Better Communication Complexity for Local SGD

Ahmed Khaled*
Cairo University
akregeb@gmail.com

Konstantin Mishchenko
KAUST†
konstantin.mishchenko@kaust.edu.sa

Peter Richtárik
KAUST
peter.richtarik@kaust.edu.sa

Abstract

We revisit the local Stochastic Gradient Descent (local SGD) method and prove new convergence rates. We close the gap in the theory by showing that it works under unbounded gradients and extend its convergence to weakly convex functions. Furthermore, by changing the assumptions, we manage to get new bounds that explain in what regimes local SGD is faster that its non-local version. For instance, if the objective is strongly convex, we show that, up to constants, it is sufficient to synchronize M times in total, where M is the number of nodes. This improves upon the known requirement of Stich (2018) of $\sqrt{TM}$ synchronization times in total, where T is the total number of iterations, which helps to explain the empirical success of local SGD.

1 Introduction

Big data optimization problems arising in machine learning and statistics, such as the training of supervised learning models, are routinely solved in a distributed manner on a cluster of compute nodes [2]. Distributed optimization algorithms are typically iterative methods alternating local computations performed on the nodes, and expensive communication steps involving all or a subset of the nodes. Due to the need to solve such problems more efficiently, there has been a lot of recent interest in understanding the trade-offs between communication and computation, a concern which is particularly important in the federated learning setting; see [3, 10, 13].

Minibatch SGD. A popular method for solving unconstrained smooth optimization problems of the form

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

in situations when the computation of the gradient of f is expensive is minibatch SGD [7, 8]:

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

Here $\gamma_t > 0$ is the stepsize used at time t and g_t^m is an unbiased estimator of the gradient: $\mathbb{E}\left[g_t^m\right] = \nabla f(x_t)$. In a typical parameter server setup, stochastic gradients g_t^m are computed in parallel by all (or a subset of) nodes m, communicated to a parameter server, which subsequently performs the update (2) and communicates it to the nodes, and the process is repeated. As M grows, the variance

---

*Work done during an internship at KAUST.

†King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

---

Algorithm 1 Local SGD

Input: Stepsize $\gamma > 0$, initial vector $x_0 = x_0^m$ for all $m \in [M]$.
1: **for** $t = 0, 1, \ldots$ **do**
2:     **for** $m = 1, \ldots, M$ **do**
3:         Sample g_t^m such that $\mathbb{E}\left[g_t^m \mid x_t^m\right] = \nabla f(x_t^m)$.
4:         $x_{t+1}^m = \begin{cases} \frac{1}{m} \sum_{j=1}^M (x_t^j - \gamma g_t^j), & \text{if } t = t_p \text{ for some } p \in \mathbb{N} \\ x_t^m - \gamma g_t^m, & \text{otherwise.} \end{cases}$
5:     **end for**
6: **end for**

$\frac{1}{M} \sum_{m=1}^M g_t^m$ as an estimator of the gradient $\nabla f(x_t)$ decreases, which leads a decrease in the overall number of communications needed to obtain a solution of sufficient quality.

Local SGD. Note that (2) can equivalently be written in the form

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

which leads to the alternative interpretation of minibatch SGD as averaging the results of a single SGD step performed by all nodes, initiated from the same starting point x_t. This simple observation immediately leads to the natural question: can we gain by performing more than a single step of SGD on each node before averaging? By performing what we hope will be useful additional computation locally on the nodes before expensive aggregation is done, we hope to decrease the number of communication rounds needed. We have just described the local SGD method, formalized as Algorithm 1.

2 Contributions

While local SGD has been popular among practitioners for a long time [5, 12], its theoretical understanding has remained elusive until very recently [1, 9, 14, 15, 16, 17] (see Table 1). The history of the methods goes back to the convergence proof in the early work [11], but a tight convergence rate has been missing since then. Although most existing works focus on analyzing local SGD for smooth and nonconvex f, there are no analyses specialized to the smooth convex case, and only two papers which provide bounds in the smooth strongly convex case.

In this paper we obtain the first result explicitly covering the convex case, and improve dramatically upon the best known communication complexity result in the strongly convex case (see the last row of Table 1). Moreover, unlike previous results in the strongly convex case that depend on a restrictive gradient boundedness assumption, our results do not have this flaw.

