arXiv:2004.01442v1 [cs.LG] 3 Apr 2020

From Local SGD to Local Fixed Point Methods for Federated Learning

Grigory Malinovsky*[2], Dmitry Kovalev[1], Elnur Gasanov[1], Laurent Condat†[1], and Peter Richtárik[1]

[1]King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
[2]Moscow Institute of Physics and Technology

February 20, 2020

Abstract

Most algorithms for solving optimization problems or finding saddle points of convex-concave functions are fixed point algorithms. In this work we consider the generic problem of finding a fixed point of an average of operators, or an approximation thereof, in a distributed setting. Our work is motivated by the needs of federated learning. In this context, each local operator models the computations done locally on a mobile device. We investigate two strategies to achieve such a consensus: one based on a fixed number of local steps, and the other based on randomized computations. In both cases, the goal is to limit communication of the locally-computed variables, which is often the bottleneck in distributed frameworks. We perform convergence analysis of both methods and conduct a number of experiments highlighting the benefits of our approach.

1 Introduction

In the 'big data' era, the explosion in size and complexity of the data arises in parallel to a shift towards distributed computations, as modern hardware increasingly relies on the power of uniting many parallel units into one system. For distributed optimization tasks, specific issues arise, such as decentralized data storage. For instance, the huge amount of mobile phones or smart home devices in the world contain an important volume of data captured and stored on each of them. This data contains a wealth of potentially useful information to their owners, and more so if appropriate machine learning models could be trained on the heterogeneous data stored across the network of such devices. Yet, many users are increasingly sensitive to privacy concerns and prefer their data to never leave their devices. But the only way to share knowledge while not having all data in one place is to communicate to keep moving towards the solution of the overall problem. Typically, mobile phones communicate back and forth with a distant server, so that a global model is progressively improved and converges to a steady state, which is globally optimal for all users. This is precisely the purpose of the recent and rising paradigm of *federated learning* [1, 2] where typically a global supervised model is trained in a massively distributed manner over a network of heterogeneous devices. Communication, which can be costly and slow, is the main bottleneck in this framework. So, it is of primary importance to devise novel algorithmic strategies, where the computation and communication loads are balanced.

*Work done during an internship at KAUST
†Corresponding author. Contact: see https://lcondat.github.io/

A strategy increasingly used by practitioners is to make use of *local computations*; that is, more local computations are performed on each device before communication and subsequent model averaging, with the hope that this will reduce the total number of communications needed to obtain a globally meaningful solution. Thus, local gradient descent methods have been investigated [3] [4] [5]; [6], [7]. Despite their practical success, local methods are little understood and there is much to be discovered. In this paper, we don't restrict ourselves to gradient descent to minimize an average of smooth functions; we consider the much broader setting of finding a fixed point of an average of large number M of operators. Indeed, most, if not all, iterative methods are fixed point methods, which aim at finding a fixed point of some operator [8]. A prominent application setting is the following: M mobile phones are involved in the same learning task, in a collaborative way; they perform computations, based on their locally stored data, in parallel, independently on each other, without sharing their actual data. These computations are modeled as operators, which are applied repeatedly. From time to time, the smartphones communicate their information to a distant server, which aggregates all the received messages by simple averaging. Then a new version of the application running on the smartphones is released, and a new cycle of collaborative learning starts from the new model, to keep improving it again and again.

1.1 Contributions

We model the setting of communication-efficient distributed fixed point optimization as follows: we have $M \geq 1$ parallel computing nodes. The variables handled by these nodes are modeled as vectors of the Euclidean space $\mathbb{R}^d$, endowed with the classical inner product, for some $d \geq 1$. Let $\mathcal{T}_i$, for $i = 1, \ldots, M$ be operators on $\mathbb{R}^d$, which model the set of operations during one iteration. We define the average operator

$$\mathcal{T} : x \in \mathbb{R}^d \mapsto \frac{1}{M} \sum_{i=1}^{M} \mathcal{T}_i(x). \tag{1}$$

Our goal is to find a fixed point of $\mathcal{T}$; that is, a vector $x^\star \in \mathbb{R}^d$ such that

$$\mathcal{T}(x^\star) = x^\star. \tag{2}$$

This ideal object $x^\star$ should be obtained by repeatedly applying $\mathcal{T}_i$ at each node, in parallel, with averaging steps to attain a consensus solution $x^\star$. Here we consider that, after some number of iterations, each node communicates its variable to a distant server, synchroneously. Then the server computes the average of the received vectors, and brodcasts it to all nodes.

We investigate two strategies. The first one consists, for each computing node, to iterate several times some sequence of operations; we call this *local steps*. The second strategy consists in reducing the number of communication steps by sharing information only with some low probability, and doing only local computations inbetween. We analyze two algorithms, which instantiate these two ideas, and we prove their convergence. Their good performances are illustrated by experiments.

1.2 Mathematical Background

Let T be an operator on $\mathbb{R}^d$. We denote by $\mathrm{Fix}(T)$ the set of its fixed points.

T is said to be χ-Lipschitz continuous, for some $\chi \geq 0$, if, for every x and y in $\mathbb{R}^d$,

$$\|T(x) - T(y)\| \leq \chi \|x - y\|. \tag{3}$$

Furthermore, T is said to be nonexpansive if it is 1-Lipschitz continuous, and χ-contractive, if it is χ-Lipschitz continuous, for some $\chi \in [0, 1)$. If T is contractive, its fixed point exists and is unique, see the Banach–Picard Theorem 1.50 in [9].

T is said to be α-averaged, for some $\alpha \in (0, 1]$, if $T = \alpha T' + \mathrm{Id}$ for some nonexpansive operator T', where Id denotes the identity. T is said to be firmly nonexpansive if it is 1/2-averaged.

A convex function f is said to be L-smooth, for some $L > 0$, if f is differentiable and its gradient is L-Lipschitz continuous.

Algorithm 1 Local distributed fixed point method

Input: Initial estimate $\hat{x}^0 \in \mathbb{R}^d$, stepsize $\lambda > 0$, sequence of synchronization times $0 = t_0 < t_1 < \dots$
Initialize: $x_i^0 \coloneqq \hat{x}^0$, for $i = 1, \dots, M$
for $k = 0, 1, \dots$ **do**
  for $i = 1, 2, \dots, M$ in parallel **do**
    $h_i^{k+1} \coloneqq (1-\lambda)x_i^k + \lambda \mathcal{T}_i(x_i^k)$
    if $k+1 = t_n$, for some $n \in \mathbb{N}$, **then**
      Communicate h_i^{k+1} to master node
    else
      $x_i^{k+1} \coloneqq h_i^{k+1}$
    end if
  end for
  if $k+1 = t_n$, for some $n \in \mathbb{N}$, **then**
    At master node: $\hat{x}^{k+1} \coloneqq \frac{1}{M}\sum_{i=1}^M h_i^{k+1}$
    Broadcast: $x_i^{k+1} \coloneqq \hat{x}^{k+1}$ for all $i = 1, \dots, M$
  end if
end for

2 A generic distributed fixed point method with local steps

We propose Algorithm 1, shown above. $(t_n)_{n\in\mathbb{N}}$ is the sequence of integers at which communication occurs. It proceeds as follows: at every iteration, the operator $\mathcal{T}_i$ is applied at node i, with further relaxation with parameter λ. After some number of iterations, each of the M computing nodes communicates its vector to a master node, which computes their average, and broadcasts it to all nodes. Thus, the later resume computing at the next iteration from the same variable $\hat{x}^k$.

We call an *epoch* a sequence of local iterations, followed by averaging; that is, the n-th epoch, for $n \geq 1$, is the sequence of iterations of indices $k = t_{n-1}+1, \dots, t_n$ (the 0-th epoch is the initialization step $x_i^0 \coloneqq \hat{x}^0$, for $i = 1, \dots, M$). We assume that the number of iterations in each epoch, between two aggregation steps, is bounded by some integer $H \geq 1$; that is,

Assumption 2.1. *$1 \leq t_n - t_{n-1} \leq H$, for every $n \geq 1$.*

To analyse Algorithm 1, we introduce the following averaged vector:

$$\hat{x}^k = \frac{1}{M}\sum_{i=1}^M x_i^k. \tag{4}$$

Note that this vector is actually computed only when k is one of the t_n.

In the uniform case $t_n = nH$, for every $n \in \mathbb{N}$, we introduce the operator

$$\widetilde{\mathcal{T}}_\lambda = \frac{1}{M}\sum_{i=1}^M \big(\lambda\mathcal{T}_i + (1-\lambda)\mathrm{Id}\big)^H, \tag{5}$$

where $\cdot^H$ denotes the composition of an operator with itself H times.

Thus, $x_1^{nH} = \dots = x_M^{nH} = \hat{x}^{nH}$ is the variable shared by every node at the end of the n-th epoch. We have, for every $n \in \mathbb{N}$,

$$\hat{x}^{(n+1)H} = \widetilde{\mathcal{T}}_\lambda \hat{x}^{nH}. \tag{6}$$

We also assume that the following holds:

Assumption 2.2. $\mathrm{Fix}(\mathcal{T})$ *and* $\mathrm{Fix}(\widetilde{\mathcal{T}}_\lambda)$ *are nonempty.*

Note that the fixed points of $\widetilde{\mathcal{T}}_\lambda$ depend on λ. The smaller λ, the closer $\mathrm{Fix}(\mathcal{T})$ and $\mathrm{Fix}(\widetilde{\mathcal{T}}_\lambda)$. But the smaller λ, the slower the convergence, so λ controls the tradeoff between precision and speed in estimating a fixed point of $\mathcal{T}$.

We first establish the convergence of Algorithm 1, in general conditions.

Theorem 2.3. *Suppose that* $t_n = nH$*, for every* $n \in \mathbb{N}$*, and suppose that the* $\mathcal{T}_i$ *are all* α*-averaged, for some* $\alpha \in (0,1]$*. Let* $\lambda \in (0, 1/\alpha)$ *be the parameter in Algorithm 1. Then the sequence* $(\hat{x}^{nH})_{n\in\mathbb{N}}$ *converges to a fixed point* $x^\dagger$ *of* $\widetilde{\mathcal{T}}_\lambda$*. In addition, the following hold:*
(i) $\widetilde{\mathcal{T}}_\lambda$ *is* ζ*-averaged, with*

$$\zeta = \frac{H\alpha\lambda}{1 + (H-1)\alpha\lambda}. \tag{7}$$

(ii) The distance between $\hat{x}^{nH}$ *and* $x^\dagger$ *decreases at every epoch: for every* $n \in \mathbb{N}$*,*

$$\|\hat{x}^{(n+1)H} - x^\dagger\|^2 \le \|\hat{x}^{nH} - x^\dagger\|^2 - \frac{1-\zeta}{\zeta}\|\hat{x}^{(n+1)H} - \hat{x}^{nH}\|^2. \tag{8}$$

(iii) The squared differences between two successive updates are summable:

$$\sum_{n\in\mathbb{N}} \|\hat{x}^{(n+1)H} - \hat{x}^{nH}\|^2 \le \frac{\zeta}{1-\zeta}\|\hat{x}^0 - x^\dagger\|^2. \tag{9}$$

Proof. The convergence property and the property (iii) come from the application of the Krasnosel'skii–Mann theorem, see Theorem 5.15 in [9]. The properties (i) and (ii) are applications of Proposition 4.46, Proposition 4.42, and Proposition 4.35 in [9]. □

Remark 2.4. *Analysis of Theorem 2.3-(ii):* $\frac{1-\zeta}{\zeta} = \frac{1-\alpha\lambda}{H\alpha\lambda}$*, and this value multiplies* $\|\hat{x}^{(n+1)H} - \hat{x}^{nH}\|^2$*, which can be up to* H^2 *larger than* $\|\hat{x}^{k+1} - \hat{x}^k\|^2$*, for* k *in* $nH, \dots, (n+1)H-1$*. So, in favorable cases, the 'progress' made in decreasing the squared distance to* $x^\dagger$ *in an epoch of* H *iterations, in* (8)*, is as large as the one we would have if averaging occured after every iteration, which is*

$$\|\hat{x}^{(n+1)H} - x^\dagger\|^2 \le \|\hat{x}^{nH} - x^\dagger\|^2 - \frac{1-\alpha\lambda}{\alpha\lambda} \sum_{k=nH}^{(n+1)H-1} \|\hat{x}^{k+1} - \hat{x}^k\|^2. \tag{10}$$

In less favorable cases, the progress is H times smaller, corresponding to the progress in 1 iteration. Given that communication occurs only once per epoch, the ratio of convergence speed to communication burden is, roughly speaking, between 1 and H times better than the one of the baseline algorithm, which is Algorithm 1 with $H = 1$. They don't converge to the same elements, however.

