Near-Optimal Decentralized Stochastic Convex Optimization over Networks

Nitai Kluger[1*], Amit Attia[1*], Tomer Koren[1,2]

Tel Aviv University[1] Google Research[2] *Equal contribution

Blavatnik School of Computer Science and AI
Faculty of Exact Sciences
Tel Aviv University

Google

European Research Council
Established by the European Commission

TL;DR: Accelerating both the optimization and the gossip, via a one-step-delayed query, lets decentralized SGD keep the centralized $O(1/\sqrt{N})$ rate with up to $\sqrt{\rho}\,N^{3/4}$ workers (previously $\rho\sqrt{N}$), which we prove is optimal up to logarithmic factors

Motivation

Distributing SGD over M workers cuts the sequential rounds to N/M, ideally at no cost in accuracy. The **critical parallelism** is the largest M that preserves the centralized rate $O(\sigma R/\sqrt{N})$; with synchronized accelerated minibatching it is $N^{3/4}$ [5].

Synchronizing all workers at every step, through a server or all-reduce, is itself a bottleneck. Decentralized methods avoid it by gossiping only with neighbors, but workers then hold different copies of the model, and their disagreement **biases** the gradients, an effect acceleration is known to amplify [4]. The best decentralized guarantee so far was $\rho\sqrt{N}$ [3].

Global synchronization



every step waits for all M

Gossip



neighbors only; copies differ

Problem setup

Objective	minimize $f(x) = \frac{1}{M}\sum_{i=1}^{M} f_i(x)$, each f_i convex and L-smooth
Start	a common x_0 with $\|x_0 - x^*\| \le R$
Oracle	unbiased $G_i(x,\xi)$ for $\nabla f_i(x)$, with variance at most σ^2
Heterogeneity	bounded only at the optimum: $\frac{1}{M}\sum_{i=1}^{M}\|\nabla f_i(x^*)\|^2 \le \zeta_*^2$
Gossip	$X \mapsto XP$ with P symmetric, doubly stochastic, $1-\lambda_2(P) \ge \rho$
Budget	per round, 1 gradient + 1 gossip step per worker; N/M rounds

Main results vs previous work

Method	M_{crit} (up to logs)
Central server, accelerated	$N^{3/4}$
D-SGD [1]	$\sqrt{\rho}\,N^{1/4}$ if $\zeta_* \approx 0$, else $\rho\,N^{1/4}$
Gradient tracking [2]	$\sqrt{\rho}\,N^{1/4}$
DAT-SGD [3]	$\rho\sqrt{N}$
DDA-SGD (ours)	$\sqrt{\rho}\,N^{3/4}$
Lower bound (ours)	no method exceeds $\sqrt{\rho}\,N^{3/4}$

DDA-SGD: doubly accelerated

Algorithm sketch

Capitals stack the workers' vectors as columns, $X = [x_1, \dots, x_M]$; worker i holds column i

Input: P, block length B, momenta m_t, β_g, steps θ_t, η_t

for $t = 0, 1, \dots, T$ **do**

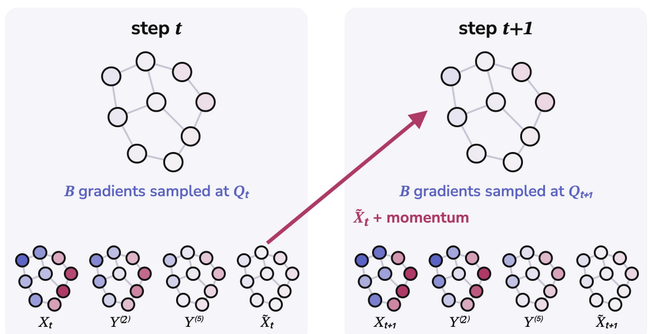
- gossip starts at the current copies: $Y^{(-1)} = Y^{(0)} = X_t$ — LOCAL
- **for** $s = 1, \dots, B$ **do** (all workers in parallel)
  - gradient at delayed query: $g_{t,i} \mathrel{+}= \frac{1}{B} G_i(q_{t,i}, \xi)$ — LOCAL
  - momentum point: $V = Y^{(s-1)} + \beta_g(Y^{(s-1)} - Y^{(s-2)})$ — LOCAL
  - gossip step: $Y^{(s)} = \tfrac{1}{2}V + \tfrac{1}{2}VP$ — GOSSIP
- mixed copies $\tilde{X}_t = Y^{(B)}$, next delayed query: $Q_{t+1} = \tilde{X}_t + m_t(1+m_{t+1})(\tilde{X}_t - \tilde{X}_{t-1})$ — LOCAL
- update: $Z_{t+1} = Z_t - \eta_t G_t$ — LOCAL
- $X_{t+1} = (1-\theta_t)X_t + \theta_t Z_{t+1}$ — LOCAL

Output: node i returns $\tilde{x}_{T,i}$, its column of $\tilde{X}_T$

Only multiplying by P communicates (one message per neighbor); everything else is local

Query Q_t
fixed for all B rounds

Gossip $Y^{(s)}$
mixing X_t round by round



below average — above average — pale = workers agree

Each node is one worker's copy. Real run of DDA-SGD on this 9-node network, $B = 8$.