Unbounded gradient	$C(T)$ f strongly convex	$C(T)$ f weakly convex	$C(T)$ f nonconvex	Reference
✓	✗	✗	$\Omega(T)$	Zhou & Cong, 8/2017, [17]
✗	$\Omega(\sqrt{MT})$	✗	✗	Stich, 5/2018, [14]
✗	✗	✗	$\Omega\left(M^{3/4}T^{3/4}\right)$	Yu et al, 7/2018, [16]
✓	✗	✗	$\Omega(M^{3/2}T^{1/2})$	Wang & Joshi, 8/2018, [15]
✓	✗	✗	$\Omega(T)$	Jiang & Agarwal, 12/2018, [9]
✗	$\Omega\left(\sqrt{MT}\right)$	✗	$\Omega\left(M^{3/4}T^{3/4}\right)$	Basu et al, 6/2019, [1]
✓	$\tilde{\Omega}(M)$	$\Omega(M^{3/2}T^{1/2})$	✗	**THIS WORK, 9/2019**

Table 1: Existing theoretical bounds for local SGD. $C(T)$ denotes the minimum number of communication steps required each T iterations to achieve a linear speedup in the number of nodes M.

An overview of related work on local stochastic gradient methods is given in Table 1.

2.1 Setting and Contributions

In this work we consider minimization problem (1) under the following assumptions:

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

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

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 when conditioned on x_t^m:

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

Note that Assumption 2 is less restrictive than the bounded gradients assumption ($\mathbb{E}\left[\|g_t^m\|^2\right] \leqslant G^2$) used in several previous analysis as shown in Table 1. Under this setting, the main contributions of this paper are:

1. If f is strongly convex, then by properly choosing stepsizes γ_t and taking the average of the local iterates $\hat{x}_t$, we can obtain $\mathbb{E}\left[\|\hat{x}_t - x_*\|^2\right] \leqslant \varepsilon$ when the total number of iterates T and the total number of communication rounds C are:

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

   where $\tilde{\Omega}(\cdot)$ indicates possibly ignoring polylogarithmic factors. This tightens the previous analysis [14], where $C = \Omega\left(\sqrt{T/M}\right)$ was required.

2. Furthermore, if f is (possibly weakly) convex, then we can guarantee $\mathbb{E}\left[f\left(\frac{1}{T}\sum_{i=0}^{T-1}\hat{x}_i\right) - f(x_*)\right] \leqslant \varepsilon$ provided that:

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

3. We support our analysis by experiments illustrating the behavior of the algorithm.

3 Convergence Theory

We denote the sequence of time stamps when synchronization happens as $(t_p)_{p=1}^\infty$. The average of all local iterates is $\hat{x}_t = \frac{1}{M}\sum_{m=1}^M x_t^m$ and that of gradients is $g_t = \frac{1}{M}\sum_{m=1}^M g_t^m$. We define the set $[M] \overset{\text{def}}{=} \{1, 2, \ldots, M\}$.

Lemma 1. Choose a stepsize $\gamma > 0$ such that $\gamma \leqslant \frac{1}{2L}$. Under Assumptions 1, and 2 we have that for Algorithm 1 with $\max_p |t_p - t_{p+1}| \leqslant H$,

$$\mathbb{E}[V_t] \leqslant H\gamma^2\sigma^2, \tag{5}$$

where $V_t \overset{\text{def}}{=} \frac{1}{M}\sum_{m=1}^M \|x_t^m - \hat{x}_t\|^2$.

Theorem 1. Suppose that Assumptions 1, and 2 hold with $\mu > 0$. Then combining Lemma 1 with techniques from [14] we can conclude that for a constant stepsize $\gamma > 0$ such that $\gamma \leqslant \frac{1}{4L}$ we have for Algorithm 1 with $\max_p |t_p - t_{p+1}| \leqslant H$,

$$\mathbb{E}\left[\|\hat{x}_T - x_*\|^2\right] \leqslant (1 - \gamma\mu)^T \|\hat{x}_0 - x_*\|^2 + \frac{\gamma\sigma^2}{\mu M} + \frac{2L\gamma^2 H\sigma^2}{\mu}, \tag{6}$$

where $\hat{x}_T = \frac{1}{M}\sum_{m=1}^M x_T^m$.