If the operators $\mathcal{T}_i$ are contractive, we have linear convergence:

Theorem 2.5. *Suppose that* $t_n = nH$*, for every* $n \in \mathbb{N}$*, and suppose that the* $\mathcal{T}_i$ *are all* χ*-contractive, for some* $\chi \in (0,1)$*. Let* $\lambda \in (0, \frac{2}{1+\chi})$ *be the parameter in Algorithm 1. Then the the fixed point* $x^\dagger$ *of* $\widetilde{\mathcal{T}}_\lambda$ *exists and is unique, and the sequence* $(\hat{x}^{nH})_{n\in\mathbb{N}}$ *converges linearly to* $x^\dagger$*. More precisely, the following hold:*
(i) $\widetilde{\mathcal{T}}_\lambda$ *is* ξ^H*-contractive, with*

$$\xi = \max\big(\lambda\chi + (1-\lambda), \lambda(1+\chi) - 1\big). \tag{11}$$

(ii) For every $n \in \mathbb{N}$*,*

$$\|\hat{x}^{(n+1)H} - x^\dagger\| \le \xi^H \|\hat{x}^{nH} - x^\dagger\|. \tag{12}$$

(iii) We have linear convergence with rate ξ*: for every* $n \in \mathbb{N}$*,*

$$\|\hat{x}^{nH} - x^\dagger\| \le \xi^{nH} \|\hat{x}^0 - x^\dagger\|. \tag{13}$$

(iv) For every $n \in \mathbb{N}$*,*

$$\|\hat{x}^{nH} - x^\dagger\| \le \frac{\xi^{nH}}{1-\xi^H} \|\hat{x}^H - \hat{x}^0\|. \tag{14}$$

(v) For every $n \in \mathbb{N}$,

$$\|\hat{x}^{nH} - x^\dagger\| \leq \frac{1}{1-\xi^H}\|\hat{x}^{(n+1)H} - \hat{x}^{nH}\|. \tag{15}$$

Proof. For every $i = 1, \ldots, M$, $\lambda\mathcal{T}_i + (1-\lambda)\mathrm{Id}$ is ξ-contractive, with $\xi = \{\lambda\chi + (1-\lambda)$ if $\lambda \leq 1$, $\lambda(1+\chi) - 1$ else$\}$. Thus, $(\lambda\mathcal{T}_i + (1-\lambda)\mathrm{Id})^H$ is ξ^H contractive. Furthermore, the average of ξ^H-contractive operators is ξ^H-contractive. Then the different properties are applications of the Banach–Picard Theorem 1.50 in [9]. □

Remark 2.6. *Convergence speed in Theorem 2.5. The convergence rate ξ with respect to the number of iterations is the same, whatever H: the distance to a fixed point is contracted by a factor of ξ after every iteration, in average. The fixed point depends on H, however.*

Remark 2.7. *choice of λ. In the conditions of Theorem 2.5, without further knowledge on the operators $\mathcal{T}_i$, we should set $\lambda = 1$, so that $\xi = \chi$, since every other choice may slow down the convergence.*

To provide further insights on the convergence speed and the relationship between the fixed points $x^\star$ and $x^\dagger$, we also perform a non-asymptotic ergodic analysis. We do not assume uniform t_n any more, we simply assume that Assumption (2.1) holds.

We introduce the following shorthand notations:

$$V_k = \frac{1}{M}\sum_{i=1}^{M}\left\|x_i^k - \hat{x}^k\right\|^2, \tag{16}$$

$$g_i(x_i^k) = x_i^k - \mathcal{T}_i(x_i^k), \tag{17}$$

$$\hat{g}^k = \hat{x}^k - \frac{1}{M}\sum_{i=1}^{M}\mathcal{T}_i(x_i^k), \tag{18}$$

$$\sigma^2 = \frac{1}{M}\sum_{i=1}^{M}\|g_i(x^\star)\|^2. \tag{19}$$

The first value measures the deviation of the iterates from their average. This value is crucial for the convergence analysis. The second and third values can be viewed as analogues of the gradient and the average gradient in our more general setting. The last value serves as a measure of variance adapted to methods with local steps.

Assumption 2.8. *Each operator $\mathcal{T}_i$ is firmly nonexpansive. Equivalently, for every x and $y \in \mathbb{R}^d$,*

$$\|\mathcal{T}_i(x) - \mathcal{T}_i(y)\|^2 \leq \|x-y\|^2 - \|\mathcal{T}_i(x) - x - \mathcal{T}_i(y) + y\|^2. \tag{20}$$

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

Lemma 2.9. *Under Assumption 2.8 and the condition $0 \leq \lambda \leq 1$ we have*

$$\begin{aligned}\|\hat{x}^{k+1} - x^\star\|^2 \leq{}& \|\hat{x}^k - x^\star\|^2 + \lambda(2-\lambda)V_k \\ &- \frac{1}{2}\lambda(1-\lambda)\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(\hat{x}^k) - g_i(x^\star)\right\|^2.\end{aligned} \tag{21}$$

Lemma 2.10. *Under Assumption 2.8 and the condition* $0 \le \lambda \le 1$ *we have*

$$V_k \le \lambda^2(H-1)\sum_{j=k_p}^{k}\frac{3}{M}\sum_{i=1}^{M}\|x_i^j - \hat{x}^j\|^2 \tag{22}$$
$$+\sum_{j=k_p}^{k}\frac{2}{M}\sum_{i=1}^{M}\|g_i(\hat{x}^j) - g_i(x^\star)\|^2 + 6\sum_{j=k_p}^{k}\sigma^2.$$

Theorem 2.11. *Suppose that* $\lambda \le \frac{1}{8\max(1,H-1)}$ *and that Assumption 2.8 holds. Then, for every* $k \in \mathbb{N}$,

$$\frac{1}{T}\sum_{k=0}^{T-1}\left\|\hat{x}^k - \mathcal{T}(\hat{x}^k)\right\|^2 \le \frac{3\|\hat{x}^0 - x^\star\|^2}{\lambda T} + 36\lambda^2(H-1)^2\sigma^2. \tag{23}$$

We would like to know how many iterations we need to achieve ε-accuracy for Algorithm 1. The next result gives us an explicit complexity.

Corollary 2.12. *Suppose that* $\lambda \le \frac{1}{8\max(1,H-1)}$ *and that Assumption 2.8 holds. Then a sufficient condition on the number* T *of iterations to reach* ε*-accuracy, for any* $\varepsilon > 0$, *is*

$$\frac{T}{H-1} \ge \frac{24\|\hat{x}^0 - x^*\|^2}{\varepsilon}\max\left\{2, \frac{3\sigma}{\sqrt{2\varepsilon}}\right\}. \tag{24}$$

Note that as long as the target accuracy is not too high, in particular $\varepsilon \ge \frac{9\sigma^2}{8}$, then

$$\frac{T}{H} = \mathcal{O}\left(\frac{\|\hat{x}^0 - x^\star\|^2}{\varepsilon}\right). \tag{25}$$

If $\varepsilon < \frac{9}{8}\sigma^2$, then the communication complexity is equal to

$$\frac{T}{H} = \mathcal{O}\left(\frac{\|\hat{x}^0 - x^\star\|^2\sigma}{\epsilon^{3/2}}\right). \tag{26}$$

Corollary 2.13. *Let* $T \in \mathbb{N}$ *and let* $H \ge 1$ *be such that* $H \le \frac{\sqrt{T}}{\sqrt{M}}$; *then* $\lambda = \frac{1}{8}\frac{\sqrt{M}}{\sqrt{T}}$ *and using the result of the previous theorem we get*

$$\frac{1}{T}\sum_{k=0}^{T-1}\left\|\hat{x}^k - \mathcal{T}(\hat{x}^k)\right\|^2 \le \frac{24\|\hat{x}^0 - x^\star\|^2}{\sqrt{MT}} + \frac{3M(H-1)^2\sigma^2}{8T}. \tag{27}$$

Hence, to get a convergence rate of $\frac{1}{\sqrt{MT}}$ we can choose the parameter H as $H = \mathcal{O}\left(T^{1/4}M^{-3/4}\right)$, which implies a total number of $\Omega\left(T^{3/4}M^{3/4}\right)$ synchronization steps. If we need a rate of $1/\sqrt{T}$, we can set a larger value $H = \mathcal{O}\left(T^{1/4}\right)$.

This result covers [4] in case of Local gradient descent.

Remark 2.14. *Case* $H = 1$. *We remark that if* $H = 1$, *i.e. communication occurs after every iteration, the last term in Theorem 2.11, which depends on* $H - 1$, *is zero. This is coherent with the fact that* $x^\dagger = x^\star$ *in that case, so that the algorithm converges to an exact fixed point of* $\mathcal{T}$. *In that sense, Theorem 3 is tight.*

Remark 2.15. *Local gradient descent (GD) case. Consider that* $\mathcal{T}_i(x_i^k) = x_i^k - \frac{1}{L}\nabla f_i(x_i^k)$, *where each convex function* f_i *is L-smooth. Then the assumptions in Theorem 2.11 are satisfied and we recover known results about local GD methods for heterogeneous data as particular cases.*

Algorithm 2 Randomized distributed fixed point method

Input: Initial estimate $\hat{x}^0 \in \mathbb{R}^d$, stepsize $\lambda > 0$, communication probability $0 < p \leq 1$
Initialize: $x_i^0 = \hat{x}^0$, for all $i = 1, \ldots, M$
for $k = 1, 2, \ldots$ **do**
  for $i = 1, 2, \ldots, M$ in parallel **do**
    $h_i^{k+1} := (1-\lambda)x_i^k + \lambda \mathcal{T}_i(x_i^k)$
  end for
  Flip a coin and
  with probability p **do**
    Communicate h_i^{k+1} to master, for $i = 1, \ldots, M$
    At master node: $\hat{x}^{k+1} := \frac{1}{M}\sum_{i=1}^{M} h_i^{k+1}$
    Broadcast: $x_i^{k+1} := \hat{x}^{k+1}$, for all $i = 1, \ldots, M$
  else, with probability $1-p$, **do**
    $x_i^{k+1} := h_i^{k+1}$, for all $i = 1, \ldots, M$
end for

3 Randomized communication-efficient distributed fixed point method

Now, we propose a second loopless algorithm, where the local steps in Algorithm 1, which can be viewed as an inner loop between two communication steps, is replaced by a probabilistic aggregation. This yields Algorithm 2, shown above. To analyse Algorithm 2, we suppose that $\mathcal{T}$ is contractive. So, we suppose that the following holds.

Assumption 3.1. *Each operator $\mathcal{T}_i$ is contractive and firmly nonexpansive*[1] *with parameter $\rho > 0$. More precisely, there exists $\rho > 0$ such that, for every $i = 1, \ldots, M$ and every $x, y \in \mathbb{R}^d$,*

$$(1+\rho)\|\mathcal{T}_i(x) - \mathcal{T}_i(y)\|^2 \leq \|x-y\|^2 - \|x - \mathcal{T}_i(x) - y + \mathcal{T}_i(y)\|^2. \tag{28}$$

In the case of gradient descent (GD) as the operator, this assumption is satisfied for strongly convex smooth functions.

The next lemma provides a recurrence property on the optimality gap, for one iteration of Algorithm 2.

Lemma 3.2. *Under Assumption 3.1 with $\rho > 0$, for every $k \in \mathbb{N}$,*

$$\|\hat{x}^{k+1} - x^\star\|^2 \leq \left(1 - \frac{\lambda\rho}{1+\rho}\right)\|\hat{x}^k - x^\star\|^2 + \frac{5}{2}\lambda V_k - \frac{1}{2}\lambda\left(\frac{1}{2} - \lambda\right)\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(x_i^k) - g_i(x^\star)\right\|^2. \tag{29}$$

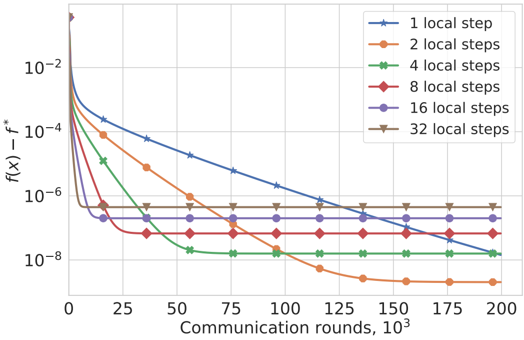
We now bound the variance V_k for one iteration, using the contraction property.

