First Analysis of Local GD on Heterogeneous Data

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 provide the first convergence analysis of local gradient descent for minimizing the average of smooth and convex but otherwise arbitrary functions. Problems of this form and local gradient descent as a solution method are of importance in federated learning, where each function is based on private data stored by a user on a mobile device, and the data of different users can be arbitrarily heterogeneous. We show that in a low accuracy regime, the method has the same communication complexity as gradient descent.

1 Introduction

We are interested in solving the optimization problem

$$\min_{x\in\mathbb{R}^d}\left\{f(x)\overset{\text{def}}{=}\frac{1}{M}\sum_{m=1}^{M}f_m(x)\right\}, \tag{1}$$

which is arises in training of supervised machine learning models. We assume that each $f_m:\mathbb{R}^d\to\mathbb{R}$ is an L-smooth and convex function and we denote by x_* a fixed minimizer of f.

Our main interest is in situations where each function is based on data available on a single device only, and where the data distribution across the devices can be arbitrarily heterogeneous. This situation arises in *federated learning*, where machine learning models are trained on data available on consumer devices, such as mobile phones. In federated learning, transfer of local data to a single data center for centralized training is prohibited due to privacy reasons, and frequent communication is undesirable as it is expensive and intrusive. Hence, several recent works aim at constructing new ways of solving (1) in a distributed fashion with as few communication rounds as possible.

Large-scale problems are often solved by first-order methods as they have proved to scale well with both dimension and data size. One attractive choice is *Local Gradient Descent*, which divides the optimization process into epochs. Each epoch starts by communication in the form of a model averaging step across all M devices.[3] The rest of each epoch does not involve any communication, and is devoted to performing a fixed number of gradient descent steps initiated from the average model, and based on the local functions, performed by all M devices independently in parallel. See Algorithm 1 for more details.

[*] Work done during an internship at KAUST.

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

[3] In the practice of federated learning, averaging is performed over a subset of devices only. Usually, only those updates are averaged which are received by a certain time window. Here we focus on an idealized scenario

Algorithm 1 Local Gradient Descent

```
Input: Stepsize $\gamma > 0$, synchronization/communication times $0 = t_0 \leqslant t_1 \leqslant t_2 \leqslant \dots$, initial
    vector $x_0 \in \mathbb{R}^d$
Initialize $x_0^m = x_0$ for all $m \in [M] \overset{\text{def}}{=} \{1, 2, \dots, M\}$
for $t = 0, 1, \dots$ do
    for $m = 1, \dots, M$ do
        $x_{t+1}^m = \begin{cases} \frac{1}{M}\sum_{j=1}^M (x_t^j - \gamma \nabla f_j(x_t^j)), & \text{if } t = t_p \text{ for some } p \in \{1, 2, \dots\} \\ x_t^m - \gamma \nabla f_m(x_t^m), & \text{otherwise.} \end{cases}$
    end for
end for
```

The stochastic version of this method is at the core of the *Federated Averaging* algorithm which has been used recently in federated learning applications, see e.g. [7, 10]. Essentially, Federated Averaging is a variant of local Stochastic Gradient Descent (SGD) with participating devices sampled randomly. This algorithm has been used in several machine learning applications such as mobile keyboard prediction [5], and strategies for improving its communication efficiency were explored in [7]. Despite its empirical success, little is known about convergence properties of this method and it has been observed to diverge when too many local steps are performed [10]. This is not so surprising as the majority of common assumptions are not satisfied; in particular, the data is typically very non-i.i.d. [10], so the local gradients can point in different directions. This property of the data can be written for any vector x and indices i, j as

$$\|\nabla f_i(x) - \nabla f_j(x)\| \gg 1.$$

Unfortunately, it is very hard to analyze local methods without assuming a bound on the dissimilarity of $\nabla f_i(x)$ and $\nabla f_j(x)$. For this reason, almost all prior work assumed bounded dissimilarity [8, 16, 17, 18] and addressed other less challenging aspects of federated learning such as decentralized communication, nonconvexity of the objective or unbalanced data partitioning. In fact, a common way to make the analysis simple is to assume Lipschitzness of local functions, $\|\nabla f_i(x)\| \leqslant G$ for any x and i. We argue that this assumption is pathological and should be avoided when seeking a meaningful convergence bound. First of all, in unconstrained strongly convex minimization this assumption can not be satisfied, making the analysis in works like [14] questionable. Second, there exists at least one method, whose convergence is guaranteed under bounded gradients [6], but in practice the method diverges [3, 12].