Corollary 1. Choosing $\gamma = \frac{1}{\mu a}$, with $a = 4\kappa + t$ for $t > 0$ and we take $T = 2a \log a$ steps. Then substituting in (6) and using that $1 - x \leqslant \exp(-x)$ and some algebraic manipulation we can conclude that,

$$\mathbb{E}\left[\|\hat{x}_T - x_*\|^2\right] \leqslant \frac{4\log^2(a)}{T^2}\|\hat{x}_0 - x_*\|^2 + \frac{2\sigma^2 \log(a)}{\mu^2 MT} + \frac{8L\sigma^2 H \log^2(a)}{\mu^3 T^2} \tag{7}$$

$$= \tilde{O}\left(\frac{\|\hat{x}_0 - x_*\|^2}{T^2} + \frac{\sigma^2}{\mu^2 MT} + \frac{\kappa \sigma^2 H}{\mu^2 T^2}\right). \tag{8}$$

where $\tilde{O}(\cdot)$ ignores logarithmic factors. We see that choosing $H = O(T/M)$ recovers the same $1/(MT)$ convergence rate of minibatch SGD up to polylogarithmic factors, and the number of communications is then $C = T/H = \tilde{\Omega}(M)$.

Using similar proof techniques, we can show the following result for weakly convex functions:

Theorem 2. Suppose that Assumptions 1, and 2 hold with $\mu = 0$ and that a constant stepsize γ such that $\gamma \geqslant 0$ and $\gamma \leqslant \frac{1}{4L}$ is chosen and that Algorithm 1 is run with up to H local iterations, for $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T}\hat{x}_T$,

$$\mathbb{E}\left[f(\bar{x}_T) - f(x_*)\right] \leqslant \frac{2}{\gamma T}\|x_0 - x_*\|^2 + \frac{2\gamma\sigma^2}{M} + 4\gamma^2 L^2 \sigma^2 H. \tag{9}$$

Corollary 2. Assume that $T \geqslant M$. Choosing $\gamma = \frac{\sqrt{M}}{4L\sqrt{T}}$, then substituting in (9) we have,

$$\mathbb{E}\left[f(\bar{x}_T) - f(x_*)\right] \leqslant \frac{8\|x_0 - x_*\|}{\sqrt{MT}} + \frac{\sigma^2}{2L\sqrt{MT}} + \frac{\sigma^2 MH}{T}. \tag{10}$$

We see that choosing $H = O(\sqrt{T}/M^{3/2})$ we recover the same $1/\sqrt{MT}$ convergence rate of mini-batch SGD, and the number of communication steps is then $C = T/H = \Omega\left(M^{3/2}T^{1/2}\right)$.

4 Experiments

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 use the 'a9a' dataset from the LIBSVM library [4] and set ℓ_2 penalty to be $\frac{1}{n}$, where n is the dataset size. The code was written in Python using MPI [6].

We ran two experiments, with stepsizes $\frac{1}{L}$ and $\frac{0.05}{L}$ and minibatch size equal 1. In both cases we observe convergence to a neighborhood, although of a different radius. Since we run the experiments on a single machine, the communication is very cheap and there is little gain in time required for convergence. However, the advantage in terms of required communication rounds is self-evident and can lead to significant time improvement under slow communication networks.
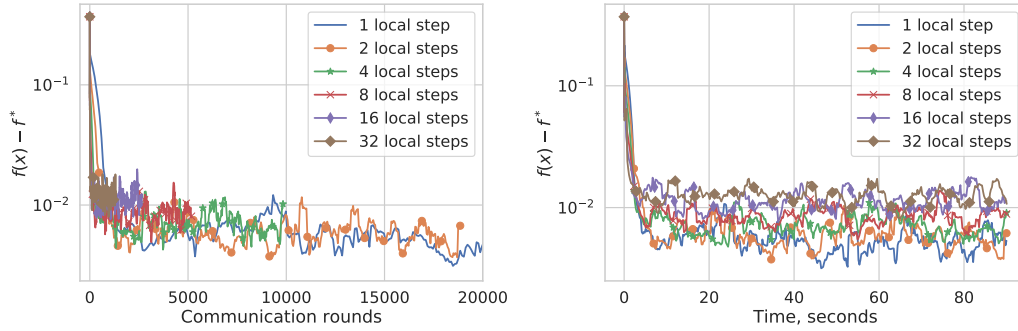
4

Figure 1: Results on regularized logistic regression, 'a9a' dataset, with stepsize $\frac{1}{L}$. All local numbers of local iterations converge to a neighborhood within a small number of communication rounds due to large stepsizes.
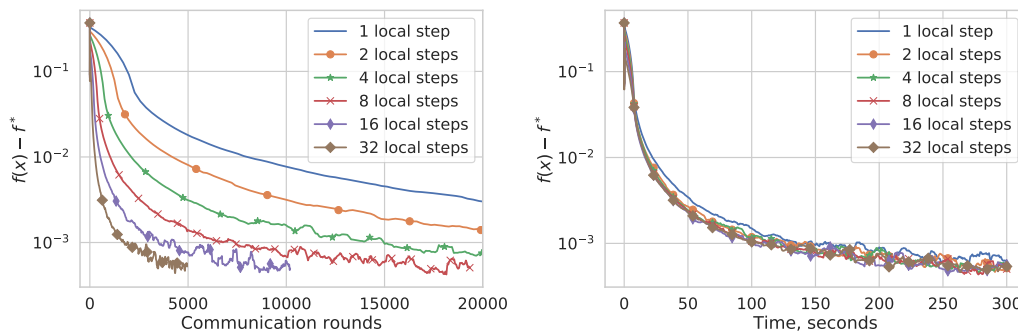