Lemma 3.3. *Under Assumption 3.1 and if $\lambda < \frac{p}{15}$ we have, for every $k \in \mathbb{N}$,*

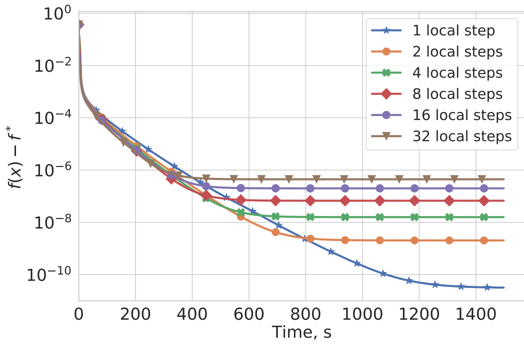
$$V_k \leq \frac{2}{p}\left(1 - \frac{p}{4} + \frac{5}{p}\lambda^2\right)V_k + 60\frac{\lambda^2}{p^2}\sigma^2 - \frac{2}{p}\mathbb{E}[V_{k+1}] + 20\frac{\lambda^2}{p^2}\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(\hat{x}^k) - g_i(x^\star)\right\|^2. \tag{30}$$

Combining the previous two lemmas, almost sure linear convergence of Algorithm 2 up to a neighborhood is established in the next theorem:

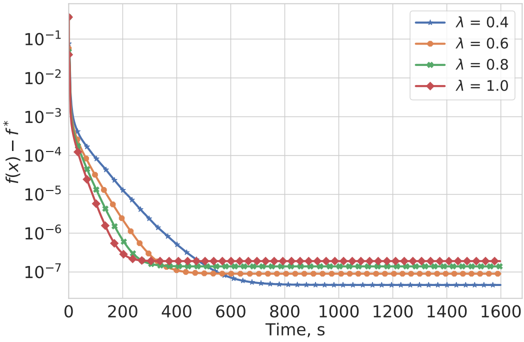
[1] In operator theory this property is known as $(1+\frac{\rho}{2})$-cocoercivity of operator.



(a)



(b)



(c)

Figure 1: We analyze the convergence of Algorithm 1 with gradient steps and uniform number H of local steps; that is, $t_n = nH$; in (a) with respect to the number of communication rounds, for different values of H, with $\lambda = 0.5$; in (b) with respect to computation time, for different values of H, with $\lambda = 0.5$; in (c) with respect to computation time, for different values of λ, with $H = 4$.

Theorem 3.4. *Let Ψ^k be the Lyapunov function, for every $k \in \mathbb{N}$,*

$$\Psi^k := \|\hat{x}^k - x^\star\|^2 + \frac{5\lambda}{p} V_k. \tag{31}$$

Under Assumption 3.1 and if $\lambda < \frac{p}{15}$, we have, for every $k \in \mathbb{N}$,

$$\mathbb{E}\Psi^k \leq \left(1 - \min\left(\frac{\lambda\rho}{1+\rho}, \frac{p}{5}\right)\right)^k \Psi^0 + \frac{150}{\min\left(\frac{\lambda\rho}{1+\rho}, \frac{p}{5}\right) p^2} \lambda^3 \sigma^2. \tag{32}$$

Since the previous theorem may be difficult to analyze, the next results gives a bound to reach ε-accuracy in Algorithm 2:

Corollary 3.5. *Under Assumption 3.1 and if $\lambda < \frac{p}{15}$, for any $\varepsilon > 0$, ε-accuracy is reached after T iterations, with*

$$T \geq \max\left\{\frac{15(1+\rho)}{\rho p}, \frac{18\sigma(1+\rho)^{\frac{1}{3}}}{p\rho^{\frac{3}{2}}\varepsilon^{\frac{1}{2}}}, \frac{40\sigma^{\frac{2}{3}}(1+\rho)}{p\rho\varepsilon^{\frac{1}{3}}}\right\} \log\frac{2\Psi_0}{\varepsilon}. \tag{33}$$

4 Experiments

4.1 Experimental setup

4.1.1 Model

Despite the fact that our approach can be applied more broadly, we focus on logistic regression, since this is one of the most important models for classification. The corresponding objective function takes the following form:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\left(-b_i a_i^\top x\right)\right) + \frac{\lambda}{2}\|x\|^2, \tag{34}$$

where $a_i \in \mathbb{R}^d$ and $b_i \in \{-1, +1\}$ are the data samples.

4.1.2 Datasets

We use the 'a9a' and 'a4a' datasets from the LIBSVM library and we set λ to be $\frac{L}{n}$, where n is the size of the dataset and L is a Lipschitz constant of ∇f.

4.1.3 Hardware and software

We implemented all algorithms in Python using the package MPI4PY in order to run the code on a truly parallel architecture. All methods were evaluated on a computer with an Intel(R) Xeon(R) Gold 6146 CPU at 3.20GHz, having 24 cores. The cores are connected to 2 sockets, with 12 cores for each of them.

4.2 Local gradient descent

We consider gradient descent (GD) steps as the operators. That is, we consider the problem of minimizing the finite sum:

$$f(x) = \frac{1}{M}\sum_{i=1}^{M} f_i(x), \tag{35}$$

where each function f_i is convex and L-smooth. We set

$$\mathcal{T}_i(x_i^k) \coloneqq x_i^k - \frac{1}{L}\nabla f_i(x_i^k). \tag{36}$$

We use $\frac{1}{L}$ as the stepsize, so that each $\mathcal{T}_i$ is firmly nonexpansive.

4.3 Local cycling GD

In this section we consider another operator, which is cycling GD. So, we consider minimizing the same function as in (35), but this time each function f_i is also a finite sum:

$$f_i = \frac{1}{N}\sum_{j=1}^{N} f_{ij}. \tag{37}$$

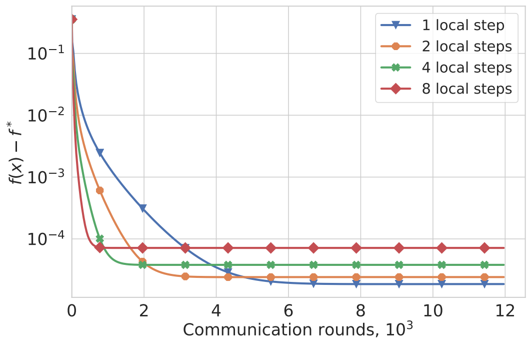
Instead of applying full gradient steps, we apply N element-wise gradient steps, in the sequential order of the data points. Thus,

$$\mathcal{T}_i(x_i^k) \coloneqq S_{i1}(S_{i2}(\dots S_{in}(x_i^k))), \tag{38}$$

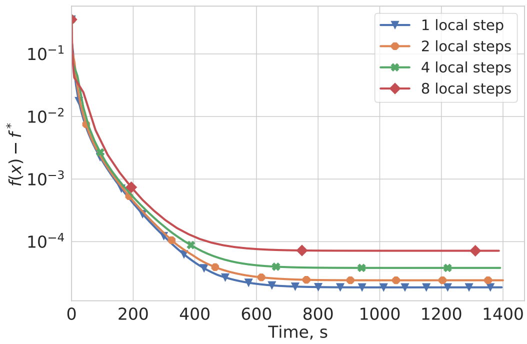
where

$$S_{ij} : y \mapsto y - \frac{1}{NL}\nabla f_{ij}. \tag{39}$$

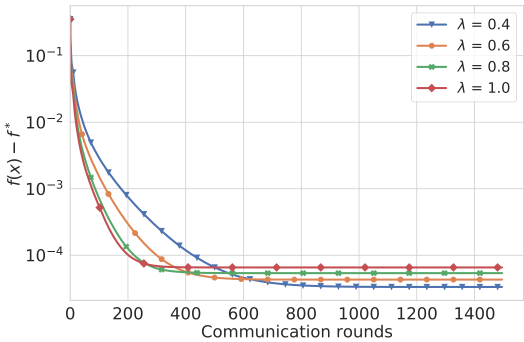
If, for each i, all functions f_{ij} have the same minimizer $x_i^\star$, then this joint minimizer is a fixed point of $\mathcal{T}_i$. Also, these operators can be shown to be firmly nonexpansive.



(a)



(b)



(c)

Figure 2: We analyze the convergence of Algorithm 1 with cyclic gradient steps and uniform number H of local steps; that is, $t_n = nH$; in (a) with respect to the number of communication rounds, for different values of H, with $\lambda = 0.5$; in (b) with respect to computation time, for different values of H, with $\lambda = 0.5$; in (c) with respect to computation time, for different values of λ, with $H = 4$.

4.4 Results

We observe a very tight match between our theory and the numerical results. As can be seen, the larger the value of the parameters H and λ, the faster the convergence at the beginning, but the larger the radius of the neighborhood. In terms of computational time, there is no big advantage, since the experiments were run on a single machine and the communication time was negligible. But in a distributed setting where communication is slow, our approach has a clear advantage. We can also observe the absence of oscillations. Hence, there is a clear advantage of local methods when only limited accuracy is required.

In the experiment with cyclic GD, the algorithm converges only to a neighbourhood of the ideal solution, even when 1 local step is used. This happens because the assumption of a joint minimizer for all i is not satisfied here. However, since the operators are firmly nonexpansive, we have convergence to a fixed point. The convergence of Algorithm 1 is illustrated with respect to the relaxation parameter λ. If λ is small, convergence is slower, but the algorithm converges to a point closer to the true solution $x^\star$. In Figure 4, we further illustrate the behavior of Algorithm 2 with respect to the probability p, for cyclic gradient descent. We can see that the fastest and most accurate convergence is obtained for an intermediate value of p, here $p = 0.2$.

The experiments with Algorithm 2 show that, with a low probability p of update, the neighborhood is substantially larger; however, with p increasing, the convergence in terms of communication rounds becomes worse. Therefore, with careful selection of the probability parameter, a significant advantage can be obtained.

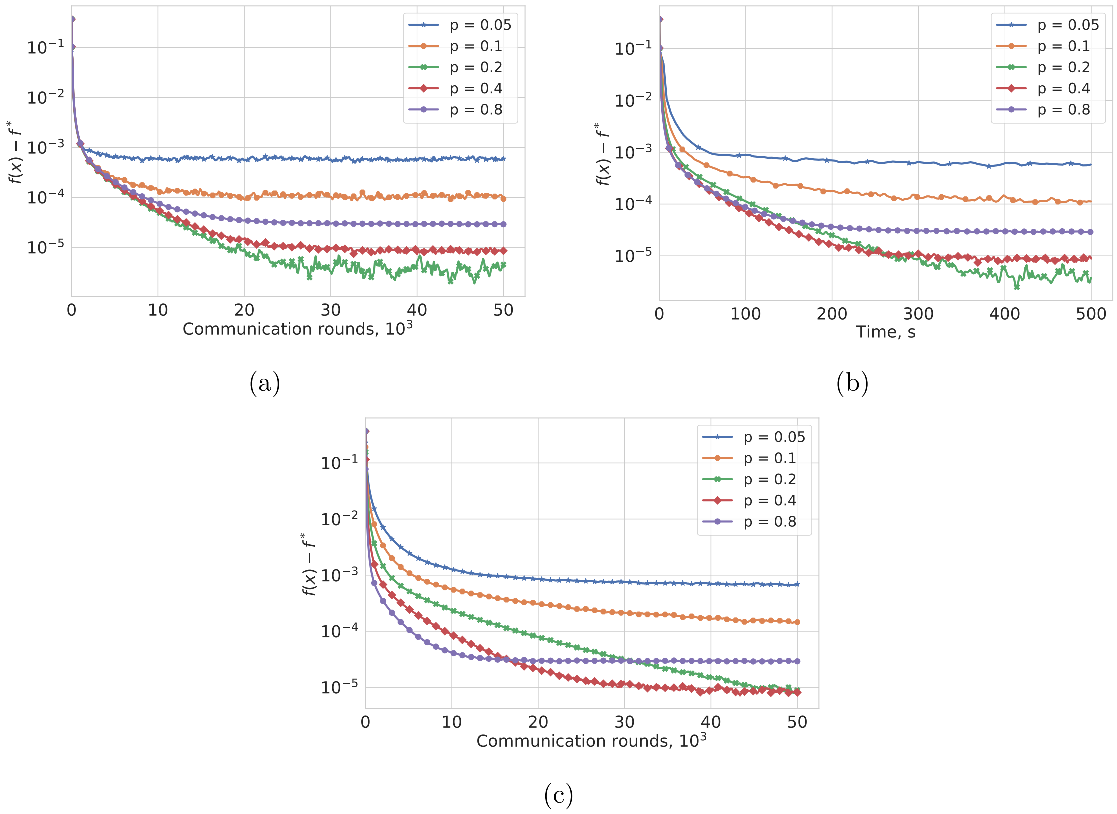


Figure 3: We analyze the convergence of Algorithm 2, with $\lambda = 0.5$; in (a) with the same gradient stepsizes, with respect to the number of communication rounds, for different values of the probability p; in (b) with the same gradient stepsizes, with respect to the computation time, for different values of the p; in (c) with gradient stepsizes proportional to p, with respect to the number of communication rounds, for different values of p.

