Moscow Institute of Physics and Technology
Phystech School of Applied Mathematics and Informatics
A. A. Kharkevich Institute for Information Transmission Problems

"Approved for defence"
Head of Department

____________________ Sobolevski A.N

«_____» ____________________ 2022 г.

A Thesis Submitted for the Degree of Master

Topic: **Methods for solving distributed saddle point problems: lower bounds, optimal and practical algorithms**

Speciality: 03.04.01 – Applied Mathematics and Physics

Student ____________ Beznosikov Aleksandr Nikolaevich

Scientific Supervisor,
D.Sc., Professor at MIPT ____________________ Gasnikov A.V.

Moscow – 2022

Abstract

This work focuses on the distributed optimization of stochastic saddle point problems. The first part of the paper is devoted to lower bounds for the cenralized and decentralized distributed methods for smooth (strongly) convex-(strongly) concave saddle-point problems as well as the near-optimal algorithms by which these bounds are achieved. Next, we present a new federated algorithm for cenralized distributed saddle point problems – Extra Step Local SGD. Theoretical analysis of the new method is carried out for strongly convex-strongly concave and non-convex-non-concave problems. In the experimental part of the paper, we show the effectiveness of our method in practice. In particular, we train GANs in a distributed manner.

Contents

1. Introduction 4
2. Settings and assumptions 7
3. Lower bounds 9
4. Optimal algorithms 11
5. Local algorithm 14
6. Experiments 16
7. Conclusion and future work 18

Bibliography 19

8. General facts and technical lemmas 24
9. Proof of Theorems from Section 3 24
10. Proof of Theorems from Section 4 35
11. Proof of Theorems from Section 5 56

1. Introduction

Distributed algorithms have already become an integral part of solving many applied tasks, including machine learning problems [1–3]. This paper also deals with distributed methods, we study the saddle point problem (SPP):

$$\min_{x \in X} \max_{y \in Y} f(x,y) := \frac{1}{M}\sum_{m=1}^{M} f_m(x,y), \tag{1}$$

where parts of the function f are distributed between M nodes, while the function f_m is known only to the node with the corresponding number m. SPPs or Min-Max problems, including distributed ones, have many applications. Here we can note the already classical and long-established applications in equilibrium theory, games and economics [4], as well as new and young trends in image deconvolution [5, 6], reinforcement and statistical learning [7, 8], adversarial training [9] and GANs [10]. In particular, a series of papers [11–16] showed the connection of the theory for convex SPPs with the training of GANs and obtained insights and useful tips for the GANs community. From the point of view of machine learning, it can be interesting when f_m is an empirical loss function of the model on the local data of the mth device. Therefore, we consider the statement of the problem (1) when we have access only to local stochastic oracle of $f_m(x,y) := \mathbb{E}_{\xi_m \sim \mathcal{D}_m} f_m(x,y,\xi_m)$, where the data ξ_m follows unknown distributions $\mathcal{D}_m$.

However, the main problem of the distributed learning tasks is not the stochasticity, but precisely the separation of the data from the devices. All f_m have access only to their own data, while transferring data to other devices may be inefficient and, moreover, impossible for privacy reasons. Therefore, to solve (1), it becomes necessary to construct a distributed algorithm that combine local computations on each of the devices and communication between them. Such an Algorithm can be organized as follows: all devices communicate only with the main device (server). This approach is called centralized. The main problem is the importance of the server – it can crash and interrupt the whole process. Therefore, along with the centralized approach, the decentralized [17] one is also popular. In this case, all devices are equal and connected to a network, communication occurs along the edges of this network.

Both centralised and decentralised methods are well developed for minimization problems. But meanwhile, the direction of distributed algorithms for SPPs is much weaker. Our work makes the following contribution to this area:

1.1. Summary of results

· **Lower bounds** We present lower bounds for distributed stochastic smooth strongly-convex-strongly-concave and convex-concave SPPs. Both centralized and decentralized cases are considered.

