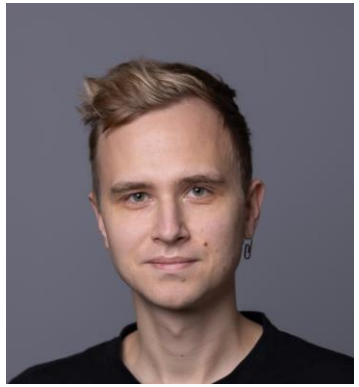
KAUST
King Abdullah University of Science and Technology

Rennala NSGD: Asynchronous Stochastic Optimization Beyond Euclidean Geometry







Igor Sokolov

Alexander Tyurin

Peter Richtárik

NeurIPS 2026
40th Conference on Neural Information Processing Systems

PDF

Igor Sokolov, Alexander Tyurin and Peter Richtárik
Rennala NSGD: Asynchronous Stochastic Optimization Beyond Euclidean Geometry.
NeurIPS 2026 (poster), arXiv:2610.38773, 2026.

Stochastic optimization in normed spaces

$$\min_{x \in E} \left\{ f(x) := \mathbb{E}_{\xi \sim \mathcal{D}} \left[f(x; \xi) \right] \right\}$$

Stochastic optimization in normed spaces

Smooth and lower bounded, possibly non-convex

$$\min_{x \in E} \{ f(x) := \mathbb{E}_{\xi \sim \mathcal{D}} [f(x; \xi)] \}$$

Finite-dimensional normed space

Data distribution

Stochastic optimization in normed spaces

Smooth and lower bounded, possibly non-convex

$$\min_{x \in E} \{ f(x) := \mathbb{E}_{\xi \sim \mathcal{D}} [f(x; \xi)] \}$$

Finite-dimensional normed space

Data distribution

Examples	Space	Update norm
Vectors	$\mathbb{R}^d$	ℓ_p norm $\|x\|_p := \left(\sum_{i=1}^d |x_i|^p \right)^{1/p}$
Matrices	$\mathbb{R}^{m \times n}$	Schatten-p norm $\|X\|_{S_p} := \left(\sum_{i=1}^r \sigma_i(X)^p \right)^{1/p}$

$1 \leq p < \infty, \qquad r = \min\{m, n\}$

Minibatch SGD in an arbitrary norm

$$x^{k+1} \leftarrow x^k - \gamma_k (g^k)^\sharp \qquad g^k := \frac{1}{B} \sum_{j=1}^{B} \nabla f(x^k; \xi_j^k)$$

Fresh stochastic gradients

Minibatch SGD in an arbitrary norm

$$x^{k+1} \leftarrow x^k - \gamma_k (g^k)^\sharp \qquad g^k := \frac{1}{B}\sum_{j=1}^{B} \nabla f(x^k; \xi_j^k)$$

Fixed measurable sharp selection

Fresh stochastic gradients

$$s^\sharp \in \partial\phi^*(s) \qquad \phi(x) := \tfrac{1}{2}\|x\|^2 \qquad \phi^*(s) = \tfrac{1}{2}\|s\|_*^2$$

Minibatch SGD in an arbitrary norm

$$x^{k+1} \leftarrow x^k - \gamma_k (g^k)^\sharp \qquad g^k := \frac{1}{B}\sum_{j=1}^{B} \nabla f(x^k; \xi_j^k)$$

Fixed measurable sharp selection

Fresh stochastic gradients

$$s^\sharp \in \partial \phi^*(s) \qquad \phi(x) := \tfrac{1}{2}\|x\|^2 \qquad \phi^*(s) = \tfrac{1}{2}\|s\|_*^2$$

Possibly non-Euclidean norm

Its conjugate norm

Minibatch SGD in an arbitrary norm

$$x^{k+1} \leftarrow x^k - \gamma_k (g^k)^\sharp \qquad g^k := \frac{1}{B}\sum_{j=1}^{B} \nabla f(x^k; \xi_j^k)$$

Fixed measurable sharp selection

Fresh stochastic gradients

$$s^\sharp \in \partial\phi^*(s) \qquad \phi(x) := \tfrac{1}{2}\|x\|^2 \qquad \phi^*(s) = \tfrac{1}{2}\|s\|_*^2$$

Connection	Update norm	Sharp direction
SGD	$\|\cdot\|_2$	g
signSGD direction	$\|\cdot\|_\infty$	$\|g\|_1 \ \mathrm{sign}(g)$
Muon direction	$\|\cdot\|_{\mathrm{op}}$	$\|G\|_{\mathrm{nuc}} UV^\mathsf{T} \quad G = U\Sigma V^\mathsf{T}$

Directions agree up to scaling. Practical Muon also uses momentum and an approximate polar map.

Noise measured in the native dual norm

$$\eta(x;\xi) := \nabla f(x;\xi) - \nabla f(x)$$

$$\mathbb{E}_{\xi\sim\mathcal{D}}[\eta(x;\xi)] = 0$$

Unbiased gradient noise

Noise measured in the native dual norm

$$\eta(x;\xi) := \nabla f(x;\xi) - \nabla f(x)$$
$$\mathbb{E}_{\xi\sim\mathcal{D}}[\eta(x;\xi)] = 0$$

Unbiased gradient noise

$$\mathbb{E}_{\xi\sim\mathcal{D}}\left[\exp\left(\frac{\|\eta(x;\xi)\|_*^2}{\sigma^2}\right)\right] \le e$$

Light-tail assumption

Noise measured in the native dual norm

$$\eta(x;\xi) := \nabla f(x;\xi) - \nabla f(x)$$
$$\mathbb{E}_{\xi\sim\mathcal{D}}[\eta(x;\xi)] = 0$$

$$\mathbb{E}_{\xi\sim\mathcal{D}}\left[\exp\left(\frac{\|\eta(x;\xi)\|_*^2}{\sigma^2}\right)\right] \leq e$$

Can we obtain sharper convergence guarantees
by measuring stochastic gradient noise directly in the dual norm,
rather than in the Euclidean norm followed by norm conversion?

