Better Communication Complexity for Local SGD

Ahmed Khaled [1] Konstantin Mishchenko [2] Peter Richtárik [2]

[1] Cairo University [2] KAUST

Distributed Stochastic Optimization

We consider the optimization problem

$$\min_{x \in \mathbb{R}^d} f(x) \qquad (1)$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is smooth and convex and d is large. We assume that there is a solution $x_* \in \mathbb{R}^d$ of Problem (1). Problems such as (1) routinely arise in machine learning and optimization and are solved in a distributed manner on clusters of computing nodes typically connected to a central parameter server.

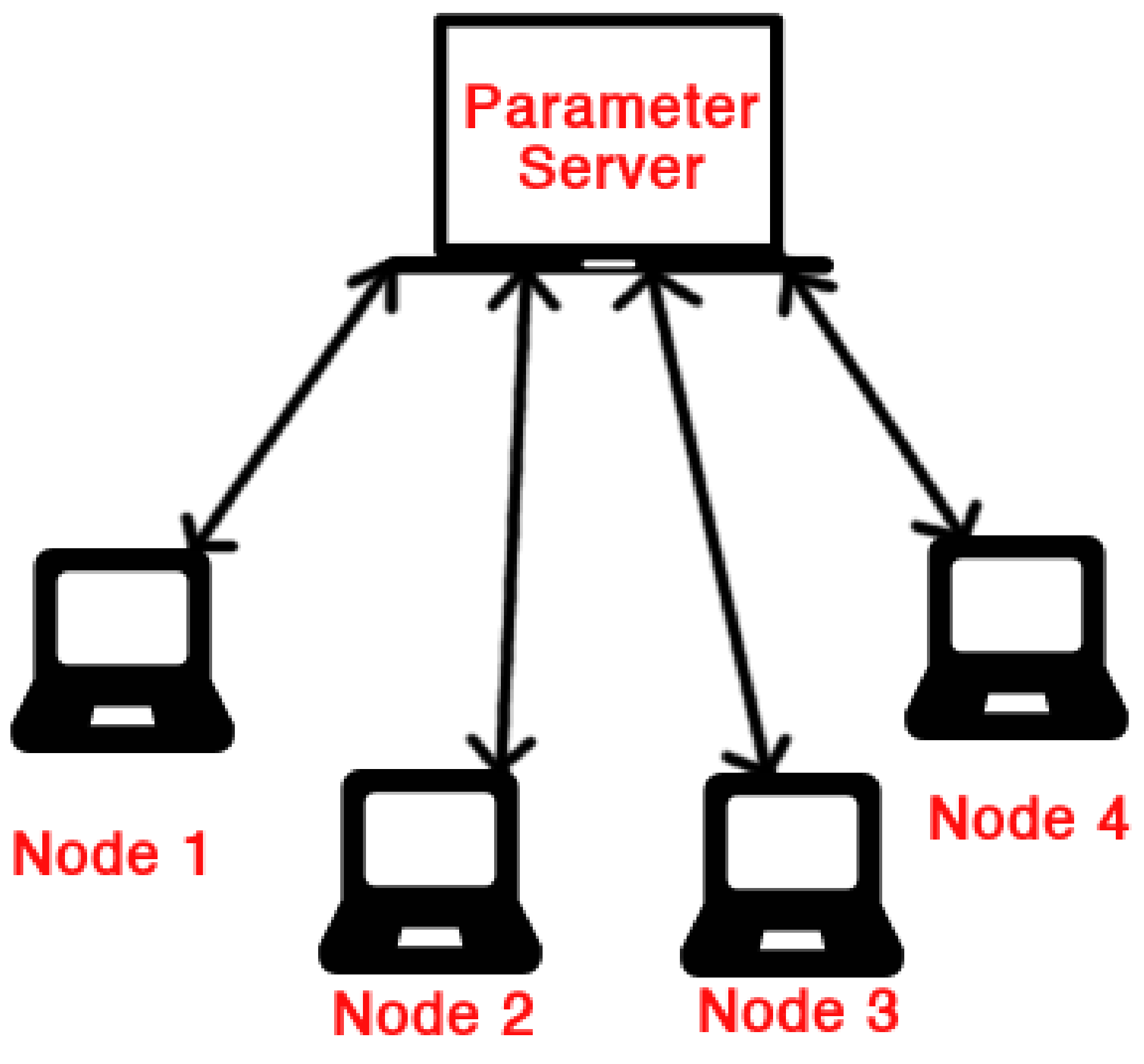

Figure 1:Parameter Server Setting

Stochastic Gradient Descent

One of the most popular methods in practice for solving (1) is **Minibatch Stochastic Gradient Descent (SGD)**. Minibatch SGD applied to problem (1) takes the form

$$x_{t+1} = x_t - \frac{\gamma_t}{M} \sum_{1 \leqslant m \leqslant M} g_t^m. \qquad (2)$$

