On the Optimal Time Complexities in Decentralized Stochastic Asynchronous Optimization

Alexander Tyurin KAUST, AIRI, Skoltech

Peter Richtárik KAUST

1. Distributed Stochastic Optimization Problem

Our objective is to address the nonconvex distributed optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

devices / workers

model parametes / features

Loss on local data $\mathcal{D}_i$ stored on device i

$f_i(x) = \mathrm{E}_{\xi\sim\mathcal{D}_i}[f_i(x;\xi)]$

2. Homogeneous Setting

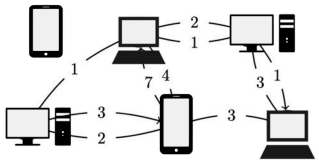
Homogeneous Setting: all workers store the same data,

i.e., $\mathcal{D}_i = \mathcal{D}$ and $f_i = f$ for all $i \in [n]$

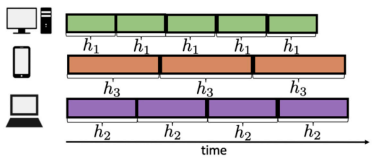
3. Decentralized Setup with Communication and Computations Times

We have n nodes with the associated computation times $\{h_i\}$, and communications times $\{\rho_{i\to j}\}$:
(i) It takes less or equal to $h_i \in [0,\infty]$ seconds to compute a stochastic gradient by the i^{th} node.
(ii) less or equal $\rho_{i\to j} \in [0,\infty]$ seconds to send *directly* a vector $v \in \mathbb{R}^d$ from the i^{th} node to the j^{th} node

Communication Takes Time

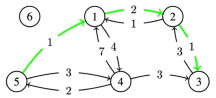


Computation Takes Time



For this setup, it would be convenient to define **the distance of the shortest path from worker i to worker j** :

$$\tau_{i\to j} := \min_{\text{path}\in P_{i\to j}} \sum_{(u,v)\in\text{path}} \rho_{u\to v} \in [0,\infty].$$



4. Assumptions

We work with the following standard assumption from smooth nonconvex stochastic optimization literature.

Assumption 1. *f is differentiable and L–smooth, i.e., $\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|\ \forall x,y\in\mathbb{R}^d$.*

Assumption 2. *There exist $f^* \in \mathbb{R}$ such that $f(x) \ge f^*$ for all $x \in \mathbb{R}^d$.*

Assumption 3. *For all $x \in \mathbb{R}^d$, stochastic gradients $\nabla f(x;\xi)$ are unbiased and σ^2-variance-bounded, i.e., $\mathbb{E}_\xi[\nabla f(x;\xi)] = \nabla f(x)$ and $\mathbb{E}_\xi[\|\nabla f(x;\xi) - \nabla f(x)\|^2] \le \sigma^2$, where $\sigma^2 \ge 0$. We also assume that computation and communication times are statistically independent of stochastic gradients.*

5. Goal

We want to find a stationary point of the optimization problem:

Find a (possibly random) vector $x \in \mathbb{R}^d$ such that

$$\mathbb{E}\left[\|\nabla f(x)\|^2\right] \le \varepsilon$$

6. Nearly Optimal Fragile SGD Method and Lower Bound

Let us define an auxiliary function called equilibrium time:

Definition 2 (Equilibrium Time). A mapping $t^* : \mathbb{R}_{\ge 0}\times\mathbb{R}^n_{\ge 0}\times\mathbb{R}^n_{\ge 0} \to \mathbb{R}_{\ge 0}$ with inputs s (scalar), $[h_i]_{i=1}^n$ (vector), and $[\bar\tau_i]_{i=1}^n$ (vector) is called the *equilibrium time* if it is defined as follows. Find a permutation π that sorts $\max\{\bar\tau_i, h_i\}$ as $\max\{\bar\tau_{\pi_1}, h_{\pi_1}\} \le \cdots \le \max\{\bar\tau_{\pi_n}, h_{\pi_n}\}$. Then the mapping returns the value

$$t^*(s,[h_i]_{i=1}^n,[\bar\tau_i]_{i=1}^n) \equiv \min_{k\in[n]} \left\{\max\{\bar\tau_{\pi_k}, h_{\pi_k}\}, s\left(\sum_{i=1}^{k}\frac{1}{h_{\pi_i}}\right)^{-1}\right\} \in [0,\infty].$$

Summary of the results:

Table 1: **Homogeneous Case** (1). The time complexities to get an ε-stationary point in the nonconvex setting. We assume that $\tau_{i\to j} = \tau_{j\to i}$ for all $i,j \in [n]$ in this table. Abbr.: σ^2 is defined as $\mathbb{E}_\xi[\|\nabla f(x;\xi) - \nabla f(x)\|^2] \le \sigma^2$ for all $x \in \mathbb{R}^d$, L is a smoothness constant of f, $\Delta := f(x^0) - f^*$.