Noise measured in the native dual norm

$$\eta(x;\xi) := \nabla f(x;\xi) - \nabla f(x)$$
$$\mathbb{E}_{\xi\sim\mathcal{D}}[\eta(x;\xi)] = 0$$

$$\mathbb{E}_{\xi\sim\mathcal{D}}\left[\exp\left(\frac{\|\eta(x;\xi)\|_*^2}{\sigma^2}\right)\right] \le e$$

Can we obtain sharper convergence guarantees
by measuring stochastic gradient noise directly in the dual norm,
rather than in the Euclidean norm followed by norm conversion?

The answer depends on the geometry of the space
and the stochastic noise model.

Oracle Complexity Result

$$N_* = \tilde{\mathcal{O}}\left(\frac{L\Delta}{\varepsilon} + \frac{\kappa\, L\Delta\, \sigma^2}{\varepsilon^2} \right)$$

Oracle Complexity Result

$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|$

$\Delta = f(x^0) - f_{\inf}$

Regularity of the dual space

Noise in the dual norm

$$N_* = \tilde{\mathcal{O}}\left(\frac{L\Delta}{\varepsilon} + \frac{\kappa L\Delta\sigma^2}{\varepsilon^2}\right)$$

First-order stationarity condition

$$\min_{0\leq k<T} \|\nabla f(x^k)\|_*^2 \leq \varepsilon$$

with probability $1 - \delta$

Geometry and Noise: A Direct Comparison

$$N_* = \widetilde{\mathcal{O}}\left(\frac{L\Delta}{\varepsilon} + \frac{\kappa\, L\Delta\, \sigma^2}{\varepsilon^2}\right)$$

Geometry and Noise: A Direct Comparison

$$N_* = \tilde{\mathcal{O}}\left(\frac{L\Delta}{\varepsilon} + \frac{\kappa\, L\Delta\, \sigma^2}{\varepsilon^2}\right)$$

Statistical term

Geometry and Noise: A Direct Comparison

$$N_* = \widetilde{\mathcal{O}}\left(\frac{L\Delta}{\varepsilon} + \frac{\kappa\, L\Delta\, \sigma^2}{\varepsilon^2} \right)$$

Statistical term

Perturbed-gradient model

$$\nabla f(x;\xi) = \nabla f(x) + \zeta(\xi)$$

Independent coordinates

$\zeta_i = \nu u_i, \quad u_i \in \{-1, +1\}$

$\zeta_i = \nu u_i, \quad u_i \sim \mathcal{N}(0,1)$

Dual space $(E^*, \|\cdot\|_*)$	**Euclidean / Frobenius + norm conversion**	**Geometry-native Rademacher**	**Geometry-native Gaussian**
$(\mathbb{R}^d, \|\cdot\|_1)$	d^2	d^3	d^3
$(\mathbb{R}^d, \|\cdot\|_2)$	d	d	d
$(\mathbb{R}^d, \|\cdot\|_\infty)$	d	$\ln(ed)$	$(\ln(ed))^2$

Orders shown; common factor (logarithmic confidence terms omitted): $\nu^2 L\Delta/\varepsilon^2$

Rennala NSGD: Semi-Asynchronous Optimization

Algorithm 1 Rennala NSGD (server)

Input: initial point $x^0 \in E$, stepsizes $\{\gamma_k > 0\}$, minibatch size B, number of iterations T.
Start Algorithm 2 on all N workers.
for $k = 0, 1, \ldots, T-1$ **do**
  Set $g^k = 0$ and $b = 0$.
  while $b < B$ **do**
    Receive the next gradient message (g, k') from any worker.
    if $k' = k$ **then**
      $g^k \leftarrow g^k + \frac{1}{B} g$ and $b \leftarrow b + 1$.
    end if
    if $b < B$ **then**
      Send (x^k, k) to that worker.
    end if
  end while
  $x^{k+1} = x^k - \gamma_k \left(g^k\right)^\sharp$.
  Send $(x^{k+1}, k+1)$ to the worker whose gradient completed the minibatch.
end for

Algorithm 2 Rennala NSGD (worker loop)

Set $g = 0$ and $k' = -1$.
loop
  Send (g, k') to the server.
  Receive (x^k, k) from the server.
  Set $k' = k$.
  Sample $\xi \sim \mathcal{D}$.
  Compute $g = \nabla f(x^k; \xi)$.
end loop

Wall-Clock Time Complexity

$$T_{\text{total}} = \widetilde{\mathcal{O}}\left(\min_{S \in [N]} \left[\left(\frac{1}{S} \sum_{i=1}^{S} \frac{1}{\tau_i} \right)^{-1} \left(\frac{L\Delta}{\varepsilon} + \frac{\kappa L \Delta \sigma^2}{S\varepsilon^2} \right) \right] \right)$$

More details can be found in the paper.

Thank you for your attention.