Here $\gamma_t > 0$ is the stepsize used at time t and g_t^m is an unbiased estimator of the gradient: $\mathbb{E}\left[g_t^m\right] = \nabla f(x_t)$. The stochastic gradients g_t^m are computed in parallel by all nodes m, communicated to a parameter server, which performs (2) and communicates the result to each of the nodes, then the process is repeated until convergence.

Linear speedup

We say that a distributed algorithm shows a *linear speedup* in the number of nodes M if doubling the number of nodes leads to halving the time to convergence. The theoretical analysis of Minibatch SGD shows that it attains a *linear speedup* in the number of nodes M [1].

In Minibatch SGD, we communicate once per computed stochastic gradient. **Can we communicate less?**

Local SGD

We sample multiple gradients on each node and take *multiple SGD steps locally then average at the end.* The result is an algorithm that communicates once every H steps rather than once every step.

```
Algorithm 1 Local SGD
Input: Stepsize $\gamma > 0$, initial vector $x_0 = x_0^m$ for all
    $m \in [M]$, synchronization interval $H$.
for $t = 0, 1, \dots$ do
  for $m = 1, \dots, M$ do
    Sample local stochastic gradient $g_t^m$ such that
        $\mathbb{E}\left[g_t^m \mid x_t^m\right] = \nabla f\left(x_t^m\right).$
    if $t + 1$ is a multiple of H then
      Communicate local nodes to parameter server,
      average them and communicate them back to
      each node
          $x_{t+1}^m = \frac{1}{m} \sum_{j=1}^{M} (x_t^j - \gamma g_t^j).$
    else
      Take one step of SGD locally on each node
          $x_{t+1}^m = x_t^m - \gamma g_t^m.$
    end if
  end for
end for
```

Assumptions

Assumption 1: f is L-smooth and μ-strongly convex (we allow $\mu = 0$). That is, for all $x, y \in \mathbb{R}^d$ we have:

$$f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2 \leqslant f(x)$$
$$f(x) \leqslant f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2.$$

Assumption 2: The stochastic gradients $(g_t^m)_{t \geqslant 0, m \in [M]}$ are unbiased estimates of the true gradient with uniformly bounded variance

$$\mathbb{E}\left[g_t^m\right] = \nabla f\left(x_t^m\right) \text{ and}$$
$$\mathbb{E}\left[\left\|g_t^m - \nabla f\left(x_t^m\right)\right\|^2\right] \leqslant \sigma^2 \text{ for all } t \geqslant 0 \text{ and } m \in [M].$$

Convergence under $\mu > 0$

Let $\kappa \overset{\text{def}}{=} L/\mu \geqslant 1$. By properly choosing stepsizes γ_t we can obtain for the average of the local iterates $\hat{x}_t$ that $\mathbb{E}\left[\|\hat{x}_t - x_*\|^2\right] \leqslant \varepsilon$ when the total number of iterates T and the total number of communication rounds $C \overset{\text{def}}{=} T/H$ are:

$$T = \tilde{\Omega}\left(\frac{\sigma^2}{\varepsilon M}\right) \text{ and } C = \Omega(\kappa M), \qquad (3)$$

where $\tilde{\Omega}(\cdot)$ indicates possibly ignoring polylogarithmic factors. Clearly the analysis shows that there is a linear speedup in the number of nodes M.

Constant number of communications When the number of nodes M is fixed, we only need a *constant* number of communication rounds regardless of the total number of local steps T. This tightens the previous analysis [2], where $C = \Omega\left(\kappa\sqrt{T/M}\right)$ was required.

Convergence under $\mu = 0$

For $\bar{x}_T = \frac{1}{MT} \sum_{t=1}^{T} \sum_{m=1}^{M} x_t^m$ we have that $f(x) - f(x_*) \leqslant \varepsilon$ provided that

$$T = \Omega\left(\frac{\sigma^4}{M\varepsilon^2}\right) \text{ and } C = \Omega\left(\sqrt{TM^3}\right).$$

This result is new: the setting with $\mu = 0$ was not considered explicitly in prior work. There is clearly a **linear speedup** in the number of nodes M: the total number of iterations needed halves when M doubles, but we have to pay the price of communicating more often.