Method	The Worst-Case Time Complexity Guarantees	Comment
Minibatch SGD	$\max\left\{\max_{i,j\in[n]}\tau_{i\to j}, \max_{i\in[n]} h_i\right\}\left(\frac{L\Delta}{\varepsilon} + \frac{\sigma^2 L\Delta}{n\varepsilon^2}\right)$	Suboptimal since, for instance, it "linearly"[c] depends on $\max_{i\in[n]} h_i$
SWIFT (Bornstein et al., 2023)	—[b]	Suboptimal since, for instance, it "linearly"[c] depends on $\max_{i\in[n]} h_i$
Asynchronous SGD (Even et al., 2024)	—[b]	Suboptimal, for instance, even if $\tau_{i\to j} = 0\ \forall i,j\in[n]$
Fragile SGD (Corollary 1)	$\frac{L\Delta}{\varepsilon}\min_{j\in[n]} t^*(\sigma^2/\varepsilon,[h_i]_{i=1}^n,[\tau_{i\to j}]_{i=1}^n)$[a]	Optimal up to $\log n$ factor
Lower Bound (Theorem 1)	$\frac{1}{\log n+1}\frac{L\Delta}{\varepsilon}\min_{j\in[n]} t^*(\sigma^2/\varepsilon,[h_i]_{i=1}^n,[\tau_{i\to j}]_{i=1}^n)$[a]	—

[a] The mapping t^* is defined in Definition 2.
[b] It is not trivial to infer the time complexities for these methods. However, in Section 5.4, we discuss some cases where it is transparent that the obtained results are suboptimal.
[c] Meaning that the corresponding time complexity $\to\infty$ if $\max_{i\in[n]} h_i \to \infty$.

Lower Bound

We analyze virtually all possible decentralized methods, including gossip algorithms, asynchronous and centralized methods

Protocol 1 Virtually All Decentralized Methods

1: Init $S_i = \emptyset$ (all available information) on worker i for all $i \in [n]$
2: Run the following two loops in each worker in parallel
3: **while** True **do**
4: Calculate a new point x_i^k based on S_i (takes 0 seconds)
5: Calculate a stochastic gradient $\nabla f(x_i^k;\xi)$ (or $\nabla f_i(x_i^k;\xi)$) $\xi\sim\mathcal{D}_\xi$ (takes h_i seconds)
6: Atomic add $\nabla f(x_i^k;\xi)$ (or $\nabla f_i(x_i^k;\xi)$) to S_i (atomic operation, takes 0 seconds)
7: **end while**
8: **while** True **do**
9: Send[a] any vector from $\mathbb{R}^d$ based on S_i to any worker j and go to the next step of this loop without waiting (takes $\tau_{i\to j}$ seconds to send; worker j adds this vector to S_j)
10: **end while**

(a): When we prove the lower bounds, we allow algorithms to send as many vectors as they want in parallel from worker i to worker j for all $i \ne j \in [n]$.

Theorem 1 (Informal Lower Bound). *Consider Protocol 1. It is impossible to design a method that converges faster than*

$$\Omega\left(\frac{1}{\log n+1}\frac{L\Delta}{\varepsilon}\min_{j\in[n]} t^*(\sigma^2/\varepsilon,[h_i]_{i=1}^n,[\tau_{i\to j}]_{i=1}^n)\right)$$

seconds.

The lower problem reduces to the analysis of the concentration of the time series $y^T := \min_{j\in[n]} y_j^T$ and $y_j^T := \min_{i\in[n]}\{y_i^{T-1} + h_i\eta_i^T + \tau_{i\to j}\}$, where $y_i^0 = 0$ for all $i \in [n]$, and $\{\eta_i^k\}$ are i.i.d. geometric random variables. This analysis is not trivial due to the $\min_{i\in[n]}$ operations and requires new proof techniques.

Fragile SGD

The formal description of Fragile SGD is presented in the paper. The idea is pretty simple. **All workers do three jobs in parallel: calculate stochastic gradients, receive vectors, and send vectors through spanning trees. A pivot worker aggregates all stochastic gradients in g^k and, at some moment, does $x^{k+1} = x^k - \gamma g^k$.**

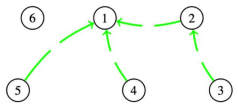


Figure 1: Node 1 is a pivot worker.

For instance, once the first *parallel process* of worker 3 finishes a calculation of a stochastic gradient, it aggregates it locally to the buffer vector and then immediately starts calculating a new stochastic gradient. *Another parallel process* of worker 3 takes the buffer vector and sends it to worker 2. Worker 2, while also calculating a stochastic gradient, receives the buffer vector and aggregates to its buffer vector, which it will send worker 1 once the communication channel is available.