Finally, under the bounded gradients assumption we have

$$\|\nabla f_i(x) - \nabla f_j(x)\| \leqslant \|\nabla f_i(x)\| + \|\nabla f_j(x)\| \leqslant 2G. \tag{2}$$

In other words, we lose control over the difference between the functions. Since G bounds not just dissimilarity, but also the gradients themselves, it makes the statements less insightful or even vacuous. For instance, it is not going to be tight if the data is actually i.i.d. since G in that case will remain a positive constant. In contrast, we will show that the rate should depend on a much more meaningful quantity,

$$\sigma_f^2 \overset{\text{def}}{=} \frac{1}{M}\sum_{m=1}^M \|\nabla f_m(x_*)\|^2,$$

where x_* is a minimizer of f. Obviously, σ_f is always finite and it serves as a natural measure of variance in local methods. On top of that, it allows us to obtain bounds that are tight in case the data is actually i.i.d. We note that an attempt to get more general convergence statement has been made in [13], but sadly their guarantee is strictly worse than that of minibatch Stochastic Gradient Descent (SGD), making their contribution minor.

We additionally note that the bound in the mentioned work [8] not only uses bounded gradients, but also provides a pessimistic $\mathcal{O}(H^2/T)$ rate, where H is the number of local steps in each epoch, and T is the total number of steps of the method. Indeed, this requires H to be $\mathcal{O}(1)$ to make the

where averaging is done across all devices. We focus on this simpler situation first as even this is not currently understood theoretically.

rate coincide with that of SGD for strongly convex functions. The main contribution of that work, therefore, is in considering partial participation as in Federated Averaging.

When the data is identically distributed and stochastic gradients are used instead of full gradients on each node, the resulting method has been explored extensively in the literature under different names, see e.g. [1, 14, 15, 19]. [11] proposed an asynchronous local method that converges to the exact solution without decreasing stepsizes, but its benefit from increasing H is limited by constant factors. [9] seems to be the first work to propose a local method, but no rate was shown in that work.

2 Convergence of Local GD

2.1 Assumptions and notation

Before introducing our main result, let us first formulate explicitly our assumptions.

Assumption 1. The set of minimizers of (1) is nonempty. Further, for every $m \in [M] \overset{\text{def}}{=} \{1, 2, \ldots, M\}$, f_m is convex and L-smooth. That is, for all $x, y \in \mathbb{R}^d$ the following inequalities are satisfied:

$$0 \leqslant f_m(x) - f_m(y) - \langle \nabla f_m(y), x - y \rangle \leqslant \tfrac{L}{2} \|x - y\|^2.$$

Further, we assume that Algorithm 1 is run with a bounded synchronization interval. That is, we assume that

$$H \overset{\text{def}}{=} \max_{p \geqslant 0} |t_p - t_{p+1}|$$

is finite. Given local vectors $x_t^1, x_t^2, \ldots, x_t^M \in \mathbb{R}^d$, we define the average iterate, iterate variance and average gradient at time t as

$$\hat{x}_t \overset{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} x_t^m \qquad V_t \overset{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} \|x_t^m - \hat{x}_t\|^2 \qquad g_t \overset{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} \nabla f_m(x_t^m), \tag{3}$$

respectively. The Bregman divergence with respect to f is defined via

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

Note that in the case $y = x_*$, we have $D_f(x, x_*) = f(x) - f(x_*)$.

2.2 Analysis

The first lemma enables us to find a recursion on the optimality gap for a single step of local GD:

Lemma 1. Under Assumption 1 and for any $\gamma \geqslant 0$ we have

$$\|r_{t+1}\|^2 \leqslant \|r_t\|^2 + \gamma L (1 + 2\gamma L) V_t - 2\gamma (1 - 2\gamma L) D_f(\hat{x}_t, x_*), \tag{4}$$

where $r_t \overset{\text{def}}{=} \hat{x}_t - x_*$. In particular, if $\gamma \leqslant \frac{1}{4L}$, then $\|r_{t+1}\|^2 \leqslant \|r_t\|^2 + \frac{3}{2}\gamma L V_t - \gamma D_f(\hat{x}_t, x_*)$.

