Adaptive Batch Sizes Using Non-Euclidean Gradient Noise Scales for Sign & Spectral Descent

Hiroki Naganuma[1,2] Shagun Gupta[3] Youssef Briki[2] Ioannis Mitliagkas[1,2] Irina Rish[1,2] Parameswaran Raman[3] Hao-Jun Michael Shi[3]
[1]Mila, Quebec AI Institute [2]Université de Montréal [3]Meta Superintelligence Labs

TL;DR: Design principled gradient noise scale (GNS) based on the optimizer's dual metric to determine the batch size B at each step. Reaches the same loss in fewer steps by increasing the batch size during training.

Motivation

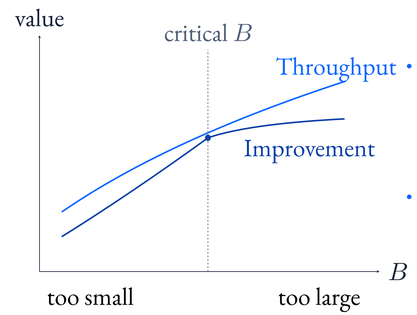
- Large batch sizes improve throughput, but impact model quality (poor statistical efficiency).
- A principled approach estimates the Critical Batch Size (CBS) using Gradient Noise Scale (GNS) to determine the batch size B.
- Q: How do we extend GNS for non-Euclidean optimizers?

Limitations of Euclidean GNS

$$\mathcal{B}_{\ell_2} = \frac{\mathrm{tr}(C_k)}{\|\nabla\mathcal{L}(x_k)\|_2^2}, \qquad B_k = \theta^{-2}\mathcal{B}(x_k), \qquad \theta \in (0,1].$$

- Euclidean signal. $\|\nabla\mathcal{L}(x_k)\|_2^2$ measures true gradient signal in ℓ_2.
- Euclidean noise. $\mathrm{tr}(C_k) = \mathbb{E}[\|g_k - \nabla\mathcal{L}(x_k)\|_2^2]$ measures stochastic gradient noise in the same geometry.
- Mismatch. For SIGNUM and MUON, the steepest-descent geometry is not ℓ_2; making Euclidean measurements incompatible.

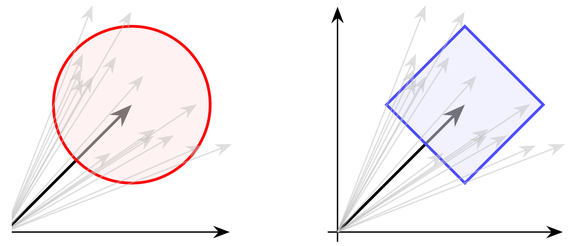
Critical Batch Size



- Before critical B. More samples reduce the variance in the gradient estimator.
- After critical B. More samples add redundant information.
- Instantaneous. GNS should be defined for each step k, not across entire training run.
- Moving target. As gradient noise and signal change, target batch size B_k shifts.
- Tolerance θ. $\theta \in (0, 1]$ sets the target noise-to-signal ratio, not a raw noise level.

Proposal: Replace $\mathcal{B}_{\ell_2}$ with $\mathcal{B}_*$: Euclidean → dual norm.

Geometry-Aware GNS



SGD: ℓ_2 Sign / Spectral

Key idea. Measure GNS in the optimizer's dual norm.

$$\mathcal{B}_*(x_k) = \left(\frac{\text{dual-norm noise}}{\text{dual-norm signal}}\right)^2, \qquad B_k = \theta^{-2}\mathcal{B}_*(x_k), \qquad \theta \in (0,1].$$

Class	Primal	Dual	GNS	Covariance
SGD	ℓ_2	ℓ_2	$\mathrm{tr}(C_k)/\|\nabla\mathcal{L}(x_k)\|_2^2$	$C_k = \mathbb{E}_k[(g_k - \nabla\mathcal{L}(x_k))(g_k - \nabla\mathcal{L}(x_k))^T]$
Sign-based	ℓ_∞	ℓ_1	$\|\sigma_k\|_1^2/\|\nabla\mathcal{L}(x_k)\|_1^2$	$[\sigma_k]_i^2 = \mathbb{E}_k\left[([\nabla\ell(x_k;\xi) - \nabla\mathcal{L}(x_k)]_i)^2\right]$
Spectral	$\mathcal{S}_\infty$	$\mathcal{S}_1$	$\|C_{\mathrm{row}}^{1/2}\|_{\mathcal{S}_1}^2/\|\nabla\mathcal{L}(x_k)\|_{\mathcal{S}_1}^2$	$C_{\mathrm{row},k} = \mathbb{E}_k[(G_k - \nabla\mathcal{L}(X_k))(G_k - \nabla\mathcal{L}(X_k))^\top]$