Figure 2: Results on regularized logistic regression, 'a9a' dataset, with stepsize $\frac{0.05}{L}$. With more local iterations, fewer communication rounds are required to get to a neighborhood of the solution.

5

References

[1] Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-SGD: Distributed SGD with Quantization, Sparsification, and Local Computations. *arXiv:1906.02367*, 2019.

[2] Tal Ben-Nun and Torsten Hoefler. Demystifying Parallel and Distributed Deep Learning: An In-Depth Concurrency Analysis. *arXiv:1802.09941*, 2018.

[3] Sebastian Caldas, Jakub Konečný, H. Brendan McMahan, and Ameet Talwalkar. Expanding the Reach of Federated Learning by Reducing Client Resource Requirements. *arXiv:1812.07210*, 2018.

[4] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[5] Gregory F. Coppola. *Iterative parameter mixing for distributed large-margin training of structured predictors for natural language processing*. PhD thesis, University of Edinburgh, UK, 2015.

[6] Lisandro D. Dalcin, Rodrigo R. Paz, Pablo A. Kler, and Alejandro Cosimo. Parallel distributed computing using Python. *Advances in Water Resources*, 34(9):1124–1139, 2011.

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

[8] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General Analysis and Improved Rates. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[9] Peng Jiang and Gagan Agrawal. A Linear Speedup Analysis of Distributed Deep Learning with Sparse and Quantized Communication. *Advances in Neural Information Processing Systems 31*, (NeurIPS):2530–2541, 2018.

[10] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated Learning: Strategies for Improving Communication Efficiency. In *NIPS Private Multi-Party Machine Learning Workshop*, 2016.

[11] L. Mangasarian. Parallel Gradient Distribution in Unconstrained Optimization. *SIAM Journal on Control and Optimization*, 33(6):1916–1925, 1995.

[12] Ryan McDonald, Keith Hall, and Gideon Mann. Distributed Training Strategies for the Structured Perceptron. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 456–464, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5.

[13] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the 20 th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017. JMLR: W&CP volume 54*, 2016.

[14] Sebastian U. Stich. Local SGD Converges Fast and Communicates Little. *arXiv:1805.09767*, 2018.

[15] Jianyu Wang and Gauri Joshi. Cooperative SGD: A Unified Framework for the Design and Analysis of Communication-Efficient SGD Algorithms. *arXiv:1808.07576*, 2018.

[16] Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel Restarted SGD with Faster Convergence and Less Communication: Demystifying Why Model Averaging Works for Deep Learning. *arXiv:1807.06629*, 2018.

[17] Fan Zhou and Guojing Cong. On the Convergence Properties of a K-step Averaging Stochastic Gradient Descent Algorithm for Nonconvex Optimization. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2018-July, pages 3219–3227, 2018.

Supplementary Material

A Basic Facts and Notation

We denote the sequence of time stamps when synchronization happens as $(t_p)_{p=1}^\infty$. Given stochastic gradients $g_t^1, g_t^2, \ldots, g_t^M$ at time $t \geqslant 0$ we define

$$g_t \overset{\text{def}}{=} \frac{1}{M} \sum_{m=1}^M g_t^m, \qquad \bar{g}_t^m \overset{\text{def}}{=} \mathbb{E}\left[g_t^m\right] = \nabla f(x_t^m), \qquad \bar{g}_t \overset{\text{def}}{=} \mathbb{E}\left[g_t\right].$$

Throughout the proofs, we will use the variance decomposition that holds for any random vector X with finite second moment:

$$\mathbb{E}\left[\|X\|^2\right] = \mathbb{E}\left[\|X - \mathbb{E}\left[X\right]\|^2\right] + \|\mathbb{E}\left[X\right]\|^2.$$

In particular, its version for vectors with finite number of values gives

$$\frac{1}{M} \sum_{m=1}^M \|X_m\|^2 = \frac{1}{M} \sum_{m=1}^M \left\| X_m - \frac{1}{M} \sum_{i=1}^M X_i \right\|^2 + \left\| \frac{1}{M} \sum_{m=1}^M X_m \right\|^2. \tag{11}$$