We now bound the sum of the variances V_t *over an epoch*. An epoch-based bound is intuitively what we want since we are only interested in the points $\hat{x}_{t_p}$ produced at the end of each epoch.

Lemma 2. Suppose that Assumption 1 holds and let $p \in \mathbb{N}$, define $v = t_{p+1} - 1$ and suppose Algorithm 1 is run with a synchronization interval $H \geqslant 1$ and a constant stepsize $\gamma > 0$ such that $\gamma \leqslant \frac{1}{4LH}$. Then the following inequalities hold:

$$\sum_{t=t_p}^{v} V_t \leqslant 5L\gamma^2 H^2 \sum_{i=t_p}^{v} D_f(\hat{x}_i, x_*) + \sum_{i=t_p}^{v} 8\gamma^2 H^2 \sigma_f^2,$$

$$\sum_{t=t_p}^{v} \frac{3}{2} L V_t - D_f(\hat{x}_t, x_*) \leqslant -\frac{1}{2} \sum_{t=t_p}^{v} D_f(\hat{x}_i, x_*) + \sum_{t=t_p}^{v} 12L\gamma^2 H^2 \sigma_f^2.$$

Combining the previous two lemmas, the convergence of local GD is established in the next theorem:

Theorem 1. For local GD run with a constant stepsize $\gamma > 0$ such that $\gamma \leqslant \frac{1}{4LH}$ and under Assumption 1, we have

$$f(\bar{x}_T) - f(x_*) \leqslant \frac{2\|x_0 - x_*\|^2}{\gamma T} + 24\gamma^2 \sigma_f^2 H^2 L, \tag{5}$$

where $\bar{x}_T \overset{\text{def}}{=} \frac{1}{T}\sum_{t=0}^{T-1} \hat{x}_t$.

2.3 Local GD vs GD

In order to interpret the above bound, we may ask: how many communication rounds are sufficient to guarantee $f(\bar{x}_T) - f(x_*) \leqslant \epsilon$? To answer this question, we need to minimize $\frac{T}{H}$ subject to the constraints $0 < \gamma \leqslant \frac{1}{4L}$, $\frac{2\|x_0 - x_*\|^2}{\gamma T} \leqslant \frac{\epsilon}{2}$, and $24\gamma^2\sigma_f^2 H^2 L \leqslant \frac{\epsilon}{2}$, in variables T, H and γ. We can easily deduce from the constraints that

$$\frac{T}{H} \geqslant \frac{16\|x_0 - x_*\|^2}{\epsilon} \max\left\{L, \sigma_f \sqrt{\frac{3L}{\epsilon}}\right\}. \tag{6}$$

On the other hand, this lower bound is achieved by *any* $0 < \gamma \leqslant \frac{1}{4L}$ as long as we pick

$$T = T(\gamma) \overset{\text{def}}{=} \frac{4\|x_0 - x_*\|^2}{\epsilon\gamma} \qquad \text{and} \qquad H = H(\gamma) \overset{\text{def}}{=} \frac{1}{4\max\left\{L, \sigma_f\sqrt{\frac{3L}{\epsilon}}\right\}\gamma}.$$

The smallest H achieving this lower bound is $H(\frac{1}{4L}) = \min\left\{1, \sqrt{\frac{\epsilon L}{3\sigma_f^2}}\right\}$.

Further, notice that as long as the target accuracy is not too high, in particular $\epsilon \geqslant \frac{3\sigma^2}{L}$, then $\max\left\{L, \sigma_f\sqrt{3L/\epsilon}\right\} = L$ and (6) says that the number of communications of local GD (with parameters set as $H = H(\gamma)$ and $T = T(\gamma)$) is equal to

$$\frac{T}{H} = \mathcal{O}\left(\frac{L\|x_0 - x_*\|^2}{\epsilon}\right),$$

which is the same as the number of iterations (i.e., communications) of gradient descent. If $\epsilon < \frac{3\sigma^2}{L}$, then (6) gives the communication complexity

$$\frac{T}{H} = \mathcal{O}\left(\frac{\sqrt{L}\sigma_f}{\epsilon^{3/2}}\right).$$

2.4 Local GD vs Minibatch SGD

Equation (5) shows a clear analogy between the convergence of local GD and the convergence rate of minibatch SGD, establishing a $1/T$ convergence to a neighborhood depending on the expected noise at the optimum σ_f^2, which measures how dissimilar the functions f_m are from each other at the optimum x_*.