5 Conclusion

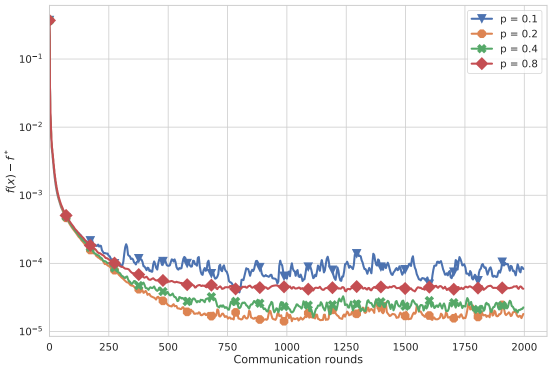
In this work, the first of its kind, we have proposed two strategies to reduce the communication burden in a generic distributed setting, where a fixed point of an average of operators is sought. We have shown that they improve the convergence speed, while achieving the goal of reducing the communication load. At convergence, only an approximation of the ideal fixed point is attained, but if a medium-precision solution is sufficient and the process is stopped before convergence, the proposed algorithms are particularly adequate.

In future work, we will generalize the setting to randomized fixed point operators, to generalize stochastic gradient descent approaches. We will also investigate compression [10, 11] of the communicated variables, with or without variance reduction, in combination with locality.

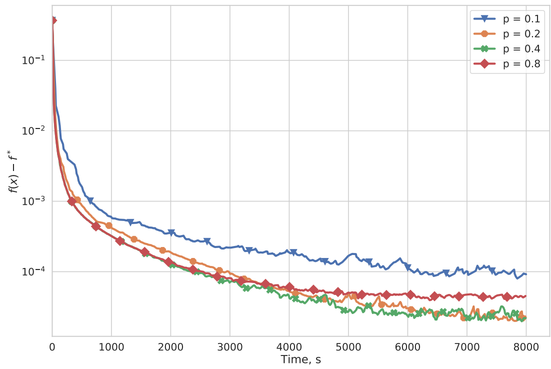
References

[1] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *NIPS Workshop on Private Multi-Party Machine Learning*, 2016. [Online]. Available: https://arxiv.org/abs/1610.05492

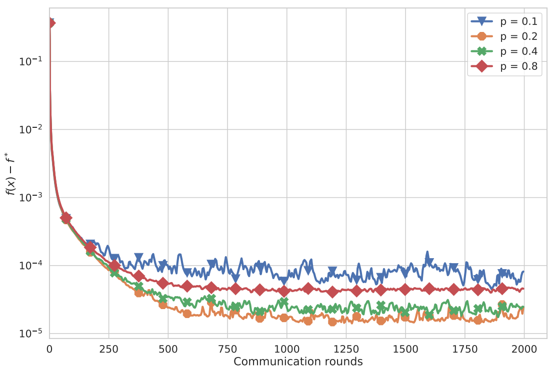
[2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson *et al.*, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.



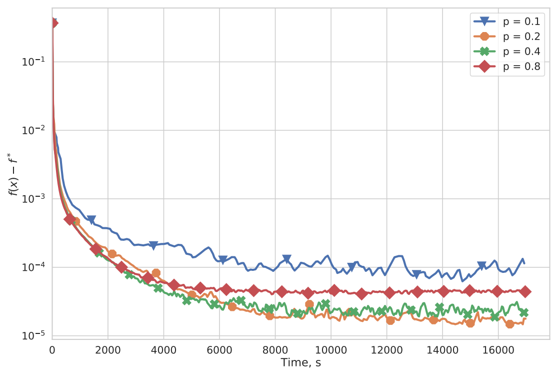
(a)



(b)



(c)



(d)

Figure 4: We analyze the convergence of Algorithm 2 for cyclic gradient descent, with $\lambda = 0.5$; in (a) with the same gradient stepsizes, with respect to the number of communication rounds, for different values of the probability p; in (b) with the same gradient stepsizes, with respect to the computation time, for different values of the p; in (c) with gradient stepsizes proportional to p, with respect to the number of communication rounds, for different values of p; in (d) with gradient stepsizes proportional to p, with respect to the computation time, for different values of p.

[3] S. U. Stich, “Local SGD Converges Fast and Communicates Little,” in *International Conference on Learning Representations*, 2019.

[4] A. Khaled, K. Mishchenko, and P. Richtárik, “First analysis of local GD on heterogeneous data,” in *NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality*, 2019, pp. 1–11.

[5] ——, “Tighter theory for local SGD on identical and heterogeneous data,” in *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, 2020.

[6] C. Ma, J. Konečnỳ, M. Jaggi, V. Smith, M. I. Jordan, P. Richtárik, and M. Takáč, “Distributed optimization with arbitrary local solvers,” *Optimization Methods and Software*, vol. 32, no. 4, pp. 813–848, 2017.

[7] F. Haddadpour and M. Mahdavi, “On the convergence of local descent methods in federated learning,” *arXiv preprint arXiv:1910.14425*, 2019.

[8] H. Bauschke, R. Burachik, P. Combettes, V. Elser, D. Luke, and H. Wolkowicz, Eds., *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer, 2011.

[9] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed. New York: Springer, 2017.

[10] A. Khaled and P. Richtárik, "Gradient descent with compressed iterates," in *NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality*, 2019.

[11] S. Chraibi, A. Khaled, D. Kovalev, P. Richtárik, A. Salim, and M. Takáč, "Distributed fixed point methods with compressed iterates," *ArXiv*, vol. abs/1912.09925, 2019.

Appendix

A Notations and Basic Facts **14**
A.1 Notations . . . 14
A.2 Basic Facts . . . 14
A.3 Technical lemmas . . . 15

B Analysis of Algorithm 1 **16**
B.1 Proof of Lemma 2.9 . . . 16
B.2 Proof of Lemma 2.10 . . . 18
B.3 Proof of Theorem 2.11 . . . 20
B.4 Proof of Corollary 2.12 . . . 22

C Analysis of Algorithm 2 **22**
C.1 Proof of Lemma 3.2 . . . 22
C.2 Proof of Lemma 3.3 . . . 25
C.3 Proof of Theorem 3.4 . . . 26
C.4 Proof of Corollary 3.5 . . . 27

A Notations and Basic Facts

A.1 Notations

Let $\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_n$ be operators on $\mathbb{R}^d$.

Let us summarize here the notations used throughout the paper:

$$\begin{aligned}
\mathcal{T}(x) = \frac{1}{M}\sum_{i=1}^{M}\mathcal{T}_i(x) - &\ \textbf{averaging operator},\\
x^\star = \mathcal{T}(x^\star) - &\ \textbf{fixed point},\\
\hat{x}^k = \frac{1}{M}\sum_{i=1}^{M}x_i^k - &\ \textbf{mean point},\\
\sigma^2 = \frac{1}{M}\sum_{i=1}^{M}\|g_i(x^\star)\|^2 - &\ \textbf{variance for locality},\\
V_k = \frac{1}{M}\sum_{i=1}^{M}\left\|x_i^k - \hat{x}^k\right\|^2 - &\ \textbf{deviation from average},\\
g_i(x_i^k) = x_i^k - \mathcal{T}_i(x_i^k) - &\ \textbf{local residual},\\
\hat{g}^k = \hat{x}^k - \frac{1}{M}\sum_{i=1}^{M}\mathcal{T}_i(x_i^k) - &\ \textbf{residual for mean point},\\
\rho - &\ \textbf{contraction parameter},\\
\lambda - &\ \textbf{relaxation parameter},\\
H - &\ \textbf{bound for the number of local steps in Alg. 1},\\
p - &\ \textbf{probability of communication in Alg. 2}.
\end{aligned}$$

A.2 Basic Facts

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

$$f\left(\frac{1}{M}\sum_{m=1}^{M}x^m\right) \le \frac{1}{M}\sum_{m=1}^{M}f\left(x^m\right). \tag{40}$$

In particular, with $f(x) = \|x\|^2$, we obtain

$$\left\|\frac{1}{M}\sum_{m=1}^{M}x_m\right\|^2 \le \frac{1}{M}\sum_{m=1}^{M}\|x_m\|^2. \tag{41}$$

Facts from linear algebra.
We will use the following important properties:

$$\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2, \text{ for every } x, y \in \mathbb{R}^d, \tag{42}$$

$$\|x+y\|^2 \ge \frac{1}{2}\|y\|^2 - \|x\|^2, \text{ for every } x, y \in \mathbb{R}^d, \tag{43}$$

$$2\langle a, b\rangle \le \zeta\|a\|^2 + \zeta^{-1}\|b\|^2 \text{ for all } a, b \in \mathbb{R}^d \text{ and } \zeta > 0, \tag{44}$$

$$\frac{1}{M}\sum_{m=1}^{M}\|X_m\|^2 = \frac{1}{M}\sum_{m=1}^{M}\left\|X_m - \frac{1}{M}\sum_{i=1}^{M}X_i\right\|^2 + \left\|\frac{1}{M}\sum_{m=1}^{M}X_m\right\|^2. \tag{45}$$

A.3 Technical lemmas

Technical Lemma 1. If $\mathcal{T}$ is firmly nonexpansive, then

$$\langle \mathcal{T}(x) - x - \mathcal{T}(y) + y, x - y \rangle \leq -\|x - \mathcal{T}(x) - y + \mathcal{T}(y)\|^2 \tag{46}$$

and

$$\|\mathcal{T}(x) - \mathcal{T}(y)\|^2 \leq \|x - y\|^2. \tag{47}$$

Proof.

$$\begin{aligned}
\|\mathcal{T}(x) - \mathcal{T}(y)\|^2 &= \|x - y\|^2 + 2\langle \mathcal{T}(x) - x - \mathcal{T}(y) + y, x - y \rangle + \|\mathcal{T}(x) - x - \mathcal{T}(y) + y\|^2 \\
&\leq \|x - y\|^2 - \|\mathcal{T}(x) - x - \mathcal{T}(y) + y\|^2 \\
&= \|x - y\|^2 - \|x - \mathcal{T}(x) - y + \mathcal{T}(y)\|^2.
\end{aligned}$$

We have

$$\begin{aligned}
\|x-y\|^2 + 2\langle \mathcal{T}(x) - x - \mathcal{T}(y) + y, x - y \rangle + \|\mathcal{T}(x) - x - \mathcal{T}(y) + y\|^2 &\leq \|x - y\|^2 - \|x - \mathcal{T}(x) - y + \mathcal{T}(y)\|^2 \\
2\langle \mathcal{T}(x) - x - \mathcal{T}(y) + y, x - y \rangle + \|\mathcal{T}(x) - x - \mathcal{T}(y) + y\|^2 &\leq -\|x - \mathcal{T}(x) - y + \mathcal{T}(y)\|^2 \\
2\langle \mathcal{T}(x) - x - \mathcal{T}(y) + y, x - y \rangle &\leq -2\|x - \mathcal{T}(x) - y + \mathcal{T}(y)\|^2 \\
\langle \mathcal{T}(x) - x - \mathcal{T}(y) + y, x - y \rangle &\leq -\|x - \mathcal{T}(x) - y + \mathcal{T}(y)\|^2.
\end{aligned}$$