As a consequence of (11) we have that,

$$\mathbb{E}\left[\|X - \mathbb{E}\left[X\right]\|^2\right] \leqslant \mathbb{E}\left[\|X\|^2\right].$$

Proposition 1 (Jensen's inequality). For any convex function f and any vectors $x^1, \ldots, x^M$ we have

$$f\left(\frac{1}{M} \sum_{m=1}^M x^m\right) \leqslant \frac{1}{M} \sum_{m=1}^M f(x^m).$$

We denote the Bregman divergence associated with function f and arbitrary x, y as

$$D_f(x, y) \overset{\text{def}}{=} f(x) - f(y) - \langle \nabla f(y), x - y \rangle.$$

Proposition 2. If f is L-smooth and convex, then for any x and y it holds

$$\|\nabla f(x) - \nabla f(y)\|^2 \leqslant 2L D_f(x, y). \tag{12}$$

If f satisfies Assumption 1, then

$$f(x) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|y - x\|^2 \leqslant f(y) \qquad \forall x, y \in \mathbb{R}^d. \tag{13}$$

B Proof of Lemma 1

Proof. Let $t \in \mathbb{N}$ be such that $t_p \leqslant t < t_{p+1}$. Recall that for a time t such that $t_p \leqslant t < t_{p+1}$ we have $x_{t+1}^m = x_t^m - \gamma g_t^m$ and $\hat{x}_{t+1} = \hat{x}_t - \gamma g_t$. Hence for the expectation conditional on $x_t^1, x_t^2, \ldots, x_t^M$ we have:

$$\begin{aligned}
\mathbb{E}\left[\|x_{t+1}^m - \hat{x}_{t+1}\|^2\right] &= \|x_t^m - \hat{x}_t\|^2 + \gamma^2 \mathbb{E}\left[\|g_t^m - g_t\|^2\right] - 2\gamma \mathbb{E}\left[\langle x_t^m - \hat{x}_t, g_t^m - g_t \rangle\right] \\
&= \|x_t^m - \hat{x}_t\|^2 + \gamma^2 \mathbb{E}\left[\|g_t^m - g_t\|^2\right] - 2\gamma \langle x_t^m - \hat{x}_t, \nabla f(x_t^m) \rangle \\
&\quad + 2\gamma \langle x_t^m - \hat{x}_t, \bar{g}_t \rangle.
\end{aligned}$$

Averaging both sides and letting $V_t = \frac{1}{M} \sum_m \|x_t^m - \hat{x}_t\|^2$, we have

$$\begin{aligned}
\mathbb{E}\left[V_{t+1}\right] &= V_t + \frac{\gamma^2}{M} \sum_m \mathbb{E}\left[\|g_t^m - g_t\|^2\right] - \frac{2\gamma}{M} \sum_m \langle x_t^m - \hat{x}_t, \nabla f(x_t^m) \rangle + 2\gamma \underbrace{\langle \hat{x}_t - \hat{x}_t, \bar{g}_t \rangle}_{=0} \\
&= V_t + \frac{\gamma^2}{M} \sum_m \mathbb{E}\left[\|g_t^m - g_t\|^2\right] - \frac{2\gamma}{M} \sum_m \langle x_t^m - \hat{x}_t, \nabla f(x_t^m) \rangle. \tag{14}
\end{aligned}$$

Now note that by expanding the square we have,

$$\mathbb{E}\left[\|g_t^m - g_t\|^2\right] = \mathbb{E}\left[\|g_t^m - \bar{g}_t\|^2\right] + \mathbb{E}\left[\|\bar{g}_t - g_t\|^2\right] + 2\mathbb{E}\left[\langle g_t^m - \bar{g}_t, \bar{g}_t - g_t\rangle\right]. \tag{15}$$

We decompose the first term in the last equality again by expanding the square,

$$\mathbb{E}\left[\|g_t^m - \bar{g}_t\|^2\right] = \mathbb{E}\left[\|g_t^m - \bar{g}_t^m\|^2\right] + \|\bar{g}_t^m - \bar{g}_t\|^2 + 2\mathbb{E}\left[\langle g_t^m - \bar{g}_t^m, \bar{g}_t^m - \bar{g}_t\rangle\right]$$

$$= \mathbb{E}\left[\|g_t^m - \bar{g}_t^m\|^2\right] + \|\bar{g}_t^m - \bar{g}_t\|^2 + 2\underbrace{\langle \bar{g}_t^m - \bar{g}_t^m, \bar{g}_t^m - \bar{g}_t\rangle}_{=0}$$

$$= \mathbb{E}\left[\|g_t^m - \bar{g}_t^m\|^2\right] + \|\bar{g}_t^m - \bar{g}_t\|^2.$$