The analogy between SGD and local GD extends further to the convergence rate, as the next corollary shows:

Corollary 1. Choose H such that $H \leqslant \frac{\sqrt{T}}{\sqrt{M}}$, then $\gamma = \frac{\sqrt{M}}{4L\sqrt{T}} \leqslant \frac{1}{4HL}$, and hence applying the result of the previous theorem

$$f(\bar{x}_T) - f(x_*) \leqslant \frac{8L\|x_0 - x_*\|^2}{\sqrt{MT}} + \frac{3M\sigma_f^2 H^2}{2LT}.$$

To get a convergence rate of $1/\sqrt{MT}$ we can choose $H = O\left(T^{1/4}M^{-3/4}\right)$, which implies a total number of $\Omega(T^{3/4}M^{3/4})$ communication steps. If a rate of $1/\sqrt{T}$ is desired instead, we can choose a larger $H = O\left(T^{1/4}\right)$.

3 Experiments

To verify the theory, we run our experiments on logistic regression with ℓ_2 regularization and datasets taken from the LIBSVM library [2]. We use a machine with 24 Intel(R) Xeon(R) Gold 6146 CPU @ 3.20GHz cores and we handle communication via the MPI for Python package [4].

Since our architecture leads to a very specific trade-off between computation and communication, we also provide plots for the case the communication time relative to gradient computation time is higher or lower. In all experiments, we use full gradients ∇f_m and constant stepsize $\frac{1}{L}$. The amount of ℓ_2 regularization was chosen of order $\frac{1}{n}$, where n is the total amount of data. The data partitioning is not i.i.d. and is done based on the index in the original dataset.

We observe a very tight match between our theory and numerical results. In cases where communication is significantly more expensive than gradient computation, local methods are much faster for imprecise convergence. This was not a big advantage though with our architecture, mainly because full gradients took a lot of time to be computed.

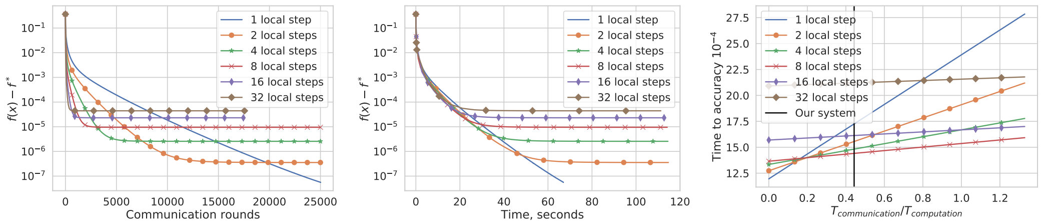

Figure 1: Convergence of local GD methods with different number of local steps on the 'a5a' dataset. 1 local step corresponds to fully synchronized gradient descent and it is the only method that converges precisely to the optimum. The left plot shows convergence in terms of communication rounds, showing a clear advantage of local GD when only limited accuracy is required. The mid plot, however, illustrates that wall-clock time might improve only slightly and the right plot shows what changes with different communication cost.

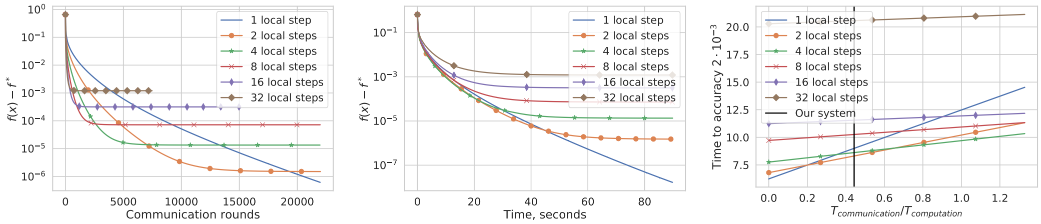

Figure 2: Same experiment as in Figure 3, performed on the 'mushrooms' dataset.

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] 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.

[3] Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing Noise in GAN Training with Variance Reduced Extragradient. *arXiv preprint arXiv:1904.08598*, 2019.

