Rennala NSGD: Asynchronous Stochastic Optimization Beyond Euclidean Geometry

Igor Sokolov[1] Alexander Tyurin[2, 3] Peter Richtárik[1]

[1]KAUST [2]AXXX [3]Applied AI Institute

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

The problem

Smooth and lower bounded, possibly non-convex

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

Finite-dimensional normed space — Data distribution

Examples	Space E	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: $\|X\|_{S_p} := \left(\sum_{i=1}^r \sigma_i(X)^p\right)^{1/p}$

Here, $1 \le p < \infty$, $r := \min\{m, n\}$, and $\sigma_i(X)$ are the singular values of X. For $p = \infty$, the norms are $\max_i |x_i|$ and $\max_i \sigma_i(X)$, respectively.

Minibatch SGD in an arbitrary norm

Definition 1 (Sharp operator). The sharp operator is a fixed Borel-measurable selection $s^\sharp \in \partial\phi^*(s)$, where $\phi(x) := \frac{1}{2}\|x\|^2$ and $\phi^*(s) = \frac{1}{2}\|s\|_*^2$, with a possibly non-Euclidean norm $\|\cdot\|$ and its dual norm $\|\cdot\|_*$.

$$x^{k+1} \leftarrow x^k - \gamma_k (g^k)^\sharp, \qquad g^k := \frac{1}{B}\sum_{j=1}^B \nabla f(x^k;\xi_j^k) \in E^*. \tag{7}$$

At each iteration, draw B fresh i.i.d. samples, independently of the past.

Update norm	$g^\sharp$	Connection
$\|\cdot\|_2$	g	SGD
$\|\cdot\|_\infty$	$\|g\|_1 \operatorname{sign}(g)$	signSGD direction
$\|\cdot\|_{\text{op}}$	$\|g\|_{\text{nuc}} UV^\top$	Muon direction

Directions agree up to scaling. In the matrix row, $g = U\Sigma V^\top$ is a compact SVD retaining only positive singular values, and $\|\cdot\|_{\text{nuc}}$ is the nuclear norm. The sign is componentwise, with $\operatorname{sign}(0) = 0$. Practical Muon also uses momentum and an approximate polar map.

Central question

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.

Assumptions

Assumption 1 (Smoothness). $f : E \to \mathbb{R}$ is differentiable and, for some $L > 0$,

$$\|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\|, \quad \forall x, y \in E.$$

Assumption 2 (Lower bound). There is $f^{\inf} \in \mathbb{R}$ such that

$$f(x) \ge f^{\inf} \quad \forall x \in E.$$

Assumption 3 (Dual-norm light-tailed oracle). For every $x \in E$, the stochastic gradient is unbiased, $\mathbb{E}_{\xi\sim\mathcal{D}}[\nabla f(x;\xi)] = \nabla f(x)$, and satisfies

$$\mathbb{E}_{\xi\sim\mathcal{D}}\left[\exp\left(\|\nabla f(x;\xi) - \nabla f(x)\|_*^2/\sigma^2\right)\right] \le \exp(1), \tag{8}$$

for some $\sigma > 0$.

Regularity framework

Definition 2 (Regularity of the dual space). $(E^*, \|\cdot\|_*)$ is κ-regular if there exist an auxiliary norm $\|\cdot\|_{*,+}$ and $\kappa_+ \in [1, \kappa]$ such that $\psi(s) := \frac{1}{2}\|s\|_{*,+}^2$ is continuously differentiable and

$$\psi(s+t) \le \psi(s) + \langle\nabla\psi(s), t\rangle + \tfrac{\kappa_+}{2}\|t\|_{*,+}^2, \qquad \|s\|_*^2 \le \|s\|_{*,+}^2 \le \frac{\kappa}{\kappa_+}\|s\|_*^2 \qquad \forall s, t \in E^*.$$

Paper: Definition 1 and Remark 1, specialized to E^*. Framework: Juditsky and Nemirovski (2008) [1].