Plugging this into (15) we have,

$$\mathbb{E}\left[\|g_t^m - g_t\|^2\right] = \mathbb{E}\left[\|g_t^m - \bar{g}_t^m\|^2\right] + \|\bar{g}_t^m - \bar{g}_t\|^2 + \mathbb{E}\left[\|\bar{g}_t - g_t\|^2\right] + 2\mathbb{E}\left[\langle g_t^m - \bar{g}_t, \bar{g}_t - g_t\rangle\right].$$

Now average over m:

$$\frac{1}{M}\sum_m \mathbb{E}\left[\|g_t^m - g_t\|^2\right] = \frac{1}{M}\sum_m \mathbb{E}\left[\|g_t^m - \bar{g}_t^m\|^2\right] + \frac{1}{M}\sum_m \|\bar{g}_t^m - \bar{g}_t\|^2 + \mathbb{E}\left[\|\bar{g}_t - g_t\|^2\right]$$

$$- 2\mathbb{E}\left[\|\bar{g}_t - g_t\|^2\right],$$

where we used that by definition $\frac{1}{M}\sum_{m=1}^M g_t^m = g_t$. Hence,

$$\frac{1}{M}\sum_m \mathbb{E}\left[\|g_t^m - g_t\|^2\right] = \frac{1}{M}\sum_m \mathbb{E}\left[\|g_t^m - \bar{g}_t^m\|^2\right] + \frac{1}{M}\sum_m \|\bar{g}_t^m - \bar{g}_t\|^2 - \mathbb{E}\left[\|\bar{g}_t - g_t\|^2\right]$$

$$\leqslant \frac{1}{M}\sum_m \mathbb{E}\left[\|g_t^m - \bar{g}_t^m\|^2\right] + \frac{1}{M}\sum_m \|\bar{g}_t^m - \bar{g}_t\|^2. \tag{16}$$

Now note that for the first term in (16) we have by Assumption 2,

$$\mathbb{E}\left[\|g_t^m - \bar{g}_t^m\|^2\right] = \mathbb{E}\left[\|g_t^m - \nabla f(x_t^m)\|^2\right] \leqslant \sigma^2. \tag{17}$$

For the second term in (16) we have

$$\|\bar{g}_t^m - \bar{g}_t\|^2 = \|\bar{g}_t^m - \nabla f(\hat{x}_t)\|^2 + \|\nabla f(\hat{x}_t) - \bar{g}_t\|^2 + 2\langle \bar{g}_t^m - \nabla f(\hat{x}_t), \nabla f(\hat{x}_t) - \bar{g}_t\rangle.$$

Averaging over m,

$$\frac{1}{M}\sum_{m=1}^M \|\bar{g}_t^m - \bar{g}_t\|^2 = \frac{1}{M}\sum_m \|\bar{g}_t^m - \nabla f(\hat{x}_t)\|^2 + \|\nabla f(\hat{x}_t) - \bar{g}_t\|^2 + 2\langle \bar{g}_t - \nabla f(\hat{x}_t), \nabla f(\hat{x}_t) - \bar{g}_t\rangle$$

$$= \frac{1}{M}\sum_m \|\bar{g}_t^m - \nabla f(\hat{x}_t)\|^2 + \|\nabla f(\hat{x}_t) - \bar{g}_t\|^2 - 2\|\nabla f(\hat{x}_t) - \bar{g}_t\|^2$$

$$= \frac{1}{M}\sum_m \|\bar{g}_t^m - \nabla f(\hat{x}_t)\|^2 - \|\nabla f(\hat{x}_t) - \bar{g}_t\|^2 \leqslant \frac{1}{M}\sum_m \|\bar{g}_t^m - \nabla f(\hat{x}_t)\|^2,$$

where we used the fact that $\frac{1}{M}\sum_m \bar{g}_t^m = \bar{g}_t$, which comes from the linearity of expectation. Now we bound $\|\bar{g}_t^m - \nabla f(\hat{x}_t)\|^2$ in the last inequality by smoothness and then use that Jensen's inequality implies $\sum_{m=1}^M f(\hat{x}_t) - f(x_t^m) \leqslant 0$,

$$\frac{1}{M}\sum_m \|\bar{g}_t^m - \nabla f(\hat{x}_t)\|^2 = \frac{1}{M}\sum_m \|\nabla f(x_t^m) - \nabla f(\hat{x}_t)\|^2$$

$$\overset{(12)}{\leqslant} \frac{1}{M}\sum_m 2L(f(\hat{x}_t) - f(x_t^m) - \langle \hat{x}_t - x_t^m, \nabla f(x_t^m)\rangle)$$

$$\leqslant -\frac{2L}{M}\sum_m \langle x_t^m - \hat{x}_t, \nabla f(x_t^m)\rangle. \tag{18}$$