Experimental Results

We run experiments on ℓ_2 regularized logistic regression problem with $M = 20$ nodes, each with Intel(R) Xeon(R) Gold 6146 CPU @3.20GHz core. We set ℓ_2 penalty to be $\frac{1}{n}$, where n is the dataset size.

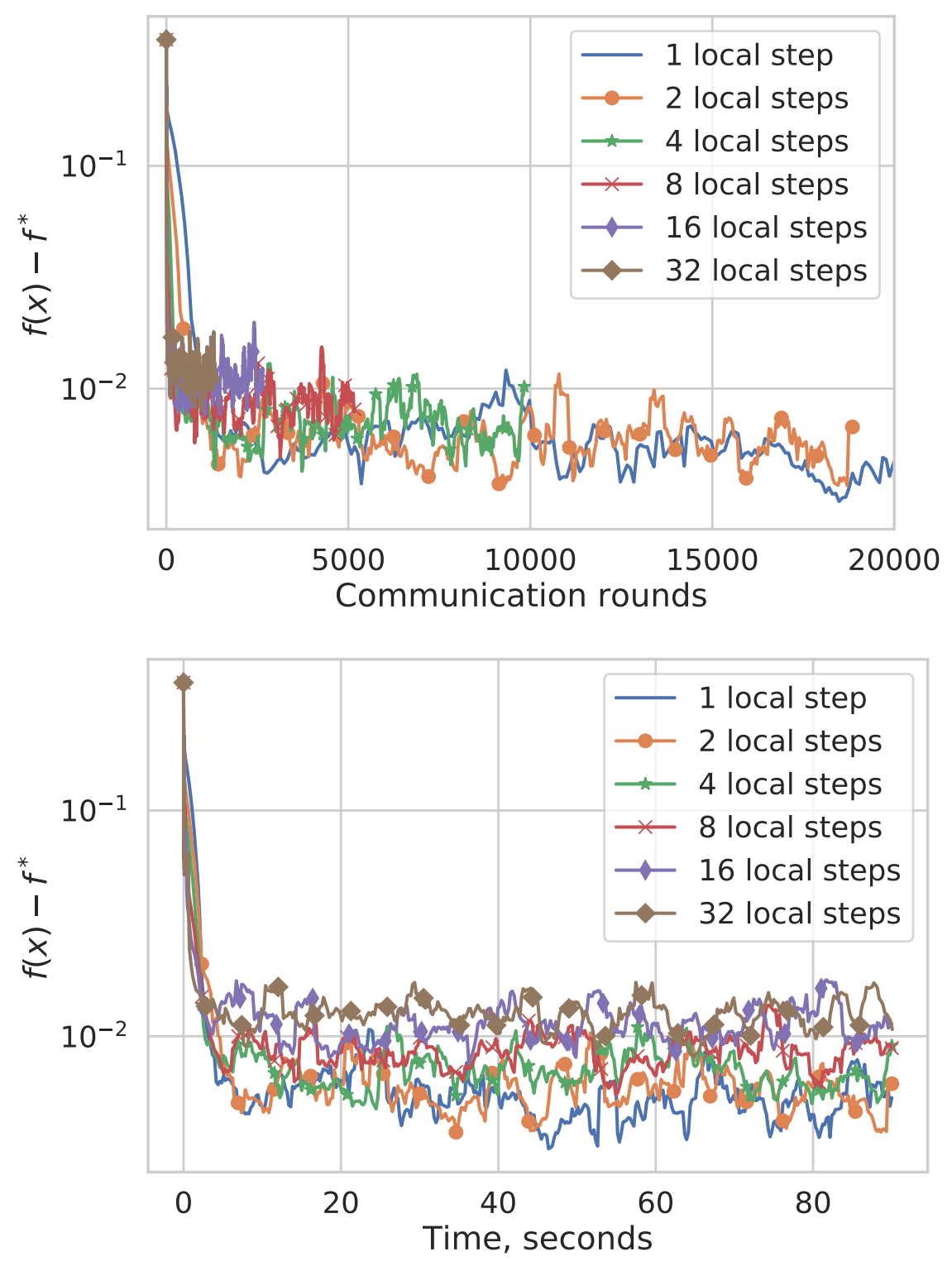

Figure 2:All local iterates converge to a neighborhood within a small number of communication rounds due to large stepsizes.