References

[1] Koloskova, Loizou, Boreiri, Jaggi, Stich. A unified theory of decentralized SGD with changing topology and local updates. ICML 2020.

[2] Koloskova, Lin, Stich. An improved analysis of gradient tracking for decentralized machine learning. NeurIPS 2021.

[3] Eisen, Dorfman, Levy. Enhancing parallelism in decentralized stochastic convex optimization. ICML 2025.

[4] Devolder, Glineur, Nesterov. First-order methods of smooth convex optimization with inexact oracle. Math. Programming 2014.

[5] Dekel, Gilad-Bachrach, Shamir, Xiao. Optimal distributed online prediction using mini-batches. JMLR 2012.

Main results

Theorem 1 (upper bound). With $B = \tilde{\Theta}(1/\sqrt{\rho})$ gossip rounds per step and Nesterov momenta, every node i outputs $\tilde{x}_{T,i}$ with

$$\mathbb{E}[f(\tilde{x}_{T,i}) - f(x^*)] = \tilde{O}\Big(\frac{\sigma R}{\sqrt{N}} + \frac{LR^2M^2}{\rho N^2}\Big)$$

Statistical term: the centralized rate. **Network term:** accelerated in both N and ρ. $\tilde{O}$ hides $\mathrm{polylog}(N/\rho, (\zeta_* + \sigma)/LR)$ factors, the only place heterogeneity appears.

$\Longrightarrow$ the centralized rate $O\big(\frac{\sigma R}{\sqrt{N}}\big)$ is kept for all $M \le \sqrt{\rho}\,N^{3/4}$ workers

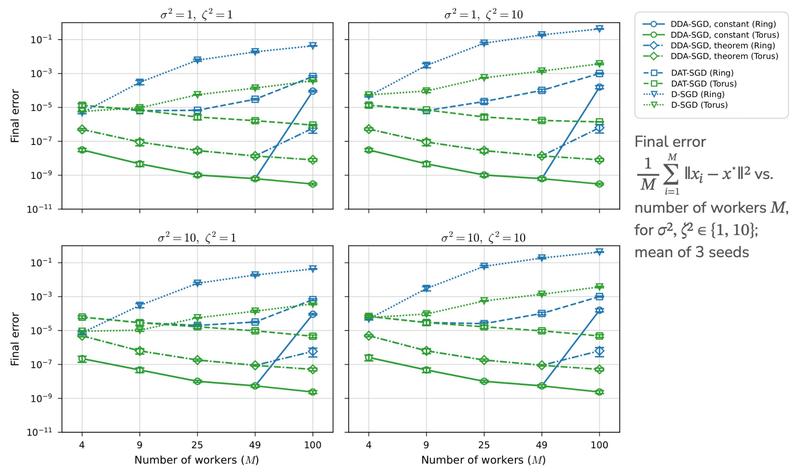
Theorem 2 (lower bound). On the M-node path, every deterministic linear-span method after S gossip rounds has

$$f(\tilde{x}_S) - f(x^*) = \Omega\Big(LR^2 \min\{1, \frac{1}{\rho S^2}\}\Big)$$

With $S = \Theta(N/M)$ this is $\Omega(LR^2M^2/\rho N^2)$; with the $\Omega(\sigma R/\sqrt{N})$ statistical limit, Theorem 1 is **optimal up to logarithmic factors**.

Experiments

Least squares ($d = 50$) on ring and torus networks, $M = 4, \dots, 100, 10^5$ rounds, vs. D-SGD [1] and DAT-SGD [3]: DDA-SGD has the lowest error in **every** configuration.



Final error $\frac{1}{M}\sum_{i=1}^{M}\|x_i - x^*\|^2$ vs. number of workers M, for $\sigma^2, \zeta^2 \in \{1, 10\}$; mean of 3 seeds

6×–1.6·10⁵×
lower error than DAT-SGD with constant momenta (median 1.6·10³×)

12.7×–7.3·10³×
lower error than DAT-SGD with the Theorem 1 schedule (median 94×)

100×
lower error on the torus from $M = 4$ to 100 workers ($\sigma^2 = \zeta^2 = 1$)

×1.00
median change in error when ζ^2 grows 10× (D-SGD: ×10)