The last inequality follows from the definition. □

Technical Lemma 2. Let $\rho > 0$. Let $\mathcal{T}$ be a contractive and firmly nonexpansive operator; that is, for every $x, y \in \mathbb{R}^d$,

$$(1 + \rho)\|\mathcal{T}(x) - \mathcal{T}(y)\|^2 \leq \|x - y\|^2 - \|x - \mathcal{T}(x) - y + \mathcal{T}(y)\|^2. \tag{48}$$

Then

$$\langle x - y, \mathcal{T}_i(x) - x + y - \mathcal{T}_i(y) \rangle \leq -\left(\frac{\rho}{2(1 + \rho)} \|x - y\|^2 + \frac{2 + \rho}{2(1 + \rho)} \|x - \mathcal{T}_i(x) - y + \mathcal{T}_i(y)\|^2 \right). \tag{49}$$

Proof.

$$\begin{aligned}
\|\mathcal{T}_i(x) - \mathcal{T}_i(y)\|^2 &= \|x - y - x + \mathcal{T}_i(x) + y - \mathcal{T}_i(y)\|^2 \\
&= \|x - y\|^2 - 2\langle x - y, x - \mathcal{T}_i(x) - y + \mathcal{T}_i(y) \rangle + \|x - \mathcal{T}_i(x) - y + \mathcal{T}_i(y)\|^2.
\end{aligned}$$

We have

$$(1+\rho)\|\mathcal{T}_i(x) - \mathcal{T}_i(y)\|^2 = (1+\rho)\|x-y\|^2 + (1+\rho)\|x - \mathcal{T}_i(x) - y + \mathcal{T}_i(y)\|^2 - 2(1+\rho)\langle x - y, x - \mathcal{T}_i(x) - y + \mathcal{T}_i(y) \rangle.$$

Since

$$\begin{aligned}
&(1 + \rho)\|x - y\|^2 + (1 + \rho)\|x - \mathcal{T}_i(x) - y + \mathcal{T}_i(y)\|^2 \\
&\qquad \leq \|x - y\|^2 - \|x - \mathcal{T}_i(x) - y + \mathcal{T}_i(y)\|^2 + 2(1 + \rho)\langle x - y, x - \mathcal{T}_i(x) - y + \mathcal{T}_i(y) \rangle,
\end{aligned}$$

we have

$$\begin{aligned}
2(1 + \rho)\langle x - y, x - \mathcal{T}_i(x) - y + \mathcal{T}_i(y) \rangle &\geq \rho\|x - y\|^2 + (2 + \rho)\|x - \mathcal{T}_i(x) - y + \mathcal{T}_i(y)\|^2 \\
\langle x - y, \mathcal{T}_i(x) - x + y - \mathcal{T}_i(y) \rangle &\leq -\left(\frac{\rho}{2(1 + \rho)} \|x - y\|^2 + \frac{2 + \rho}{2(1 + \rho)} \|x - \mathcal{T}_i(x) - y + \mathcal{T}_i(y)\|^2 \right).
\end{aligned}$$

□

B Analysis of Algorithm 1

B.1 Proof of Lemma 2.9

Under Assumption 2.8 and under the condition $0 \le \lambda \le 1$, we have

$$\|\hat{x}^{k+1} - x^\star\|^2 \le \|\hat{x}^k - x^\star\|^2 + \lambda(2-\lambda)V_k - \frac{1}{2}\lambda(1-\lambda)\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(\hat{x}^k) - g_i(x^\star)\right\|^2. \quad (50)$$

Proof.

$$\begin{aligned}
\|\hat{x}^{k+1} - x^\star\|^2 &= \|\hat{x}^{k+1} - \hat{x}^k + \hat{x}^k - x^\star\|^2 \\
&= \|\hat{x}^k - x^\star\|^2 + 2\langle \hat{x}^{k+1} - \hat{x}^k, \hat{x}^k - x^\star\rangle + \|\hat{x}^{k+1} - \hat{x}^k\|^2 \\
&= \|\hat{x}^k - x^\star\|^2 + 2\langle (1-\lambda)\hat{x}^k + \lambda\frac{1}{M}\sum_{i=1}^{M}\mathcal{T}_i(x_i^k) - \hat{x}^k, \hat{x}^k - x^\star\rangle \\
&\quad + \|(1-\lambda)\hat{x}^k + \lambda\frac{1}{M}\sum_{i=1}^{M}\mathcal{T}_i(x_i^k) - \hat{x}^k\|^2 \\
&= \|\hat{x}^k - x^\star\|^2 + 2\lambda\langle\frac{1}{M}\sum_{i=1}^{M}\left(\mathcal{T}_i(x_i^k) - \hat{x}^k\right), \hat{x}^k - x^\star\rangle \\
&\quad + \lambda^2\|\frac{1}{M}\sum_{i=1}^{M}\left(\mathcal{T}_i(x_i^k) - \hat{x}^k\right)\|^2 \\
&= \|\hat{x}^k - x^\star\|^2 + 2\lambda\frac{1}{M}\sum_{i=1}^{M}\langle\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star, \hat{x}^k - x^\star\rangle \\
&\quad + \lambda^2\|\frac{1}{M}\sum_{i=1}^{M}\left(\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\right)\|^2 \\
&= 2\lambda\frac{1}{M}\sum_{i=1}^{M}\left[\langle\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star, x_i^k - x^\star\rangle\right. \\
&\quad \left.+ \langle\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star, \hat{x}^k - x_i^k\rangle\right] \\
&\quad + \|\hat{x}^k - x^\star\|^2 + \lambda^2\|\frac{1}{M}\sum_{i=1}^{M}\left(\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\right)\|^2 \\
&= \|\hat{x}^k - x^\star\|^2 + 2\lambda\frac{1}{M}\sum_{i=1}^{M}\langle\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star, x_i^k - x^\star\rangle \\
&\quad + 2\lambda\frac{1}{M}\sum_{i=1}^{M}\langle\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star, \hat{x}^k - x_i^k\rangle \\
&\quad + \lambda^2\left\|\frac{1}{M}\sum_{i=1}^{M}\left(\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\right)\right\|^2.
\end{aligned}$$

Using Technical Lemma 1,

$$\|\hat{x}^{k+1} - x^\star\|^2 \le \|\hat{x}^k - x^\star\|^2 - 2\lambda \frac{1}{M}\sum_{i=1}^{M} \|\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\|^2$$
$$+ 2\lambda \frac{1}{M}\sum_{i=1}^{M} \langle \mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star, \hat{x}^k - x_i^k\rangle$$
$$+ \lambda^2 \frac{1}{M} \| \sum_{i=1}^{M} \left(\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\right) \|^2$$

Using the inequality (44)

$$\begin{aligned}
\|\hat{x}^{k+1} - x^\star\|^2 &\le \|\hat{x}^k - x^\star\|^2 - 2\lambda \frac{1}{M}\sum_{i=1}^{M} \|\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\|^2 \\
&\quad + \lambda^2 \frac{1}{M}\sum_{i=1}^{M} \|\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\|^2 \\
&\quad + 2\lambda \frac{1}{M}\sum_{i=1}^{M} \left(\frac{1}{2}\|\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\|^2 + \frac{1}{2}\|\hat{x}^k - x_i^k\|^2\right) \\
&= \|\hat{x}^k - x^\star\|^2 - \lambda(1-\lambda)\frac{1}{M}\sum_{i=1}^{M} \|\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\|^2 + \lambda \frac{1}{M}\sum_{i=1}^{M}\|\hat{x}^k - x_i^k\|^2 \\
&= \|\hat{x}^k - x^\star\|^2 - \lambda(1-\lambda)\frac{1}{M}\sum_{i=1}^{M} \|\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\|^2 + \lambda V_k.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\|\hat{x}^{k+1} - x^\star\|^2 \le \|\hat{x}^k - x^\star\|^2 + \lambda V_k \\
&\quad - \lambda(1-\lambda)\frac{1}{M}\sum_{i=1}^{M} \left\|\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star + \mathcal{T}_i(\hat{x}^k) - \mathcal{T}_i(\hat{x}^k) + \hat{x}^k - \hat{x}^k\right\|^2 \\
&= \|\hat{x}^k - x^\star\|^2 + \lambda V_k \\
&\quad - \lambda(1-\lambda)\frac{1}{M}\sum_{i=1}^{M} \left\| \left(\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(\hat{x}^k) + \hat{x}^k\right) + \left(\mathcal{T}_i(\hat{x}^k) - \hat{x}^k - \mathcal{T}_i(x^\star) + x^\star\right) \right\|^2.
\end{aligned}$$

Using the inequality (43)

$$\begin{aligned}
\|\hat{x}^{k+1} - x^\star\|^2 &\le \|\hat{x}^k - x^\star\|^2 - \frac{1}{2}\lambda(1-\lambda)\frac{1}{M}\sum_{i=1}^{M} \left\|\mathcal{T}_i(\hat{x}^k) - \hat{x}^k - \mathcal{T}_i(x^\star) + x^\star\right\|^2 \\
&\quad + (1-\lambda)\lambda\frac{1}{M}\sum_{i=1}^{M} \left\|\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(\hat{x}^k) + \hat{x}^k\right\|^2 + \lambda V_k \\
&\le \|\hat{x}^k - x^\star\|^2 - \frac{1}{2}\lambda(1-\lambda)\frac{1}{M}\sum_{i=1}^{M} \left\|\mathcal{T}_i(\hat{x}^k) - \hat{x}^k - \mathcal{T}_i(x^\star) + x^\star\right\|^2 + \lambda(2-\lambda)V_k \\
&= \|\hat{x}^k - x^\star\|^2 - \frac{1}{2}\lambda(1-\lambda)\frac{1}{M}\sum_{i=1}^{M} \left\|\hat{x}^k - \mathcal{T}_i(\hat{x}^k) + \mathcal{T}_i(x^\star) - x^\star\right\|^2 + \lambda(2-\lambda)V_k.
\end{aligned}$$

□

B.2 Proof of Lemma 2.10

In this section, we prove the following extended version of Lemma 2.10: Under Assumption 2.8 and under the condition $0 \le \lambda \le 1$, we have

$$V_k \le \lambda^2 (H-1) \sum_{j=k_p}^{k} \frac{3}{M} \sum_{i=1}^{M} \|x_i^j - \hat{x}^j\|^2 + \sum_{j=k_p}^{k} \frac{2}{M} \sum_{i=1}^{M} \|g_i(\hat{x}^j) - g_i(x^\star)\|^2 + 6 \sum_{j=k_p}^{k} \sigma^2. \tag{51}$$

Moreover, for $\lambda \le \frac{1}{8\max(1,H-1)}$, we have

$$\sum_{k=k_p}^{k_{p+1}-1} \left(-\frac{1}{2}\lambda(1-\lambda)\frac{1}{M}\sum_{i=1}^{M} \|\hat{x}^k - \mathcal{T}_i(\hat{x}^k) + \mathcal{T}_i(x^\star) - x^\star\|^2 + \lambda(2-\lambda)V_k \right) \le -\frac{\lambda}{3} \sum_{k=k_p}^{k_{p+1}-1} \frac{1}{M} \sum_{i=1}^{M} \|\hat{x}^k - \mathcal{T}_i(\hat{x}^k) + \mathcal{T}_i(x^\star) - x^\star\|^2 + 12\lambda^3\sigma^2 \sum_{k=k_p}^{k_{p+1}-1} \sigma^2. \tag{52}$$

Proof.

$$\begin{aligned} V_k &= \frac{1}{M}\sum_{i=1}^{M} \left\|x_i^k - \hat{x}^k\right\|^2 \\ &= \frac{1}{M}\sum_{i=1}^{M} \|x_i^{k_p} - \hat{x}^{k_p} - \lambda \sum_{j=k_p}^{k} g_i(x_i^k) - \hat{g}^j\|^2 \\ &= \lambda^2 \frac{1}{M}\sum_{i=1}^{M} \Big\| \sum_{j=k_p}^{k} (g_i(x_i^j) - \hat{g}^j) \Big\|^2 \\ &= \lambda^2 \frac{1}{M}\sum_{i=1}^{M} (k-k_p) \sum_{j=k_p}^{k} \|g_i(x_i^k) - \hat{g}^j\|^2. \end{aligned}$$

Using the property (45),

$$\begin{aligned} V_k &\le \lambda^2 (H-1)\frac{1}{M}\sum_{i=1}^{M}\sum_{j=k_p}^{k} \|g_i(x_i^j) - \hat{g}^j\|^2 \\ &\le \lambda^2 (H-1)\frac{1}{M}\sum_{i=1}^{M}\sum_{j=k_p}^{k} \|g_i(x_i^j)\|^2. \end{aligned}$$