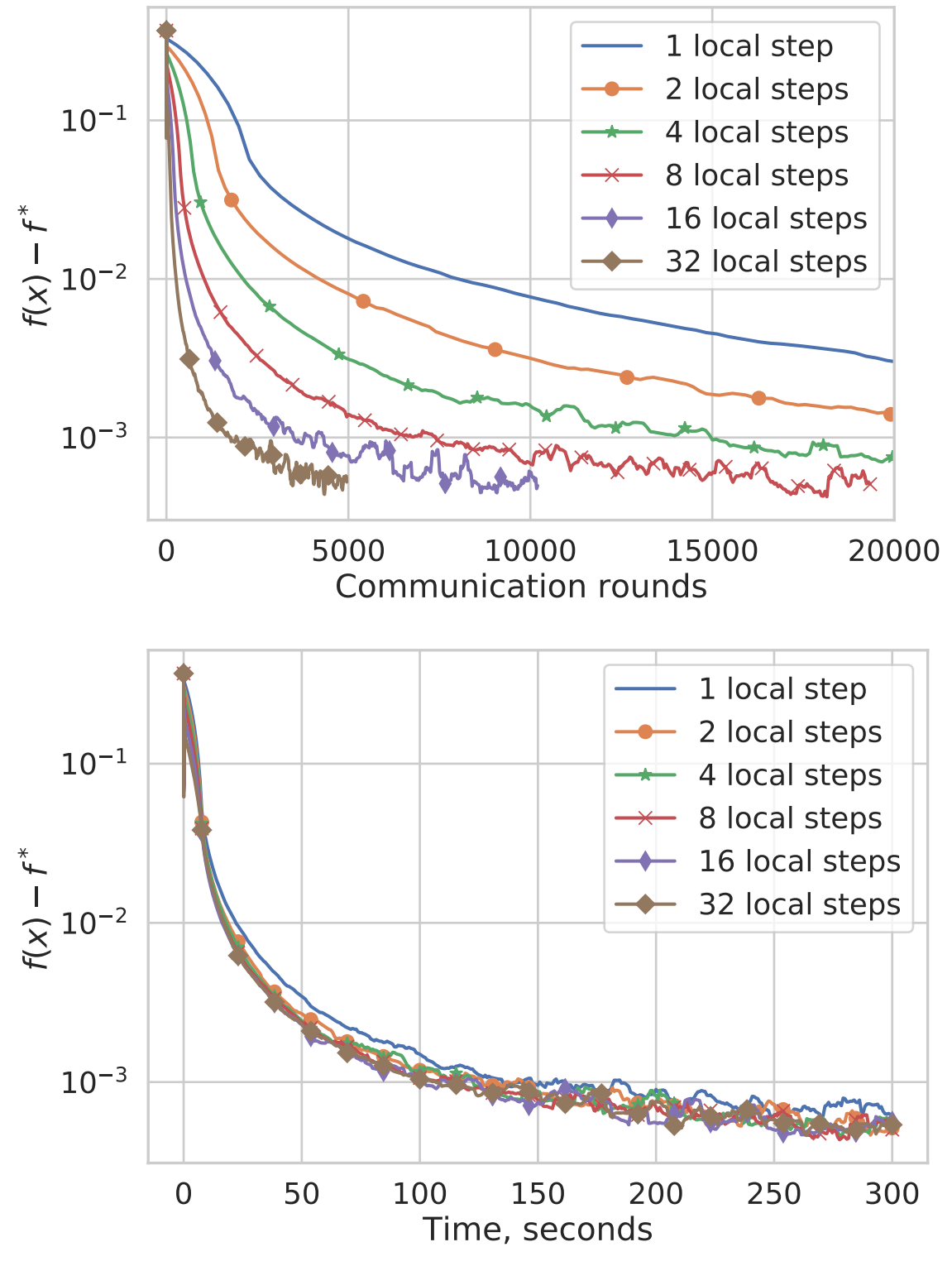

Figure 3:With more local iterations, fewer communication rounds are required to get to a neighborhood of the solution.

References

[1] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao.
Optimal Distributed Online Prediction using Mini-Batches.
arXiv:1012.1367, 2010.

[2] Sebastian U. Stich.
Local SGD converges fast and communicates little.
In *International Conference on Learning Representations*, 2019.

Acknowledgements

Icons used include Computer icon by WPZOOM under Creative Commons (Attribution-Share Alike 3.0 Unported) and the Computer icon by Danilo Demarco.