Space $(E^*, \|\cdot\|_*)$	κ scaling	κ_+	Auxiliary norm $\|\cdot\|_{*,+}$
Any d-dimensional space	$\Theta(d)$ (worst case)	1	$\|Q_{\text{J}}^{-1}\cdot\|_2$, $Q_{\text{J}} := \operatorname{arg\,max}_{Q=Q^\top\succ 0,\, Q\mathbb{B}_2\subseteq B_*} \log\det Q$ [2]
$(\mathbb{R}^d, \|\cdot\|_p)$, $1 \le p < 2$	$\Theta(d^{2/p-1})$	1	$d^{1/p-1/2}\|\cdot\|_2$
$(\mathbb{R}^d, \|\cdot\|_p)$, $2 \le p \le \infty$	$\Theta(\min\{p, \ln(ed)\})$	$\rho - 1$	$\|\cdot\|_{\rho:=\min\{p,\,\widehat{\rho}(d)\}}$ [1]
$(\mathbb{R}^{m\times n}, \|\cdot\|_{S_p})$, $1 \le p < 2$	$\Theta(r^{2/p-1})$	1	$r^{1/p-1/2}\|\cdot\|_{S_2}$
$(\mathbb{R}^{m\times n}, \|\cdot\|_{S_p})$, $2 \le p \le \infty$	$\Theta(\min\{p, \ln(er)\})$	$\max\{2, \rho - 1\}$	$\|\cdot\|_{S_{\rho:=\min\{p,\,\max\{3,\,\widehat{\rho}(r)\}\}}}$ [1]

[1] $\widehat{\rho}(D) := \ln D + \sqrt{(\ln D)^2 - 2\ln D}$.
[2] Choose a basis to identify E^* with $\mathbb{R}^d$. $B_* := \{s : \|s\|_* \le 1\}$, $\mathbb{B}_2 := \{s : \|s\|_2 \le 1\}$. The John ellipsoid of B_* is $\mathcal{E}_{\text{J}} := Q_{\text{J}}\mathbb{B}_2$.
Paper: Table 3; Appendix I. $d > 8$, $r := \min\{m, n\} > 8$. Optimal κ scaling (worst-case universal row); admissible κ_+. Schatten $p = 2$: $\kappa_+ = 1$ suffices.

For dual ℓ_p^d spaces with $1 \le p < 2$, κ-regularity recovers classical ℓ_p–ℓ_2 norm equivalence:

$$\|s\|_*^2 \le \|s\|_{*,+}^2 = d^{2/p-1}\|s\|_2^2 \le d^{2/p-1}\|s\|_*^2 \iff \|s\|_2^2 \le \|s\|_*^2 \le d^{2/p-1}\|s\|_2^2.$$

For $2 < p \le \infty$, we compare the tuned proxy $\|\cdot\|_\rho$ from the table with the Euclidean proxy:

$$\text{Tuned:} \quad \|s\|_*^2 \le \|s\|_\rho^2 \le e^2\|s\|_*^2, \qquad \text{Euclidean:} \quad \|s\|_*^2 \le \|s\|_2^2 \le d^{1-2/p}\|s\|_*^2.$$

At $p = 2$, both proxies coincide. Analogous dimension dependence holds for Schatten-p norms, with $r = \min\{m, n\}$ and the Frobenius proxy. Paper: Appendix I.1–I.2, Propositions 3–4.

Key complexity result

Theorem 1 (Oracle complexity). Under Assumptions 1–3, let κ be a regularity constant of $(E^*, \|\cdot\|_*)$ and set $\Delta := f(x^0) - f^{\inf}$. Run (7) with $\gamma_k \equiv \frac{1}{2L}$, $T = \lceil\frac{128L\Delta}{15\varepsilon}\rceil$, and $B = \max\left\{\left\lceil\frac{16\sigma^2}{\varepsilon}\left(\sqrt{2\kappa} + \sqrt{6\log(T/\delta)}\right)^2\right\rceil, 1\right\}$. Then, with probability at least $1 - \delta$, after at most

$$N_* := BT = \widetilde{\mathcal{O}}\left(\frac{L\Delta}{\varepsilon} + \frac{\kappa L\Delta\sigma^2}{\varepsilon^2}\right) \tag{9}$$

stochastic-gradient evaluations, the method produces an iterate satisfying $\min_{0\le k<T}\|\nabla f(x^k)\|_*^2 \le \varepsilon$.

The bound holds for $0 < \varepsilon \le L\Delta$ and $\delta \in (0, 1)$. The tilde hides logarithmic factors in T and $1/\delta$. Paper: Theorem 1 and Corollary 1.