Using (44), we have

$$\begin{aligned} \left\|g_i(x_i^k)\right\|^2 &\le (1+c_1)\left\|g_i(x_i^k) - g_i(\hat{x}^k)\right\|^2 + \left(1+c_1^{-1}\right)\left\|g_i(\hat{x}^k)\right\|^2 \\ &\le (1+c_1)\left\|g_i(x_i^k) - g_i(\hat{x}^k)\right\|^2 + \left(1+c_1^{-1}\right)(1+c_2)\left\|g_i(\hat{x}^k) - g_i(x_*)\right\|^2 \\ &\quad + \left(1+c_1^{-1}\right)\left(1+c_2^{-1}\right)\|g_i(x_*)\|^2. \end{aligned}$$

Setting $\lambda = 2$ and $\beta = \frac{1}{3}$, we get

$$\begin{aligned} &3\left\|g_i(x_i^k) - g_i(\hat{x}^k)\right\|^2 + 2\left\|g_i(\hat{x}^k) - g_i(x_*)\right\|^2 + 6\|g_i(x_*)\|^2 \\ &= 3\left\|x_i^k - \mathcal{T}_i(x_i^k) - \hat{x}^k + \mathcal{T}_i(\hat{x}^k)\right\|^2 + 2\left\|g_i(\hat{x}^k) - g_i(x_*)\right\|^2 + 6\|g_i(x_*)\|^2. \end{aligned}$$

Then

$$\frac{1}{M}\sum_{i=1}^{M}\|g_i(x_i^k)\|^2 \leq 3\frac{1}{M}\sum_{i=1}^{M}\|x_i^k - \hat{x}^k\|^2 + 2\frac{1}{M}\sum_{i=1}^{M}\|g_i(\hat{x}^k) - g_i(x^\star)\|^2 + 6\sigma^2.$$

So, we have

$$V_k \leq \lambda^2(H-1)\sum_{j=k_p}^{k}\left(3\frac{1}{M}\sum_{i=1}^{M}\|x_i^j - \hat{x}^j\|^2 + 2\frac{1}{M}\sum_{i=1}^{M}\|g_i(\hat{x}^j) - g_i(x^\star)\|^2 + 6\sigma^2\right)$$

We get by summation:

$$\begin{aligned}
\sum_{k=k_p}^{k_{p+1}-1} V_k &\leq \lambda^2(H-1)\sum_{k=k_p}^{k_{p+1}-1}\sum_{j=k_p}^{k}\left(3\frac{1}{M}\sum_{i=1}^{M}\|x_i^j - \hat{x}^j\|^2 + 2\frac{1}{M}\sum_{i=1}^{M}\|g_i(\hat{x}^j) - g_i(x^\star)\|^2 + 6\sigma^2\right) \\
&\leq \lambda^2(H-1)\sum_{k=k_p}^{k_{p+1}-1}\sum_{j=k_p}^{k_{p+1}-1}\left(3\frac{1}{M}\sum_{i=1}^{M}\|x_i^j - \hat{x}^j\|^2 + 2\frac{1}{M}\sum_{i=1}^{M}\|g_i(\hat{x}^j) - g_i(x^\star)\|^2 + 6\sigma^2\right) \\
&\leq \lambda^2(H-1)^2\sum_{j=k_p}^{k_{p+1}-1}\left(3V_k + 2\frac{1}{M}\sum_{i=1}^{M}\|g_i(\hat{x}^j) - g_i(x^\star)\|^2 + 6\sigma^2\right)
\end{aligned}$$

Thus,

$$\begin{aligned}
(1-3\lambda^2(H-1)^2)\sum_{k=k_p}^{k_{p+1}-1} V_k &\leq \lambda^2(H-1)^2\sum_{j=k_p}^{k_{p+1}-1}\left(2\frac{1}{M}\sum_{i=1}^{M}\|g_i(\hat{x}^j) - g_i(x^\star)\|^2 + 6\sigma^2\right) \\
\sum_{k=k_p}^{k_{p+1}-1} V_k &\leq \frac{\lambda^2(H-1)^2}{(1-3\lambda^2(H-1)^2)}\sum_{j=k_p}^{k_{p+1}-1}\left(2\frac{1}{M}\sum_{i=1}^{M}\|g_i(\hat{x}^j) - g_i(x^\star)\|^2 + 6\sigma^2\right).
\end{aligned}$$

Using $\lambda \leq \frac{1}{8\max(1,H-1)}$, we get

$$\sum_{k=k_p}^{k_{p+1}-1} V_k \leq \frac{16}{15}\lambda^2(H-1)^2\sum_{j=k_p}^{k_{p+1}-1}\left(2\frac{1}{M}\sum_{i=1}^{M}\|g_i(\hat{x}^j) - g_i(x^\star)\|^2 + 6\sigma^2\right).$$

Using this result

$$\sum_{k=k_p}^{k_{p+1}-1}\left(-\frac{1}{2}\lambda(1-\lambda)\frac{1}{M}\sum_{i=1}^{M}\|\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\|^2+\lambda(2-\lambda)V_k\right)$$

$$=-\frac{1}{2}\lambda(1-\lambda)\sum_{k=k_p}^{k_{p+1}-1}\frac{1}{M}\sum_{i=1}^{M}\|\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\|^2+\lambda(2-\lambda)\sum_{k=k_p}^{k_{p+1}-1}V_k$$

$$\le-\frac{1}{2}\lambda(1-\lambda)\sum_{k=k_p}^{k_{p+1}-1}\frac{1}{M}\sum_{i=1}^{M}\|\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\|^2$$

$$+\frac{16}{15}\lambda(2-\lambda)\lambda^2(H-1)^2\sum_{k=k_p}^{k_{p+1}-1}\left(\frac{2}{M}\sum_{i=1}^{M}\|g_i(\hat{x}^k)-g_i(x^\star)\|^2+6\sigma^2\right)$$

$$=-\left(\frac{1}{2}\lambda(1-\lambda)-\frac{16}{15}\lambda(2-\lambda)\lambda^2(H-1)^2\right)\sum_{k=k_p}^{k_{p+1}-1}\frac{1}{M}\sum_{i=1}^{M}\|\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\|^2$$

$$+6\lambda(2-\lambda)\frac{16}{15}\lambda^2(H-1)^2\sum_{k=k_p}^{k_{p+1}-1}\sigma^2$$

$$\le-\frac{\lambda}{3}\sum_{k=k_p}^{k_{p+1}-1}\frac{1}{M}\sum_{i=1}^{M}\|\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\|^2+12\lambda^3(H-1)^2\sum_{k=k_p}^{k_{p+1}-1}\sigma^2.$$

□

B.3 Proof of Theorem 2.11

Suppose that $\lambda \le \frac{1}{8\max(1,H-1)}$ and that Assumption 2.8 holds. Then, for every $k \in \mathbb{N}$,

$$\frac{1}{T}\sum_{k=0}^{T-1}\left\|\hat{x}^k-\mathcal{T}(\hat{x}^k)\right\|^2\le\frac{3\|\hat{x}^0-x^\star\|^2}{\lambda T}+36\lambda^2(H-1)^2\sigma^2. \tag{53}$$

Proof. Using statment of lemma 2.9:

$$\|\hat{x}^{k+1}-x^\star\|^2\le\|\hat{x}^k-x^\star\|^2-\frac{1}{2}\lambda(1-\lambda)\frac{1}{M}\sum_{i=1}^{M}\left\|\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\right\|^2+\lambda(2-\lambda)V_k.$$

Summing up these inequalities gives

$$\sum_{k=0}^{T-1}\|\hat{x}^{k+1}-x^\star\|^2\le\sum_{k=0}^{T-1}|\hat{x}^k-x^\star\|^2$$

$$+\sum_{k=0}^{T-1}\left(-\frac{1}{2}\lambda(1-\lambda)\frac{1}{M}\sum_{i=1}^{M}\left\|\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\right\|^2+\lambda(2-\lambda)V_k\right).$$

Considering this and using (52)

$$
\begin{aligned}
&\sum_{k=0}^{T-1}\left(-\frac{1}{2}\lambda(1-\lambda)\frac{1}{M}\sum_{i=1}^{M}\left\|\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\right\|^2+\lambda(2-\lambda)V_k\right)\\
&=\sum_{s=1}^{p}\sum_{j=k_{s-1}}^{k_s-1}\left(-\frac{1}{2}\lambda(1-\lambda)\frac{1}{M}\sum_{i=1}^{M}\left\|\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\right\|^2+\lambda(2-\lambda)V_k\right)\\
&\le\sum_{s=1}^{p}\sum_{j=k_p}^{k_p-1}\left(-\frac{1}{2}\lambda(1-\lambda)\frac{1}{M}\sum_{i=1}^{M}\left\|\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\right\|^2+\lambda(2-\lambda)V_k\right)\\
&\le\sum_{s=1}^{p}\left(-\frac{\lambda}{3}\sum_{k=k_p}^{k_{p+1}-1}\frac{1}{M}\sum_{i=1}^{M}\left\|\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\right\|^2\right)+\sum_{s=1}^{p}\left(12\lambda^3(H-1)^2\sum_{k=k_p}^{k_{p+1}-1}\sigma^2\right)\\
&\le 12\lambda^3(H-1)^2\sum_{k=0}^{T-1}\sigma^2-\frac{\lambda}{3}\sum_{k=0}^{T-1}\frac{1}{M}\sum_{i=1}^{M}\left\|\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\right\|^2.
\end{aligned}
$$

Hence,

$$
\sum_{k=0}^{T-1}\|\hat{x}^{k+1}-x^\star\|^2\le\sum_{k=0}^{T-1}|\hat{x}^k-x^\star\|^2+12\lambda^3(H-1)^2\sum_{k=0}^{T-1}\sigma^2-\frac{\lambda}{3}\sum_{k=0}^{T-1}\frac{1}{M}\sum_{i=1}^{M}\left\|\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\right\|^2
$$

Telescoping this sum:

$$
\frac{\lambda}{3}\sum_{k=0}^{T-1}\frac{1}{M}\sum_{i=1}^{M}\left\|\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\right\|^2\le\|\hat{x}^0-x^\star\|^2-\|\hat{x}^T-x^\star\|^2+12\lambda^3(H-1)^2\sum_{k=0}^{T-1}\sigma^2.
$$

Using Jensen's inequality (41):

$$
\left\|\hat{x}^k-\frac{1}{M}\sum_{i=1}^{M}\mathcal{T}_i(\hat{x}^k)\right\|^2=\left\|\frac{1}{M}\sum_{i=1}^{M}\left(\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\right)\right\|^2\le\frac{1}{M}\sum_{i=1}^{M}\left\|\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\right\|^2.
$$

Finally, we have

$$
\begin{aligned}
\frac{\lambda}{3}\sum_{k=0}^{T-1}\frac{1}{M}\sum_{i=1}^{M}\left\|\hat{x}^k-\mathcal{T}_i(\hat{x}^k)+\mathcal{T}_i(x^\star)-x^\star\right\|^2&\le\|\hat{x}^0-x^\star\|^2+12\lambda^3(H-1)^2\sum_{k=0}^{T-1}\sigma^2\\
\frac{\lambda}{3}\sum_{k=0}^{T-1}\left\|\hat{x}^k-\frac{1}{M}\sum_{i=1}^{M}\mathcal{T}_i(\hat{x}^k)\right\|^2&\le\|\hat{x}^0-x^\star\|^2+12\lambda^3(H-1)^2T\sigma^2\\
\frac{1}{T}\sum_{k=0}^{T-1}\left\|\hat{x}^k-\frac{1}{M}\sum_{i=1}^{M}\mathcal{T}_i(\hat{x}^k)\right\|^2&\le\frac{3\|\hat{x}^0-x^\star\|^2}{\lambda T}+36\lambda^2(H-1)^2\sigma^2\\
\frac{1}{T}\sum_{k=0}^{T-1}\left\|\hat{x}^k-\mathcal{T}(\hat{x}^k)\right\|^2&\le\frac{3\|\hat{x}^0-x^\star\|^2}{\lambda T}+36\lambda^2(H-1)^2\sigma^2.
\end{aligned}
$$

