(x) = \operatorname{sign}(x)\min(|x|, c)$.

$$\operatorname{clip}_c^{\mathrm{sp}}(\mathbf{X}) \;:=\; \mathbf{U}\,\operatorname{diag}\big(\operatorname{clip}_c(\sigma_1),\ldots,\operatorname{clip}_c(\sigma_q)\big)\,\mathbf{V}^\top$$

By construction $\|\operatorname{clip}_c^{\mathrm{sp}}(\mathbf{X})\|_2 \leq c$.

Post-Spectral Clipping

Apply $\operatorname{clip}_{c_k}^{\mathrm{sp}}$ to the optimizer's update $\mathbf{U}_k$. For a scaling factor $\alpha > 0$:

$$\mathbf{X}_{k+1} \;=\; (1-\lambda\eta_k)\,\mathbf{X}_k \;-\; \alpha\,\eta_k\;\operatorname{clip}_{c_k}^{\mathrm{sp}}(\mathbf{U}_k).$$

The effective update direction has spectral norm bounded by αc_k — an explicit control.

Equivalence with Composite Frank–Wolfe

Suppose $\mathbf{U}_k = \mathbf{M}_k$, Post-Spectral Clipping can be reformulated as composite Frank–Wolfe:

$$\begin{aligned}\mathbf{V}_{k+1} \;&\in\; \arg\min_{\mathbf{X}\in Q_2}\big\{\langle\mathbf{M}_k,\mathbf{X}\rangle + \psi(\mathbf{X})\big\},\\ \mathbf{X}_{k+1} \;&=\; (1-\gamma_k)\,\mathbf{X}_k + \gamma_k\,\mathbf{V}_{k+1},\end{aligned}$$

with $\gamma_k = \lambda\eta_k$, $c_k \equiv \lambda D_2/\alpha$ and $\psi(\mathbf{X}) = \frac{\lambda}{2\alpha}\|\mathbf{X}\|_F^2$. The method implicitly solves:

$$\begin{aligned}&\min_{\mathbf{X}\in Q_2}\big\{F(\mathbf{X}) := f(\mathbf{X}) + \psi(\mathbf{X})\big\},\\ &Q_2 := \{\mathbf{X}\in\mathbb{R}^{m\times n} : \|\mathbf{X}\|_2 \leq D_2\}.\end{aligned}$$

A Family of Spectral Optimizers. With different choices of ψ, we obtain various new updates rules (see discussions in the paper). With $\psi(\mathbf{X}) = \frac{b}{2}\|\mathbf{X}\|_\infty^2$, an approximate FW solution gives $\mathbf{V}_{k+1} \approx -\frac{\|\mathbf{M}_k\|_1}{b}\operatorname{clip}_c^{\mathrm{sp}}(\operatorname{sign}(\mathbf{M}_k))$ — spectral clipping on top of $\operatorname{sign}(\mathbf{M}_k) \Rightarrow$ **Spectra-Signum**.

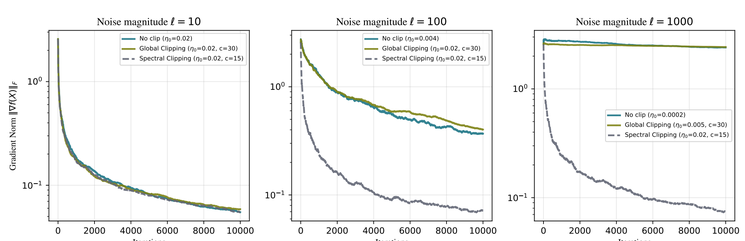
Connection to Muon. Let $\alpha \to \infty$ and $c = 1/\alpha$. Then $\frac{1}{c}\operatorname{clip}_c^{\mathrm{sp}}(\mathbf{X}) = \operatorname{orth}(\mathbf{X}) = \mathbf{U}_X\mathbf{V}_X^\top$, so Post-Spectral Clipping *recovers Muon* as the unregularized special case when $\mathbf{U}_k = \mathbf{M}_k$.

Pre-Spectral Clipping — Robustness to spectral spikes

We can also apply $\operatorname{clip}_c^{\mathrm{sp}}$ to the *raw* stochastic gradients. Suppose $\mathbf{g} = \mathbf{G} + \mathbf{N}$ with $\mathbf{N} = \ell\,\mathbf{U}_N\mathbf{V}_N^\top$, where $\mathbf{G}$ is the signal, $\ell \gg \|\mathbf{G}\|_2$ is the spike magnitude, and $\mathbf{U}_N \in \mathbb{R}^{m\times r}$, $\mathbf{V}_N \in \mathbb{R}^{n\times r}$ are zero-mean random matrices with orthonormal columns. By matrix perturbation:

$$\operatorname{clip}_c^{\mathrm{sp}}\big(\mathbf{G} + \ell\,\mathbf{U}_N\mathbf{V}_N^\top\big) \;\approx\; \mathbf{G} \;+\; c\,\mathbf{U}_N\mathbf{V}_N^\top.$$

With spectral clipping, variance reduced from $r\ell^2$ to rc^2 and true signal preserved. Global clipping cannot achieve both — it either suppresses the signal or fails to bound variance.