Euclidean proxy vs. native proxy

We compare how the noise norm in Assumption 3 affects the oracle bound (9) for the same perturbed-gradient oracle on $(E^*, \|\cdot\|_*) = (\mathbb{R}^d, \|\cdot\|_p)$:

$$\nabla f(x;\xi) = \nabla f(x) + \zeta(\xi), \qquad \begin{aligned}\text{Rademacher: } & \zeta = \nu\chi, \quad \Pr(\chi_i = \pm 1) = \tfrac{1}{2},\\ \text{Gaussian: } & \zeta = \nu\omega, \quad \omega_i \sim \mathcal{N}(0, 1).\end{aligned}$$

Here $\nu > 0$; the d coordinates are independent, with fresh noise at each oracle call.

Space $(E^*, \|\cdot\|_*)$	Euclidean [1]	Geometry-native, [2] Rademacher	Geometry-native, [2] Gaussian
$(\mathbb{R}^d, \|\cdot\|_1)$	$\widetilde{\mathcal{O}}(d^2L\Delta/\varepsilon^2)$	$\widetilde{\mathcal{O}}(d^3L\Delta/\varepsilon^2)$	$\widetilde{\mathcal{O}}(d^3L\Delta/\varepsilon^2)$
$(\mathbb{R}^d, \|\cdot\|_2)$	$\widetilde{\mathcal{O}}(dL\Delta/\varepsilon^2)$	$\widetilde{\mathcal{O}}(dL\Delta/\varepsilon^2)$	$\widetilde{\mathcal{O}}(dL\Delta/\varepsilon^2)$
$(\mathbb{R}^d, \|\cdot\|_\infty)$	$\widetilde{\mathcal{O}}(dL\Delta/\varepsilon^2)$	$\widetilde{\mathcal{O}}(\ln(ed)L\Delta/\varepsilon^2)$	$\widetilde{\mathcal{O}}((\ln(ed))^2L\Delta/\varepsilon^2)$

[1] Complexity derived using (8) in $\|\cdot\|_2$, then conversion to the dual norm induced by an arbitrary update norm $\|\cdot\|$.
[2] Complexity derived using (8) directly in the native dual norm $\|\cdot\|_*$ and the κ-regularity framework.
Paper: Table 1; Definitions 9–10; Appendix J.1. Noise terms only; ν^2 and $\widetilde{\mathcal{O}}(L\Delta/\varepsilon)$ omitted. Asymptotic orders: green smaller, red larger, black equal.

Rennala NSGD

Computing setup [2]. A server coordinates N parallel workers sharing the objective f and access to the oracle $\nabla f(x;\xi)$, $\xi \sim \mathcal{D}$, with fresh independent samples across calls and workers. After receiving a query, worker i takes a fixed τ_i seconds per gradient, with $0 < \tau_1 \le \cdots \le \tau_N$. Communication and server processing, including sharp updates, are instantaneous. Messages are processed sequentially; simultaneous returns follow a fixed worker order.

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
  $x^{k+1} = x^k - \gamma_k (g^k)^\sharp$.
  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

Theorem 2 (Wall-clock complexity). Under the computing setup above and the assumptions of Theorem 1, run Rennala NSGD with the prescribed γ_k, T, B. Then, with probability at least $1 - \delta$, after at most

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

seconds, the method produces an iterate satisfying $\min_{0\le k<T}\|\nabla f(x^k)\|_*^2 \le \varepsilon$.

Here S is an integer; $0 < \varepsilon \le L\Delta$ and $\delta \in (0, 1)$; the tilde hides logarithmic factors in T and $1/\delta$. Paper: Theorem 2.

The bound selects the best number of fastest workers; the algorithm needs neither their speeds nor the minimizing S.

References

[1] A. Juditsky and A. S. Nemirovski. *Large deviations of vector-valued martingales in 2-smooth normed spaces.* arXiv:0809.0813, 2008.

[2] A. Tyurin and P. Richtárik. *Optimal time complexities of parallel stochastic optimization methods under a fixed computation model.* NeurIPS 36, 2023.

Links

arXiv — Project page — NEURAL INFORMATION PROCESSING SYSTEMS