· **Optimal algorithms** Next, we obtain the near-optimal algorithms. They are near-optimal from a theoretical point of view, because their upper bounds reach lower bounds (up to numerical constants and logarithmic factors). For the centralized problem, we construct our method based on Extra Step method [18, 19] (classical and optimal method for non-distributed SPPs) with the right size of batches. In decentralized case, we also use Extra Step method as base, but communication takes place with the help of accelerated (gossip) consensus procedure [20].

To sum up and compare the lower and upper bounds see Table 1.

· **Local method** We also present an extra-step modification of Local SGD [21, 22], one of the most popular methods in Federated Learning [23, 24]. More recently, other versions of the Local SGD methods for SPPs have appeared [25, 26]. All of the methods presented in these papers are based on gradient descent-ascent, but it is known that such methods, even in the non-distributed case, diverge for the most common SPPs [11, 27]. Our method is based on the classic method for smooth SPPs – Extra Step algorithm, which makes it stand out from the competitors.

· **Non-convex-non-concave analysis** We analyze our new algorithms: near-optimal and local, not only in convex-concave case, but even in the non-convex-non-concave case under minty assumption [28, 29]. Minty is the weakest additional assumption for a non-convex-non-concave problem found in the literature. But even with the these minty assumption, there are not many analyses of distributed methods [30, 31]. In particular, our analysis covers the estimates of the decentralized but deterministic method from [30], and also generalizes and overlaps the estimates for the stochastic method for homogeneous data ($f_m = f$) from [31].

· **Experiments** The first part of our experiments on classical bilinear problem is devoted to the comparison of the optimal centralized method and the method based on Local SGD, as well as comparison of our local method with competitors [25, 26]. The second part is devoted to the use of Local SGD and Local Adam techniques for training GANs in a homogeneous and heterogeneous cases.

	lower	upper
	centralized	
sc	$\Omega\left(R_0^2\exp\left(-\frac{32\mu K}{L\Delta}\right)+\frac{\sigma^2}{\mu^2 MT}\right)$	$\tilde{O}\left(R_0^2\exp\left(-\frac{\mu K}{4L\Delta}\right)+\frac{\sigma^2}{\mu^2 MT}\right)$
c	$\Omega\left(\frac{L\Omega_z^2\Delta}{K}+\frac{\sigma\Omega_z}{\sqrt{MT}}\right)$	$O\left(\frac{L\Omega_z^2\Delta}{K}+\frac{\sigma\Omega_z}{\sqrt{MT}}\right)$
	decentralized	
sc	$\Omega\left(R_0^2\exp\left(-\frac{128\mu K}{L\sqrt{\chi}}\right)+\frac{\sigma^2}{\mu^2 MT}\right)$	$\tilde{O}\left(R_0^2\exp\left(-\frac{\mu K}{8L\sqrt{\chi}}\right)+\frac{\sigma^2}{\mu^2 MT}\right)$
c	$\Omega\left(\frac{L\Omega_z^2\sqrt{\chi}}{K}+\frac{\sigma\Omega_z}{\sqrt{MT}}\right)$	$\tilde{O}\left(\frac{L\Omega_z^2\sqrt{\chi}}{K}+\frac{\sigma\Omega_z}{\sqrt{MT}}\right)$

Table 1: Lower and upper bounds for distributed smooth stochastic strongly-convex–strongly-concave (sc) or convex-concave (c) saddle point problems in centralized and decentralized cases. Notation: L – smoothness constant of f, μ – constant of strong convexity-strong concavity, $R_0 = \|z_0 - z^*\|_2$, Ω_z – diameter of optimization set, Δ, χ – diameter and condition number of communication graph (condition number of gossip matrix), K – number of communication rounds, T – number of local calls of gradient oracle on each node. In the convex-concave case, the bounds are in terms of the gap function, in the strongly convex-strongly concave case – in terms of the (squared) distance to the solution.