[4] 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.

[5] Andrew Hard, Kanishka Rao, Rajiv Mathews, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated Learning for Mobile Keyboard Prediction. *arXiv:1811.03604*, 2018.

[6] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational Inequalities with Stochastic Mirror-Prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[7] 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.

[8] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the Convergence of FedAvg on Non-IID Data. *arXiv:1907.02189*, 2019.

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

[10] 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.

[11] Konstantin Mishchenko, Franck Iutzeler, Jérôme Malick, and Massih-Reza Amini. A Delay-tolerant Proximal-Gradient Algorithm for Distributed Learning. In *International Conference on Machine Learning*, pages 3584–3592, 2018.

[12] Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting Stochastic Extragradient. *arXiv preprint arXiv:1905.11373*, 2019.

[13] Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. On the Convergence of Federated Optimization in Heterogeneous Networks. *arXiv:1812.06127*, 2018.

[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] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K. Leung, Christian Makaya, Ting He, and Kevin Chan. When Edge Meets Learning: Adaptive Control for Resource-Constrained Distributed Machine Learning. *arXiv:1804.05271*, 2018.

[17] 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.

[18] Hao Yu, Rong Jin, and Sen Yang. On the Linear Speedup Analysis of Communication Efficient Momentum SGD for Distributed Non-Convex Optimization. *arXiv preprint arXiv:1905.03817*, 2019.

[19] 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 for: First Analysis of Local GD on Heterogeneous Data

4 Proofs

We first provide two technical lemmas which relate the quantities $\hat{x}_t$, V_t and g_t, x_* and $\nabla f_m(x_t^m)$ for $m = 1, 2, \dots, M$. These lemmas are independent of the algorithm.

Lemma 3. If Assumption 1 holds, then

$$\|g_t\|^2 \leqslant 2L^2 V_t + 4L D_f(\hat{x}_t, x_*). \tag{7}$$

Proof. Starting with the left-hand side,

$$\begin{aligned} \|g_t\|^2 &\leqslant 2\|g_t - \nabla f(\hat{x}_t)\|^2 + 2\|\nabla f(\hat{x}_t)\|^2 \\ &= 2\left\| \frac{1}{M}\sum_{m=1}^{M} \nabla f_m(x_t^m) - \frac{1}{M}\sum_{m=1}^{M} \nabla f_m(\hat{x}_t) \right\|^2 + 2\|\nabla f(\hat{x}_t)\|^2 \\ &\leqslant \frac{2}{M}\sum_{m=1}^{M} \|\nabla f_m(x_t^m) - \nabla f_m(\hat{x}_t)\|^2 + 2\|\nabla f(\hat{x}_t)\|^2 \\ &\leqslant \frac{2L^2}{M}\sum_{m=1}^{M} \|x_t^m - \hat{x}_t\|^2 + 2\|\nabla f(\hat{x}_t)\|^2, \end{aligned}$$

where in the second inequality we have used convexity of the map $x \mapsto \|x\|^2$. The claim of the lemma follows by noting that

$$\|\nabla f(\hat{x}_t)\|^2 = \|\nabla f(\hat{x}_t) - \nabla f(x_*)\|^2 \leqslant 2L D_f(\hat{x}_t, x_*).$$

■

Lemma 4. Suppose that Assumption 1 holds. Then,

$$\frac{-2}{M}\sum_{m=1}^{M} \langle \hat{x}_t - x_*, \nabla f_m(x_t^m) \rangle \leqslant -2D_f(\hat{x}_t, x_*) + L V_t. \tag{8}$$

Proof. Starting with the left-hand side,

$$-2\langle \hat{x}_t - x_*, \nabla f_m(x_t^m)\rangle = -2\langle x_t^m - x_*, \nabla f_m(x_t^m)\rangle - 2\langle \hat{x}_t - x_t^m, \nabla f_m(x_t^m)\rangle. \tag{9}$$

The first term in (9) is bounded by convexity:

$$-\langle x_t^m - x_*, \nabla f_m(x_t^m)\rangle \leqslant f_m(x_*) - f_m(x_t^m). \tag{10}$$