$\square$

B.4 Proof of Corollary 2.12

Suppose that $\lambda \leq \frac{1}{8\max(1,H-1)}$ and that Assumption 2.8 holds. Then a sufficient condition on the number T of iterations to reach ε-accuracy, for any $\varepsilon > 0$, is

$$\frac{T}{H-1} \geq \frac{24\|\hat{x}^0 - x^*\|^2}{\varepsilon} \max\left\{2, \frac{3\sigma}{\sqrt{2\varepsilon}}\right\}. \tag{54}$$

Proof.

$$\frac{3\|\hat{x}^0 - x^\star\|^2}{\lambda T} + 36\lambda^2(H-1)^2\sigma^2 \leq \varepsilon.$$

We have

$$\frac{3\|\hat{x}^0 - x^*\|^2}{\lambda T} \leq \frac{\varepsilon}{2} \Rightarrow T \geq \frac{6\|\hat{x}^0 - x^*\|^2}{\lambda\varepsilon}$$
$$36\lambda^2(H-1)^2\sigma^2 \leq \frac{\varepsilon}{2} \Rightarrow \lambda \leq \frac{\sqrt{\varepsilon}}{6\sqrt{2}(H-1)\sigma}.$$

So, we have

$$\lambda = \min\left\{\frac{1}{8(H-1)}, \frac{\sqrt{\varepsilon}}{6\sqrt{2}(H-1)\sigma}\right\}.$$

Using this, we get

$$\frac{T}{H-1} \geq \frac{24\|\hat{x}^0 - x^*\|^2}{\varepsilon} \max\left\{2, \frac{3\sigma}{\sqrt{2\varepsilon}}\right\}.$$

$\square$

C Analysis of Algorithm 2

C.1 Proof of Lemma 3.2

Under Assumption 3.1, for every $k \in \mathbb{N}$ and $0 \leq \lambda \leq 1$, we have

$$\|\hat{x}^{k+1} - x^\star\|^2 \leq \left(1 - \frac{\lambda\rho}{1+\rho}\right)\|\hat{x}^k - x^\star\|^2 + \frac{5}{2}\lambda V_k - \frac{1}{2}\lambda\left(\frac{1}{2} - \lambda\right)\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(x_i^k) - g_i(x^\star)\right\|^2. \tag{55}$$

Proof.

$$\begin{aligned} x_i^{k+1} &= (1-\lambda)x_i^k + \lambda\mathcal{T}_i(x_i^k) \\ &= x_i^k - \lambda\left(x_i^k - \mathcal{T}_i(x_i^k)\right) \\ &= x_i^k - \lambda g_i(x_i^k). \end{aligned}$$

So, we have

$$
\begin{aligned}
\|\hat{x}^{k+1} - x^\star\|^2 &= \|\hat{x}^{k+1} - \hat{x}^k + \hat{x}^k - x^\star\|^2 \\
&= \|\hat{x}^k - x^\star\|^2 + 2\langle \hat{x}^{k+1} - \hat{x}^k, \hat{x}^k - x^\star\rangle + \|\hat{x}^{k+1} - \hat{x}^k\|^2 \\
&= \|\hat{x}^k - x^\star\|^2 + 2\langle (1-\lambda)\hat{x}^k + \lambda \frac{1}{M}\sum_{i=1}^M \mathcal{T}_i(x_i^k) - \hat{x}^k, \hat{x}^k - x^\star\rangle \\
&\quad + \Big\|(1-\lambda)\hat{x}^k + \lambda \frac{1}{M}\sum_{i=1}^M \mathcal{T}_i(x_i^k) - \hat{x}^k\Big\|^2 \\
&= \|\hat{x}^k - x^\star\|^2 + 2\lambda\langle \frac{1}{M}\sum_{i=1}^M \mathcal{T}_i(x_i^k) - \hat{x}^k, \hat{x}^k - x^\star\rangle \\
&\quad + \lambda^2\Big\|\frac{1}{M}\sum_{i=1}^M \big(\mathcal{T}_i(x_i^k) - \hat{x}^k\big)\Big\|^2 \\
&\|\hat{x}^k - x^\star\|^2 + 2\lambda \frac{1}{M}\sum_{i=1}^M \langle \mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star, \hat{x}^k - x^\star\rangle \\
&\quad + \lambda^2\Big\|\frac{1}{M}\sum_{i=1}^M \big(\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\big)\Big\|^2
\end{aligned}
$$

$$
\begin{aligned}
\|\hat{x}^{k+1} - x^\star\|^2 &= \|\hat{x}^k - x^\star\|^2 + \lambda^2\Big\|\frac{1}{M}\sum_{i=1}^M \big(\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\big)\Big\|^2 \\
&\quad + 2\lambda \frac{1}{M}\sum_{i=1}^M \big[\langle \mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star, x_i^k - x^\star\rangle + \langle \mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star, \hat{x}^k - x_i^k\rangle\big] \\
&= 2\lambda \frac{1}{M}\sum_{i=1}^M \langle \mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star, \hat{x}^k - x_i^k\rangle + \|\hat{x}^k - x^\star\|^2 \\
&\quad + \frac{2\lambda}{M}\sum_{i=1}^M \langle \mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star, x_i^k - x^\star\rangle \\
&\quad + \lambda^2\Big\|\frac{1}{M}\sum_{i=1}^M \big(\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\big)\Big\|^2.
\end{aligned}
$$

Using Technical Lemma 2,

$$
\begin{aligned}
\|\hat{x}^{k+1} - x^\star\|^2 &\le \|\hat{x}^k - x^\star\|^2 \\
&\quad - 2\lambda \frac{1}{M}\sum_{i=1}^M \left(\frac{\rho}{2(1+\rho)}\|x_i^k - x^\star\|^2\right) \\
&\quad - 2\lambda \frac{1}{M}\sum_{i=1}^M \frac{2+\rho}{2(1+\rho)}\Big\|\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\Big\|^2 \\
&\quad + 2\lambda \frac{1}{M}\sum_{i=1}^M \langle \mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star, \hat{x}^k - x_i^k\rangle \\
&\quad + \lambda^2\Big\|\frac{1}{M}\sum_{i=1}^M \big(\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\big)\Big\|^2.
\end{aligned}
$$

$$\|\hat{x}^{k+1}-x^\star\|^2 \le \|\hat{x}^k - x^\star\|^2 - \frac{\lambda\rho}{(1+\rho)}\Big\|\frac{1}{M}\sum_{i=1}^M (x_i^k - x^\star)\Big\|^2$$
$$- 2\lambda\frac{2+\rho}{2(1+\rho)}\frac{1}{M}\sum_{i=1}^M \Big\|\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\Big\|^2$$
$$+ 2\lambda\frac{1}{M}\sum_{i=1}^M \langle \mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star, \hat{x}^k - x_i^k\rangle + \lambda^2\frac{1}{M}\sum_{i=1}^M \Big\|\big(\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\big)\Big\|^2.$$

Using the inequality (44)

$$\|\hat{x}^{k+1}-x^\star\|^2 \le \|\hat{x}^k - x^\star\|^2\left(1 - \frac{\lambda\rho}{1+\rho}\right) + 2\lambda\frac{1}{M}\sum_{i=1}^M\left(\frac{1}{4}\Big\|\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\Big\|^2 + \Big\|\hat{x}^k - x_i^k\Big\|^2\right)$$
$$- 2\lambda\frac{2+\rho}{2(1+\rho)}\frac{1}{M}\sum_{i=1}^M \Big\|\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\Big\|^2 + \lambda^2\frac{1}{M}\sum_{i=1}^M \Big\|\big(\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\big)\Big\|^2$$
$$\le \|\hat{x}^k - x^\star\|^2\left(1 - \frac{\lambda\rho}{1+\rho}\right) + \left[\lambda^2 + \frac{1}{2}\lambda - \frac{\lambda(2+\rho)}{1+\rho}\right]\frac{1}{M}\sum_{i=1}^M \Big\|\big(\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star\big)\Big\|^2 + 2\lambda V_k.$$

Hence,

$$\begin{aligned}
\|\hat{x}^{k+1}-x^\star\|^2 &\le \left(1 - \frac{\lambda\rho}{1+\rho}\right)\|\hat{x}^k - x^\star\|^2 + 2\lambda V_k \\
&\quad - \lambda\left(\frac{1}{2} - \lambda\right)\frac{1}{M}\sum_{i=1}^M \Big\|\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(x^\star) + x^\star + \mathcal{T}_i(\hat{x}^k) - \mathcal{T}_i(\hat{x}^k) + \hat{x}^k - \hat{x}^k\Big\|^2 \\
&= \left(1 - \frac{\lambda\rho}{1+\rho}\right)\|\hat{x}^k - x^\star\|^2 + 2\lambda V_k \\
&\quad - \lambda\left(\frac{1}{2} - \lambda\right)\frac{1}{M}\sum_{i=1}^M \Big\|\big(\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(\hat{x}^k) + \hat{x}^k\big) + \big(\mathcal{T}_i(\hat{x}^k) - \hat{x}^k - \mathcal{T}_i(x^\star) + x^\star\big)\Big\|^2.
\end{aligned}$$

Using (43), we have

$$\begin{aligned}
\|\hat{x}^{k+1}-x^\star\|^2 &\le \left(1 - \frac{\lambda\rho}{1+\rho}\right)\|\hat{x}^k - x^\star\|^2 - \frac{1}{2}\lambda\left(\frac{1}{2} - \lambda\right)\frac{1}{M}\sum_{i=1}^M \Big\|\mathcal{T}_i(\hat{x}^k) - \hat{x}^k - \mathcal{T}_i(x^\star) + x^\star\Big\|^2 \\
&\quad + \lambda\left(\frac{1}{2} - \lambda\right)\frac{1}{M}\sum_{i=1}^M \Big\|\mathcal{T}_i(x_i^k) - x_i^k - \mathcal{T}_i(\hat{x}^k) + \hat{x}^k\Big\|^2 + 2\lambda V_k \\
&\le \left(1 - \frac{\lambda\rho}{1+\rho}\right)\|\hat{x}^k - x^\star\|^2 + \lambda\left(2 + \frac{1}{2} - \lambda\right)V_k \\
&\quad - \frac{1}{2}\lambda\left(\frac{1}{2} - \lambda\right)\frac{1}{M}\sum_{i=1}^M \Big\|\mathcal{T}_i(\hat{x}^k) - \hat{x}^k - \mathcal{T}_i(x^\star) + x^\star\Big\|^2.
\end{aligned}$$

Finally, we have

$$\|\hat{x}^{k+1}-x^\star\|^2 \le \left(1 - \frac{\lambda\rho}{1+\rho}\right)\|\hat{x}^k - x^\star\|^2 - \frac{1}{2}\lambda\left(\frac{1}{2} - \lambda\right)\frac{1}{M}\sum_{i=1}^M \Big\|\hat{x}^k - \mathcal{T}_i(\hat{x}^k) + \mathcal{T}_i(x^\star) - x^\star\Big\|^2 + \frac{5}{2}\lambda V_k.$$

□

C.2 Proof of Lemma 3.3

Under Assumption 3.1 and if $\lambda < \frac{p}{15}$, we have, for every $k \in \mathbb{N}$,

$$V_k \leq \frac{2}{p}\left(1 - \frac{p}{4} + \frac{5}{p}\lambda^2\right) V_k + 20\frac{\lambda^2}{p^2}\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(\hat{x}^k) - g_i(x^\star)\right\|^2 + 60\frac{\lambda^2}{p^2}\sigma^2 - \frac{2}{p}\mathbb{E}[V_{k+1}]. \tag{56}$$

Proof. If communication happens, $V_k = 0$. Therefore,

$$\mathbb{E}[V_{k+1}] = (1-p)\frac{1}{M}\sum_{i=1}^{M}\left\|\hat{x}^k - \lambda\hat{g}^k - x_i^k + \lambda g_i(x_i^k)\right\|^2 = (1-p)\frac{1}{M}\sum_{i}^{M}\left\|\hat{x}^k - x_i^k\right\|^2$$
$$+ (1-p)\lambda^2\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(x_i^k) - \hat{g}^k\right\|^2 + 2(1-p)\lambda\frac{1}{M}\sum_{i=1}^{M}\langle\hat{x}^k - x_i^k, g_i(x_i^k) - \hat{g}^k\rangle$$