Derivation. Measure bias in g_k in the optimizer's dual norm $\|\cdot\|_*$ based on inequality:

$$\mathbb{E}_k\langle\nabla\mathcal{L}(x_k), p_k\rangle \geq \|\nabla\mathcal{L}(x_k)\|_* - \mathbb{E}_k\|\nabla\mathcal{L}(x_k) - g_k\|_*.$$

given stochastic steepest descent direction $p_k = \arg\max_{\|p\|\leq 1}\langle g_k, p\rangle$.

Adaptive Batch Size Rule

Practical online rule

- Use small batch during learning rate warmup.
- Estimate GNS every P steps; smooth with EMA.
- Enforce monotonic batch size growth.

Distributed estimator

- Use per-rank mini-batch gradients as independent samples in data parallelism.
- Reuse gradient AllReduce; no extra backward passes.

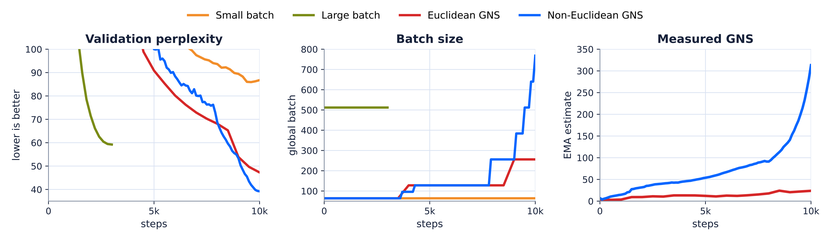
$$B_{k+1} = \max\left\{\left\lceil\frac{N_k}{\theta^2 M_k}\right\rceil, B_k\right\}, \qquad \eta_{k+1} \propto \sqrt{B_{k+1}}.$$

- N_k/M_k are smooth estimates of numerator/denominator of $\mathcal{B}_*(x_k)$.
- $\eta_k > 0$ is the learning rate.

Experimental Setup

- 160M Llama-3/C4 and SimpleViT/ImageWoof.
- Compare fixed batches, Euclidean GNS, and optimizer-matched GNS (ℓ_1 for sign-based, $\mathcal{S}_1$ for spectral).
- Step reduction is measured at the baseline minimum validation loss.

Results



160M Llama-3 on C4.
Non-Euclidean GNS grows B_k early and lowers perplexity faster.

Optimizer	Validation loss ↓				Steps red.
	B=64	B=256	Eucl.	Non-Eucl.	
SIGNSGD	$4.139_{\pm.024}$	$3.936_{\pm.006}$	$3.891_{\pm.012}$	$3.840_{\pm.001}$	22.6%
SIGNUM	$3.374_{\pm.005}$	$3.446_{\pm.008}$	$3.384_{\pm.006}$	$3.371_{\pm.003}$	66.6%
ADAMW	$3.299_{\pm.005}$	$3.323_{\pm.010}$	$3.313_{\pm.010}$	$3.303_{\pm.005}$	67.1%
SPECSGD	$3.749_{\pm.010}$	$3.940_{\pm.007}$	$3.744_{\pm.027}$	$3.729_{\pm.009}$	16.5%
MUON	$3.304_{\pm.002}$	$3.341_{\pm.002}$	$3.318_{\pm.007}$	$3.306_{\pm.003}$	66.8%

160M Llama-3 on C4.

Optimizer	Validation loss ↓			Steps red.
	B=128	B=512	Adaptive	
MSGD	$1.197_{\pm.013}$	$1.454_{\pm.012}$	$1.185_{\pm.004}$	4.1%
SGD	$1.483_{\pm.019}$	$1.647_{\pm.008}$	$1.318_{\pm.024}$	40.3%
SIGNSGD	$1.200_{\pm.009}$	$1.515_{\pm.008}$	$1.193_{\pm.006}$	27.4%
SIGNUM	$1.299_{\pm.022}$	$1.511_{\pm.005}$	$1.232_{\pm.003}$	11.3%
ADAMW	$0.975_{\pm.015}$	$1.065_{\pm.005}$	$0.934_{\pm.003}$	10.9%
MUON	$0.658_{\pm.022}$	$0.716_{\pm.006}$	$0.643_{\pm.006}$	57.7%

SimpleViT on ImageWoof.