For the second term, we use L-smoothness,

$$-\langle \hat{x}_t - x_t^m, \nabla f_m(x_t^m)\rangle \leqslant f_m(x_t^m) - f_m(\hat{x}_t) + \frac{L}{2}\|x_t^m - \hat{x}_t\|^2. \tag{11}$$

Combining (11) and (10) in (9),

$$\begin{aligned} -2\langle \hat{x}_t - x_*, \nabla f_m(x_t^m)\rangle &\leqslant 2\left(f_m(x_*) - f_m(x_t^m)\right) \\ &\quad + 2\left(f_m(x_t^m) - f_m(\hat{x}_t) + \frac{L}{2}\|x_t^m - \hat{x}_t\|^2\right) \\ &= 2\left(f_m(x_*) - f_m(\hat{x}_t) + \frac{L}{2}\|x_t^m - \hat{x}_t\|^2\right). \end{aligned}$$

Averaging over m,

$$\frac{-2}{M}\sum_{m=1}^{M}\left\langle \hat{x}_t - x_*, \nabla f_m(x_t^m)\right\rangle \leqslant -2\left(f(\hat{x}_t) - f(x_*)\right) + \frac{L}{M}\sum_{m=1}^{M}\|x_t^m - \hat{x}_t\|^2.$$

Note that the first term is the Bregman divergence $D_f(\hat{x}_t, x_*)$ and the second term is just V_t, hence

$$\frac{-2}{M}\sum_{m=1}^{M}\left\langle \hat{x}_t - x_*, \nabla f_m(x_t^m)\right\rangle \leqslant -2D_f(\hat{x}_t, x_*) + LV_t,$$

which is the claim of this lemma. ■

Proof of Lemma 1. Note that $\hat{x}_{t+1} = \hat{x}_t - \gamma g_t$ always holds. Then we have,

$$\begin{aligned}\|\hat{x}_{t+1} - x_*\|^2 &= \|\hat{x}_t - \gamma g_t - x_*\|^2\\ &= \|\hat{x}_t - x_*\|^2 + \gamma^2\|g_t\|^2 - 2\gamma\left\langle \hat{x}_t - x_*, g_t\right\rangle\\ &= \|\hat{x}_t - x_*\|^2 + \gamma^2\|g_t\|^2 - \frac{2\gamma}{M}\sum_{m=1}^{M}\left\langle \hat{x}_t - x_*, \nabla f_m(x_t^m)\right\rangle\end{aligned}$$

Let $r_t = \hat{x}_t - x_*$, then using Lemmas 3 and 4,

$$\begin{aligned}\|r_{t+1}\|^2 &\leqslant \|r_t\|^2 + \gamma^2\|g_t\|^2 - \frac{2\gamma}{M}\sum_{m=1}^{M}\left\langle \hat{x}_t - x_*, \nabla f_m(x_t^m)\right\rangle\\ &\overset{(7)}{\leqslant} \|r_t\|^2 + \gamma^2\left(2L^2V_t + 4LD_f(\hat{x}_t, x_*)\right) - \frac{2\gamma}{M}\sum_{m=1}^{M}\left\langle \hat{x}_t - x_*, \nabla f_m(x_t^m)\right\rangle\\ &\overset{(8)}{\leqslant} (1-\gamma\mu)\|r_t\|^2 + \gamma L(1+2\gamma L)V_t - 2\gamma(1-2\gamma L)D_f(\hat{x}_t, x_*).\end{aligned}$$

If $\gamma \leqslant \frac{1}{4L}$, then $1 - 2\gamma L \geqslant \frac{1}{2}$ and $1 + 2\gamma L \leqslant \frac{3}{2}$, and hence

$$\|r_{t+1}\|^2 \leqslant \|r_t\|^2 + \frac{3}{2}\gamma LV_t - \gamma D_f(\hat{x}_t, x_*),$$

as claimed. ■

Proof of Lemma 2. Let $g_t^m = \nabla f(x_t^m)$, then noting that $x_{t+1}^m = x_t^m - \gamma g_t^m$ when $t_{p+1} > t > t_p$ and $x_{t_p}^m = \hat{x}_{t_p}$ we have,