Using Young's inequality (44),

$$\mathbb{E}[V_{k+1}] \leq (1-p)\frac{1}{M}\sum_{i}^{M}\left\|\hat{x}^k - x_i^k\right\|^2 + (1-p)\lambda^2\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(x_i^k) - \hat{g}^k\right\|^2$$
$$+ \frac{p}{4}(1-p)\frac{1}{M}\sum_{i}^{M}\left\|\hat{x}^k - x_i^k\right\|^2 + (1-p)\frac{4}{p}\lambda^2\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(x_i^k) - \hat{g}^k\right\|^2.$$

Using our notations,

$$\frac{p}{2}V_k \leq \left(1 - \frac{p}{2}\right)V_k + (1-p)\left(\lambda^2 + \frac{4}{p}\lambda^2\right)\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(x_i^k) - \hat{g}^k\right\|^2 - \mathbb{E}[V_{k+1}] + (1-p)\frac{p}{4}V_k$$
$$\leq \left(1 - \frac{p}{2} + (1-p)\frac{p}{4}\right)V_k + (1-p)\lambda^2\left(1 + \frac{4}{p}\right)\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(x_i^k) - \hat{g}^k\right\|^2 - \mathbb{E}[V_{k+1}]$$
$$\leq \left(1 - \frac{p}{2} + (1-p)\frac{p}{4}\right)V_k + \frac{5}{p}(1-p)\lambda^2\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(x_i^k)\right\|^2 - \mathbb{E}[V_{k+1}].$$

Applying the same technique, we get:

$$\left\|g_i(x_i^k)\right\|^2 \leq (1+c_1)\left\|g_i(x_i^k) - g_i(\hat{x}^k)\right\|^2 + \left(1 + c_1^{-1}\right)\left\|g_i(\hat{x}^k)\right\|^2$$
$$\leq (1+c_1)\left\|g_i(x_i^k) - g_i(\hat{x}^k)\right\|^2 + \left(1 + c_1^{-1}\right)(1+c_2)\left\|g_i(\hat{x}^k) - g_i(x_*)\right\|^2$$
$$+ \left(1 + c_1^{-1}\right)\left(1 + c_2^{-1}\right)\left\|g_i(x_*)\right\|^2.$$

Setting $c_1 = 2$, $c_2 = \frac{1}{3}$, we get

$$3\left\|g_i(x_i^k) - g_i(\hat{x}^k)\right\|^2 + 2\left\|g_i(\hat{x}^k) - g_i(x_*)\right\|^2 + 6\left\|g_i(x_*)\right\|^2$$
$$= 3\left\|x_i^k - \mathcal{T}_i(x_i^k) - \hat{x}^k + \mathcal{T}_i(\hat{x}^k)\right\|^2 + 2\left\|g_i(\hat{x}^k) - g_i(x_*)\right\|^2 + 6\left\|g_i(x_*)\right\|^2.$$

By averaging,

$$\frac{1}{M}\sum_{i=1}^{M}\|g_i(x_i^k)\|^2 \leq 3\frac{1}{M}\sum_{i=1}^{M}\|x_i^k - \hat{x}^k\|^2 + 2\frac{1}{M}\sum_{i=1}^{M}\|g_i(\hat{x}^k) - g_i(x^\star)\|^2 + 6\sigma^2.$$

Using this inequality,

$$\frac{p}{2}V_k \leq \left(1-\frac{p}{2}+(1-p)\frac{p}{4}\right)V_k+\frac{5}{p}(1-p)\lambda^2\left(3V_k+2\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(\hat{x}^k)-g_i(x^\star)\right\|^2+6\sigma^2\right)-\mathbb{E}[V_{k+1}]$$

$$\leq\left(1-\frac{p}{2}+(1-p)\left(\frac{p}{4}+\frac{5}{p}\lambda^2\right)\right)V_k+\frac{5}{p}(1-p)\lambda^2\left(2\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(\hat{x}^k)-g_i(x^\star)\right\|^2+6\sigma^2\right)-\mathbb{E}[V_{k+1}]$$

$$\leq\left(1-\frac{p}{4}+\frac{5}{p}\lambda^2\right)V_k+\frac{10}{p}\lambda^2\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(\hat{x}^k)-g_i(x^\star)\right\|^2+30\frac{\lambda^2}{p}\sigma^2-\mathbb{E}[V_{k+1}].$$

Finally, we get

$$V_k\leq\frac{2}{p}\left(1-\frac{p}{4}+\frac{5}{p}\lambda^2\right)V_k+20\frac{\lambda^2}{p^2}\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(\hat{x}^k)-g_i(x^\star)\right\|^2+60\frac{\lambda^2}{p^2}\sigma^2-\frac{2}{p}\mathbb{E}[V_{k+1}].$$

□

C.3 Proof of Theorem 3.4

For every $k\in\mathbb{N}$, let Ψ^k be the Lyapunov function defined as:

$$\Psi^k:=\|\hat{x}^k-x^\star\|^2+\frac{5\lambda}{p}V_k. \tag{57}$$

Under Assumption 3.1 and if $\lambda<\frac{p}{15}$, we have, for every $k\in\mathbb{N}$,

$$\mathbb{E}\Psi^k\leq\left(1-\min\left(\frac{\lambda\rho}{1+\rho},\frac{p}{5}\right)\right)^k\Psi^0+\frac{150}{\min\left(\frac{\lambda\rho}{1+\rho},\frac{p}{5}\right)p^2}\lambda^3\sigma^2. \tag{58}$$

Proof. Using Lemma 3.2,

$$\|\hat{x}^{k+1}-x^\star\|^2\leq\left(1-\frac{\lambda\rho}{1+\rho}\right)\|\hat{x}^k-x^\star\|^2-\frac{1}{2}\lambda\left(\frac{1}{2}-\lambda\right)\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(\hat{x}^k)-g_i(x^\star)\right\|^2+\frac{5}{2}\lambda V_k.$$

Using Lemma 3.3,

$$\begin{aligned}\|\hat{x}^{k+1}-x^\star\|^2&\leq\left(1-\frac{\lambda\rho}{1+\rho}\right)\|\hat{x}^k-x^\star\|^2-\frac{1}{2}\lambda\left(\frac{1}{2}-\lambda\right)\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(\hat{x}^k)-g_i(x^\star)\right\|^2\\&\quad+\frac{5}{2}\lambda\frac{2}{p}\left(\left(1-\frac{p}{4}+\frac{5}{p}\lambda^2\right)V_k-\mathbb{E}[V_{k+1}]\right)+\frac{5}{2}\lambda\frac{20}{p^2}\lambda^2\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(\hat{x}^k)-g_i(x^\star)\right\|^2\\&\quad+60\frac{\lambda^2}{p^2}\frac{5}{2}\lambda\sigma^2\\&\leq\left(1-\frac{\lambda\rho}{1+\rho}\right)\|\hat{x}^k-x^\star\|^2+\lambda\left(\frac{50}{p^2}\lambda^2-\frac{1}{2}\left(\frac{1}{2}-\lambda\right)\right)\frac{1}{M}\sum_{i=1}^{M}\left\|g_i(\hat{x}^k)-g_i(x^\star)\right\|^2\\&\quad+\frac{5}{2}\lambda\frac{2}{p}\left(\left(1-\frac{p}{4}+\frac{5}{p}\lambda^2\right)V_k-\mathbb{E}[V_{k+1}]\right)+150\frac{\lambda^3}{p^2}\sigma^2.\end{aligned}$$

If $\lambda\leq\frac{p}{15}$, we have

$$\|\hat{x}^{k+1}-x^\star\|^2\leq\left(1-\frac{\lambda\rho}{1+\rho}\right)\|\hat{x}^k-x^\star\|^2+\frac{5\lambda}{p}\left(\left(1-\frac{p}{4}+\frac{5}{p}\lambda^2\right)V_k-\mathbb{E}[V_{k+1}]\right)+150\frac{\lambda^3}{p^2}\sigma^2.$$

We have the contraction property

$$\|\hat{x}^{k+1} - x^\star\|^2 + \frac{5\lambda}{p}\mathbb{E}[V_{k+1}] \leq \left(1 - \frac{\lambda\rho}{1+\rho}\right)\|\hat{x}^k - x^\star\|^2 + \frac{5\lambda}{p}\left(1 - \frac{p}{5}\right)V_k + 150\frac{\lambda^3}{p^2}\sigma^2.$$

Define the Lyapunov function:

$$\Psi^k = \|\hat{x}^k - x^\star\|^2 + \frac{5\lambda}{p}V_k.$$

Using the law of total expectation, we get

$$\mathbb{E}\Psi^{k+1} \leq \left(1 - \min\left(\frac{\lambda\rho}{1+\rho}, \frac{p}{5}\right)\right)\Psi^k + 150\frac{\lambda^3}{p^2}\sigma^2.$$

Finally we get

$$\mathbb{E}\Psi^T \leq \left(1 - \min\left(\frac{\lambda\rho}{1+\rho}, \frac{p}{5}\right)\right)^T \Psi^0 + \frac{150}{\min\left(\frac{\lambda\rho}{1+\rho}, \frac{p}{5}\right)}\frac{\lambda^3}{p^2}\sigma^2$$

□

C.4 Proof of Corollary 3.5

Under Assumption 3.1 and if $\lambda < \frac{p}{15}$, for any $\varepsilon > 0$, ε-accuracy is reached after T iterations, with

$$T \geq \max\left\{\frac{15(1+\rho)}{\rho p}, \frac{18\sigma(1+\rho)^{\frac{1}{3}}}{p\rho^{\frac{3}{2}}\varepsilon^{\frac{1}{2}}}, \frac{40\sigma^{\frac{2}{3}}(1+\rho)}{p\rho\varepsilon^{\frac{1}{3}}}\right\}\log\frac{2\Psi_0}{\varepsilon}. \tag{59}$$

Proof. We start from

$$\left[1 - \min\left\{\frac{\lambda\rho}{\rho+1}, \frac{p}{5}\right\}\right]^k \Psi_0 + \frac{150\lambda^3\sigma^2}{p^2\min\left\{\frac{\lambda\rho}{\rho+1}, \frac{p}{5}\right\}} \leq \varepsilon.$$

Concerning the second term,

$$150\lambda^3\sigma^2 \leq \frac{1}{2}p^2\varepsilon\min\left\{\frac{\lambda\rho}{\rho+1}, \frac{p}{5}\right\} \Rightarrow \begin{cases}150\lambda^3\sigma^2 \leq \frac{1}{2}p^2\varepsilon\frac{\lambda\rho}{\rho+1}, \\ 150\lambda^3\sigma^2 \leq \frac{p^3\varepsilon}{10}\end{cases}$$

$$\lambda \leq \min\left\{\frac{p}{18\sigma}\sqrt{\frac{\varepsilon\rho}{\rho+1}}, \frac{p\varepsilon^{\frac{1}{3}}}{40\sigma^{\frac{2}{3}}}\right\}$$

Concerning the first term and using the fact that $\lambda < \frac{p}{15}$,

$$\left[1 - \min\left\{\frac{\lambda\rho}{\rho+1}, \frac{p}{5}\right\}\right]^T \Psi_0 \leq \frac{\varepsilon}{2} \Rightarrow T \geq \max\left\{\frac{1+\rho}{\lambda\rho}, \frac{5}{p}\right\}\log\frac{2\Psi_0}{\varepsilon}$$

$$\lambda = \min\left\{\frac{p}{15}, \frac{p}{18\sigma}\sqrt{\frac{\varepsilon\rho}{\rho+1}}, \frac{p\varepsilon^{\frac{1}{3}}}{40\sigma^{\frac{2}{3}}}\right\}.$$

Finally, we get

$$\begin{aligned} T &\geq \max\left\{\frac{5}{p}, \frac{15(1+\rho)}{\rho p}, \frac{18\sigma(1+\rho)^{\frac{1}{3}}}{p\rho^{\frac{3}{2}}\varepsilon^{\frac{1}{2}}}, \frac{40\sigma^{\frac{2}{3}}(1+\rho)}{p\rho\varepsilon^{\frac{1}{3}}}\right\}\log\frac{2\Psi_0}{\varepsilon} \\ &= \max\left\{\frac{15(1+\rho)}{\rho p}, \frac{18\sigma(1+\rho)^{\frac{1}{3}}}{p\rho^{\frac{3}{2}}\varepsilon^{\frac{1}{2}}}, \frac{40\sigma^{\frac{2}{3}}(1+\rho)}{p\rho\varepsilon^{\frac{1}{3}}}\right\}\log\frac{2\Psi_0}{\varepsilon}. \end{aligned}$$

□