7. Heterogeneous Setting : Optimal Amelie SGD Method and Lower Bound

Heterogeneous Setting: all workers store different data,

i.e., $\mathcal{D}_i$ and f_i are different

Summary of the results:

Table 2: **Heterogeneous Case** (13). Time complexities to get an ε-stationary point in the nonconvex setting. Abbr.: σ^2 is defined as $\mathbb{E}_\xi[\|\nabla f_i(x;\xi) - \nabla f_i(x)\|^2] \le \sigma^2$ for all $x \in \mathbb{R}^d, i \in [n]$, L is a smoothness constant of $f = \frac{1}{n}\sum_{i=1}^n f_i$, $\Delta := f(x^0) - f^*$.

Method	The Worst-Case Time Complexity Guarantees	Comment
Minibatch SGD	$\frac{L\Delta}{\varepsilon}\max\left\{\left(1+\frac{\sigma^2}{n\varepsilon}\right)\max\{\max_{i,j\in[n]}\tau_{i\to j}, \max_{i\in[n]} h_i\}\right\}$	suboptimal if σ^2/ε is large
RelaySGD, Gradient Tracking (Vogels et al., 2021) (Liu et al., 2024)	$\ge \frac{\max_{i\in[n]} L_i\Delta}{\varepsilon}\frac{\sigma^2}{n\varepsilon}\max_{i\in[n]} h_i$	requires local L_i-smooth. of f_i, suboptimal if σ^2/ε is large (even if $\max_{i\in[n]} L_i = L$)
Asynchronous SGD (Even et al., 2024)	—	requires similarity of the functions $\{f_i\}$, requires local L_i-smooth. of f_i
Amelie SGD and Lower Bound (Thm. 7 and Cor. 2)	$\frac{L\Delta}{\varepsilon}\max\left\{\max_{i,j\in[n]}\tau_{i\to j}, \max_{i\in[n]} h_i, \frac{\sigma^2}{n\varepsilon}\left(\frac{1}{n}\sum_{i=1}^n h_i\right)\right\}$	Optimal up to a constant factor

Amelie SGD is closely related to Fragile SGD but with essential algorithmic changes to make it work with heterogeneous functions.

8. Example in the Homogeneous Setting: Line

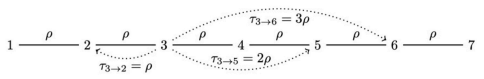


Figure 4: Line with $\rho_{i+1\to i} = \rho_{i\to i+1} = \rho$ for all $i \in [n-1]$, $\rho_{i\to j} = \infty$ otherwise. For all $i \ne j \in [n]$, edges $i \to j$ and $j \to i$ are merged and visualized with one undirected edge.

Let us consider Line graphs where we can get more explicit and interpretable formulas. Surprisingly, even in some simple cases like Line or Star graphs, as far as we know, we provide new time complexity results and insights. We can show that the optimal time complexity (up to logarithmic factors) for Line graphs in the homogeneous setting is

$$\frac{L\Delta}{\varepsilon}\left[h + \begin{cases}\sigma^2 h/\varepsilon, & \text{if } \sqrt{\sigma^2 h/\varepsilon\rho} \le 1,\\ \sqrt{\rho\sigma^2 h/\varepsilon}, & \text{if } n > \sqrt{\sigma^2 h/\varepsilon\rho} > 1,\\ \sigma^2 h/n\varepsilon, & \text{if } \sqrt{\sigma^2 h/\varepsilon\rho} \ge n\end{cases}\right]$$

seconds. There are three time complexity regimes:

- slow communication, i.e., $\sqrt{\sigma^2 h/\varepsilon\rho} \le 1$, this inequality means that ρ is so large, that communication between workers will not increase the convergence speed, and the best strategy is to work with only one worker!
- medium communication, i.e., $n > \sqrt{\sigma^2 h/\varepsilon\rho} > 1$, more than one worker will participate in the optimization process; however, *not all of them!*, some workers will not contribute since their distances $\tau_{j^*\to\cdot}$ to the pivot worker j^* are large
- fast communication, i.e., $\sqrt{\sigma^2 h/\varepsilon\rho} \ge n$, all n workers will participate in optimization because ρ is small.

9. Experiments

We focus on highlighting the results from the logistic regression experiments:

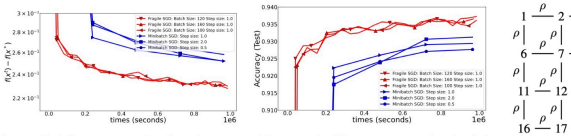
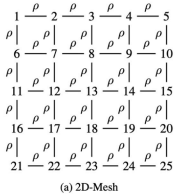


Figure 5: The communication time $\rho = 10$ seconds (Slow communication) in 2D-Mesh



(a) 2D-Mesh

On *MNIST* dataset (LeCun et al., 2010) with 100 workers, Fragile SGD is much faster and has better test accuracy than Minibatch SGD.