$$\begin{aligned}V_t = \frac{1}{M}\sum_{m=1}^{M}\|x_t^m - \hat{x}_t\|^2 &= \frac{1}{M}\sum_{m=1}^{M}\left\|x_{t_p}^m - \hat{x}_{t_p} - \gamma\sum_{i=t_p}^{t} g_i^m - g_i\right\|^2\\ &= \frac{\gamma^2}{M}\sum_{m=1}^{M}\left\|\sum_{i=t_p}^{t}\left(g_i^m - g_i\right)\right\|^2\\ &\leqslant \frac{\gamma^2}{M}\sum_{m=1}^{M}(t-t_p)\sum_{i=t_p}^{t}\|g_i^m - g_i\|^2\\ &\leqslant \frac{\gamma^2 H}{M}\sum_{m=1}^{M}\sum_{i=t_p}^{t}\|g_i^m - g_i\|^2\\ &\leqslant \frac{\gamma^2 H}{M}\sum_{m=1}^{M}\sum_{i=t_p}^{t}\|g_i^m\|^2.\end{aligned}$$

Then we have

$$\begin{aligned}\|g_i^m\|^2 &\leqslant (1+\alpha)\|g_i^m - \nabla f_m(\hat{x}_i)\|^2 + \left(1+\alpha^{-1}\right)\|\nabla f_m(\hat{x}_i)\|^2 \\ &\leqslant (1+\alpha)\|g_i^m - \nabla f_m(\hat{x}_i)\|^2 + \left(1+\alpha^{-1}\right)(1+\beta)\|\nabla f_m(\hat{x}_i) - \nabla f_m(x_*)\|^2 \\ &+ \left(1+\alpha^{-1}\right)\left(1+\beta^{-1}\right)\|\nabla f_m(x_*)\|^2.\end{aligned}$$

Putting $\alpha = 2$, $\beta = \frac{1}{3}$, we get

$$\begin{aligned}\|g_i^m\|^2 &\leqslant 3\|g_i^m - \nabla f_m(\hat{x}_i)\|^2 + 2\|\nabla f_m(\hat{x}_i) - \nabla f_m(x_*)\|^2 + 6\|\nabla f_m(x_*)\|^2 \\ &\leqslant 3L^2\|x_t^m - \hat{x}_t\|^2 + 2\left(2LD_{f_m}(\hat{x}_t, x_*)\right) + 6\|\nabla f_m(x_*)\|^2.\end{aligned}$$

Averaging with respect to m,

$$\frac{1}{M}\sum_{m=1}^{M}\|g_t^m\|^2 \leqslant 3L^2 V_t + 4LD_f(\hat{x}_t, x_*) + 6\sigma^2.$$

Hence we have,

$$V_t \leqslant \gamma^2 H \sum_{i=t_p}^{t}\frac{1}{M}\sum_{m=1}^{M}\|g_t^m\|^2 \leqslant \gamma^2 H\sum_{i=t_p}^{t}\left(3L^2V_i + 4LD_f(\hat{x}_i, x_*) + 6\sigma^2\right).$$

Summing up the above inequality as t varies from t_p to $v = t_{p+1} - 1$,

$$\begin{aligned}\sum_{t=t_p}^{v} V_t &\leqslant \gamma^2 H \sum_{t=t_p}^{v}\sum_{i=t_p}^{t}\left(3L^2V_i + 4LD_f(\hat{x}_i, x_*) + 6\sigma^2\right) \\ &\leqslant \gamma^2 H \sum_{t=t_p}^{v}\sum_{i=t_p}^{v}\left(3L^2V_i + 4LD_f(\hat{x}_i, x_*) + 6\sigma^2\right) \\ &= 3L^2\gamma^2H^2\sum_{i=t_p}^{v}V_i + 4L\gamma^2H^2\sum_{i=t_p}^{v}D_f(\hat{x}_i, x_*) + \sum_{i=t_p}^{v}6\gamma^2H^2\sigma^2.\end{aligned}$$

Noting that the sum $\sum_{t=t_p}^{v} V_t$ appears on both sides, we have

$$\left(1 - 3L^2\gamma^2H^2\right)\sum_{t=t_p}^{v}V_t \leqslant 4L\gamma^2H^2\sum_{i=t_p}^{v}D_f(\hat{x}_i, x_*) + 6\gamma^2H^2\sigma^2.$$