Plugging in (18) and (17) into (16) we have,

$$\frac{1}{M}\sum_m \mathbb{E}\left[\|g_t^m - g_t\|^2\right] \leqslant \sigma^2 - \frac{2L}{M}\sum_m \langle x_t^m - \hat{x}_t, \nabla f(x_t^m)\rangle. \tag{19}$$

Plugging (19) into (14), we get

$$\mathbb{E}\left[V_{t+1}\right] \leqslant V_t + \gamma^2\sigma^2 - \frac{2\gamma}{M}\sum_m \langle x_t^m - \hat{x}_t, \nabla f(x_t^m)\rangle + \frac{2L\gamma^2}{M}\sum_m \langle x_t^m - \hat{x}_t, \nabla f(x_t^m)\rangle$$

$$= V_t + \gamma^2\sigma^2 - \frac{2\gamma(1-\gamma L)}{M}\sum_m \langle x_t^m - \hat{x}_t, \nabla f(x_t^m)\rangle \tag{20}$$

$$\overset{(13)}{\leqslant} (1 - \gamma(1-\gamma L)\mu)\, V_t + \gamma^2\sigma^2.$$

Using that $\gamma \leqslant \frac{1}{2L}$ we can conclude,

$$\mathbb{E}\left[V_{t+1}\right] \leqslant \left(1 - \frac{\gamma\mu}{2}\right)V_t + \gamma^2\sigma^2$$

$$\leqslant V_t + \gamma^2\sigma^2.$$

Taking expectations and iterating the above inequality,

$$\mathbb{E}\left[V_t\right] \leqslant \mathbb{E}\left[V_{t_p}\right] + \gamma^2\sigma^2\,(t - t_p)$$

$$\leqslant \mathbb{E}\left[V_{t_p}\right] + \gamma^2\sigma^2 H.$$

It remains to notice that by assumption we have $V_{t_p} = 0$. ∎

C Two More Lemmas

Lemma 2. [14]. Let $(x_t^m)_{t\geqslant 0}$ be iterates generated by Algorithm 1. Suppose that f satisfies Assumption 1 and that $\gamma \leqslant \frac{1}{2L}$. Then,

$$\mathbb{E}\left[\|\hat{x}_{t+1} - x_*\|^2\right] \leqslant (1-\gamma\mu)\mathbb{E}\left[\|\hat{x}_t - x_*\|^2\right] + \gamma^2\mathbb{E}\left[\|g_t - \bar{g}_t\|^2\right]$$

$$- \frac{\gamma}{2}D_f(\hat{x}_t, x_*) + 2\gamma L V_t. \tag{21}$$

Proof. This is Lemma 3.1 in [14]. ∎

Lemma 3. Suppose that Assumption 2 holds. Then,

$$\mathbb{E}\left[\|g_t - \bar{g}_t\|^2\right] \leqslant \frac{\sigma^2}{M}.$$

Proof. This is Lemma 3.2 in [14]. Because the stochastic gradients g_t^m are independent we have that the variance of their sum is the sum of their variances, hence

$$\mathbb{E}\left[\|g_t - \bar{g}_t\|^2\right] = \frac{1}{M^2}\mathbb{E}\left[\left\|\sum_{m=1}^M g_t^m - \bar{g}_t^m\right\|^2\right] = \frac{1}{M^2}\sum_{m=1}^M \mathbb{E}\left[\|g_t^m - \bar{g}_t^m\|^2\right] \leqslant \frac{\sigma^2}{M}.$$

∎

D Proof of Theorem 1

Proof. Combining Lemma 2 and Lemma 3, we have

$$\mathbb{E}\left[\|\hat{x}_{t+1} - x_*\|^2\right] \leqslant (1-\gamma\mu)\mathbb{E}\left[\|\hat{x}_t - x_*\|^2\right] + \frac{\gamma^2\sigma^2}{M} - \frac{\gamma}{2}\mathbb{E}\left[D_f(\hat{x}_t, x_*)\right] + 2\gamma L V_t. \tag{22}$$