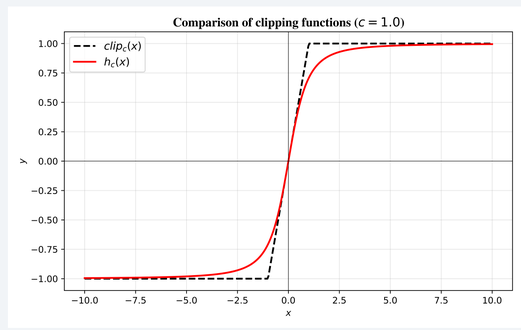
Synthetic rank-r spike noise: spectral-clipping keeps the same lr across various noise scales ℓ; vanilla SGD / global-clip significantly slow down.

SPECTRA framework

SPECTRA, a simple wrapper that consistently improves coordinate-wise optimizers (AdamW, Signum, etc.) via Post-Soft Spectral Clipping on updates and (optionally) Pre-Spectral Clipping on raw stochastic gradients.

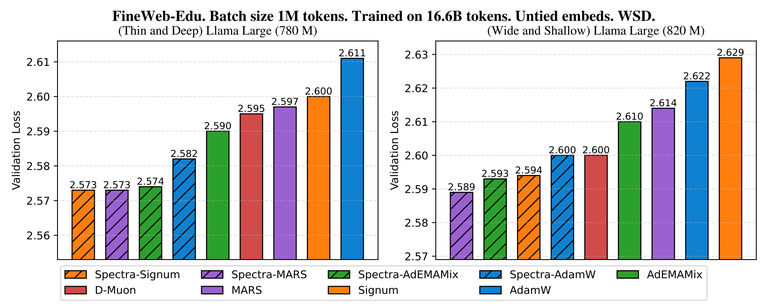
Soft Spectral Clipping. To avoid exact SVD, we use the smooth approximation $h_c(x) = x/\sqrt{1 + x^2/c^2}$ applied to singular values:

$$H_c(\mathbf{X}) \;=\; \big(\mathbf{I} + \mathbf{X}\mathbf{X}^\top/c^2\big)^{-1/2}\mathbf{X} \;\approx\; \operatorname{clip}_c^{\mathrm{sp}}(\mathbf{X})$$



Inverse square root computed via **Newton method** — pure matmul on (small $m \times m$) square matrices, GPU friendly.

LLM Pretraining



Llama-style 780M (deep & thin) and 820M (wide & shallow), Chinchilla-optimal duration (≈16.4B tokens, FineWeb-Edu).

Overhead vanishes at scale

Method	160M	780M	820M
AdamW + SPECTRA	+18.7%	+1.6%	+2.1%
Signum + SPECTRA	+18.7%	+2.0%	+2.4%
AdEMAMix + SPECTRA	+16.6%	+1.8%	+2.5%
MARS + SPECTRA	+17.8%	+1.6%	+2.0%
D-Muon (NS-based)	+8.8%	+1.4%	+2.1%
SOAP (eigendecomp.)	+49%	+21%	+10%

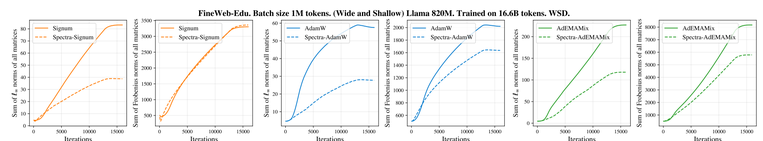
Per-step overhead vs. base optimizer (4× H100). At ≥ 780M, SPECTRA's additional cost amortizes to ∼2%.

SPECTRA wins even with extra baseline steps

	Baseline (extra steps)				SPECTRA
Optimizer	16k	+5%	+10%	+20%	16k
AdamW	2.608	2.605	2.600	2.590	**2.582**
Signum	2.600	2.594	2.589	2.580	**2.573**
AdEMAMix	2.591	2.585	2.580	2.571	**2.574**
MARS	2.598	2.592	2.587	2.579	**2.573**

780M validation loss. Baselines are given +5–20% extra training steps to compensate SPECTRA's ∼2% overhead.

Implicit weight regularization



820M weight ℓ_∞ norms over training. SPECTRA models end up with consistently smaller weight norms.