Note that because $\gamma \leqslant \frac{1}{4LH} \leqslant \frac{1}{\sqrt{15}LH}$, then our choice of γ implies that $1 - 3L^2\gamma^2H^2 \geqslant \frac{4}{5}$, hence

$$\sum_{t=t_p}^{v}V_t \leqslant 5L\gamma^2H^2\sum_{i=t_p}^{v}D_f(\hat{x}_i, x_*) + \sum_{i=t_p}^{v}\frac{15}{2}\gamma^2H^2\sigma^2.$$

For the second part, we have

$$\sum_{t=t_p}^{v}\frac{3}{2}LV_t - \sum_{i=t_p}^{v}D_f(\hat{x}_i, x_*) \leqslant \left(\frac{15}{2}L^2\gamma^2H^2 - 1\right)\sum_{i=t_p}^{v}D_f(\hat{x}_i, x_*) + \sum_{i=t_p}^{v}\frac{45}{4}L\gamma^2H^2\sigma^2.$$

Using that $\gamma \leqslant \frac{1}{4LH}$ we get that $\frac{15}{2}L^2\gamma^2H^2 - 1 \leqslant \frac{-1}{2}$, and using this we get the second claim. ■

Proof of Theorem 1. Starting with Lemma 1, we have

$$\|r_{t+1}\|^2 \leqslant \|r_t\|^2 + \gamma\left(\frac{3}{2}LV_t - D_f(\hat{x}_t, x_*)\right) \leqslant \|r_t\|^2 + \gamma\left(2LV_t - D_f(\hat{x}_t, x_*)\right).$$

Summing up these inequalities gives

$$\sum_{i=1}^{T}\|r_t\|^2 \leqslant \sum_{i=0}^{T-1}\|r_t\|^2 + \gamma\sum_{i=0}^{T-1}\left(2LV_i - D_f(\hat{x}_i, x_*)\right),$$

and using that $T = t_p$ for some $p \in \mathbb{N}$, we can decompose the second term by double counting and bound it by Lemma 2, then using double counting again,

$$\begin{aligned}
\gamma \sum_{i=0}^{T-1} \left(2LV_i - D_f(\hat{x}_i, x_*)\right) &= \gamma \sum_{k=1}^{p} \sum_{i=t_{k-1}}^{t_k - 1} \left(2LV_i - D_f(\hat{x}_i, x_*)\right) \\
&\leqslant \frac{-\gamma}{2} \sum_{k=1}^{p} \sum_{i=t_p}^{v} \left(D_f(\hat{x}_i, x_*)\right) + \sum_{k=1}^{p} \sum_{i=t_p}^{v} 12L\gamma^3 H^2 \sigma^2 \\
&= \frac{-\gamma}{2} \sum_{i=0}^{T-1} \left(D_f(\hat{x}_i, x_*)\right) + \sum_{i=0}^{T-1} 12L\gamma^3 H^2 \sigma^2.
\end{aligned}$$

Using this in the previous bound,

$$\sum_{i=1}^{T} \|r_t\|^2 \leqslant \sum_{i=0}^{T-1} \|r_t\|^2 - \frac{\gamma}{2} \sum_{i=0}^{T-1} \left(D_f(\hat{x}_i, x_*)\right) + \sum_{i=0}^{T-1} \left(12L\gamma^3 H^2 \sigma^2\right).$$

Rearranging we get,

$$\begin{aligned}
\frac{\gamma}{2} \sum_{i=0}^{T-1} \left(D_f(\hat{x}_i, x_*)\right) &\leqslant \sum_{i=0}^{T-1} \|r_t\|^2 - \sum_{i=1}^{T} \|r_t\|^2 + \sum_{i=0}^{t-1} \left(12L\gamma^3 H^2 \sigma^2\right) \\
&= \|r_0\|^2 - \|r_T\|^2 + \sum_{i=0}^{T-1} \left(12L\gamma^3 H^2 \sigma^2\right) \\
&\leqslant \|r_0\|^2 + T\left(12L\gamma^3 H^2 \sigma^2\right).
\end{aligned}$$

Dividing both sides by $\gamma T/2$ we get,

$$\frac{1}{T} \sum_{i=0}^{T-1} \left(D_f(\hat{x}_i, x_*)\right) \leqslant \frac{2r_0}{\gamma T} + 24L\gamma^2 H^2 \sigma^2.$$

Finally, using Jensen's inequality and the convexity of f we get the required claim. ■