Using Lemma 1 we can upper bound the V_t term in (22):

$$\mathbb{E}\left[\|\hat{x}_{t+1} - x_*\|^2\right] \leqslant (1 - \gamma\mu)\mathbb{E}\left[\|\hat{x}_t - x_*\|^2\right] + \frac{\gamma^2\sigma^2}{M} - \frac{\gamma}{2}\mathbb{E}\left[D_f(\hat{x}_t, x_*)\right] + 2\gamma^3 LH\sigma^2.$$

Letting $r_{t+1} = \hat{x}_{t+1} - x_*$ and we have,

$$\mathbb{E}\left[\|r_{t+1}\|^2\right] \leqslant (1 - \gamma\mu)\mathbb{E}\left[\|r_t\|^2\right] + \frac{\gamma^2\sigma^2}{M} + 2\gamma^3 LH\sigma^2.$$

Recursing the above inequality we have,

$$\mathbb{E}\left[\|r_T\|^2\right] \leqslant (1 - \gamma\mu)^T \mathbb{E}\left[\|r_0\|^2\right] + \left(\sum_{t=0}^{T-1}(1 - \gamma\mu)^t\right)\left(\frac{\gamma^2\sigma^2}{M} + 2\gamma^3 LH\sigma^2\right).$$

Using that $\sum_{t=0}^{T-1}(1 - \gamma\mu)^t \leqslant \frac{1}{\gamma\mu}$ we have,

$$\mathbb{E}\left[\|r_T\|^2\right] \leqslant (1 - \gamma\mu)^T \mathbb{E}\left[\|r_0\|^2\right] + \frac{\gamma\sigma^2}{\mu M} + \frac{2\gamma^2 LH\sigma^2}{\mu},$$

which is the claim of this theorem. $\blacksquare$

E   Proof of Theorem 2

Proof. Let $r_t = \hat{x}_t - x_*$, then putting $\mu = 0$ in Lemma 2 and combining it with Lemma 3, we have

$$\mathbb{E}\left[\|r_{t+1}\|^2\right] \leqslant \mathbb{E}\left[\|r_t\|^2\right] + \frac{\gamma^2\sigma^2}{M} - \frac{\gamma}{2}\mathbb{E}\left[D_f(\hat{x}_t, x_*)\right] + 2\gamma L V_t.$$

Further using Lemma 1,

$$\mathbb{E}\left[\|r_{t+1}\|^2\right] \leqslant \mathbb{E}\left[\|r_t\|^2\right] + \frac{\gamma^2\sigma^2}{M} - \frac{\gamma}{2}\mathbb{E}\left[D_f(\hat{x}_t, x_*)\right] + 2\gamma^3 LH\sigma^2.$$

Rearranging we have,

$$\frac{\gamma}{2}\mathbb{E}\left[D_f(\hat{x}_t, x_*)\right] \leqslant \mathbb{E}\left[\|r_t\|^2\right] - \mathbb{E}\left[\|r_{t+1}\|^2\right] + \frac{\gamma^2\sigma^2}{M} + 2\gamma^3 LH\sigma^2.$$

Averaging the above equation as t varies between 0 and $T - 1$,

$$\frac{\gamma}{2T}\sum_{t=0}^{T-1}\mathbb{E}\left[D_f(\hat{x}_t, x_*)\right] \leqslant \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|r_t\|^2\right] - \mathbb{E}\left[\|r_{t+1}\|^2\right] + \frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{\gamma^2\sigma^2}{M} + 2\gamma^3 LH\sigma^2\right)$$

$$= \frac{\|r_0\|^2 - \mathbb{E}\left[\|r_T\|^2\right]}{T} + \frac{\gamma^2\sigma^2}{M} + 2\gamma^3 LH\sigma^2$$

$$\leqslant \frac{\|r_0\|^2}{T} + \frac{\gamma^2\sigma^2}{M} + 2\gamma^3 LH\sigma^2. \tag{23}$$

By Jensen's inequality we have $D_f(\bar{x}_T, x_*) \leqslant \frac{1}{T}\sum_{t=0}^{T-1}D_f(\hat{x}_t, x_*)$. Using this in (23) we have,

$$\frac{\gamma}{2}\mathbb{E}\left[D_f(\hat{x}_t, x_*)\right] \leqslant \frac{\|r_0\|^2}{T} + \frac{\gamma^2\sigma^2}{M} + 2\gamma^3 LH\sigma^2.$$

Dividing both sides by $\gamma/2$ yields the theorem's claim. $\blacksquare$