1.2. Related works

SPPs. First, we highlight two main non-distributed algorithms for SPPs. First algorithm – Mirror Descent [32], it is customary to use in the non-smooth case. For smooth problems, Extra Step/Mirror Prox is applied [18, 19, 33]. Also, the following methods [34–36] can be noted as popular for smooth SPPs.

Lower bounds. In the non-distributed case, the lower bounds for smooth strongly convex-strongly concave case SPPs are given in [37], for convex-concave – in [38]. In smooth stochastic convex optimization, we highlight works about lower bounds [39, 40]. It is also important to note the works devoted to the lower bounds for centralized and decentralized distributed convex optimization [41, 42].

Distributed SPPs. The following works are devoted to decentralized Min-Max: in the deterministic case [30, 43, 44], in the stochastic case [31]. Let us also highlight the local methods for SPPs [25, 26] already noted earlier in Section 1.1.

2. Settings and assumptions

We consider problem (1), where the sets $X \subseteq \mathbb{R}^{n_x}$ and $Y \subseteq \mathbb{R}^{n_y}$ are convex sets. For simplicity, we introduce the set $Z = X \times Y$, $z = (x, y)$ and the operators F_m:

$$F_m(z) := F_m(x,y) := \begin{pmatrix} \nabla_x f_m(x,y) \\ -\nabla_y f_m(x,y) \end{pmatrix}. \tag{2}$$

As mentioned above, we do not have access to the oracles for $F_m(z)$, at each iteration our oracles gives only some stochastic realization $F_m(z,\xi)$. Next, we introduce the following assumptions:

Assumption 1(g). *$f(x,y)$ is L - smooth, i.e. for all $z_1, z_2 \in Z$*

$$\|F(z_1) - F(z_2)\| \leq L\|z_1 - z_2\|. \tag{3}$$

Assumption 1(l). *For all m, $f_m(x,y)$ is $L_{\max}$-smooth, i.e. for all $z_1, z_2 \in Z$*

$$\|F_m(z_1) - F_m(z_2)\| \leq L_{\max}\|z_1 - z_2\|. \tag{4}$$

Assumption 2(sc). *$f(x,y)$ is strongly-convex-strongly-concave with constant μ, if for all $z_1, z_2 \in Z$*

$$\langle F(z_1) - F(z_2), z_1 - z_2\rangle \geq \mu\|z_1 - z_2\|^2. \tag{5}$$

Assumption 2(c). *$f(x,y)$ is convex-concave, if $f(x,y)$ is strongly-convex-strongly-concave with $\mu = 0$.*

Assumption 2(nc). *f satisfies the minty assumption, if exists $z^* \in Z$ such that for all $z \in Z$*

$$\langle F(z), z - z^*\rangle \geq 0. \tag{6}$$

Assumption 3. *$F_m(z,\xi)$ is unbiased and has bounded variance, i.e. for all $z \in Z$ it holds that*

$$\mathbb{E}[F_m(z,\xi)] = F_m(z), \quad \mathbb{E}[\|F_m(z,\xi) - F_m(z)\|^2] \leq \sigma^2. \tag{7}$$

Assumption 4. *$\mathcal{Z}$ is compact bounded, i.e. for all $z, z' \in \mathcal{Z}$*

$$\|z - z'\| \leq \Omega_z. \tag{8}$$

Hereinafter, we use the standard Euclidean norm $\|\cdot\|$. We also introduce the following notation $\text{proj}_{\mathcal{Z}}(z) = \min_{u \in \mathcal{Z}} \|u - z\|$ – the Euclidean projection onto $\mathcal{Z}$.

We also assume that all devices are connected to each other in a network, which can be represented as an undirected graph $G(\mathcal{V}, \mathcal{E})$ with diameter Δ. As mentioned earlier, we are interested in several cases of distributed optimization: centralized, and decentralized. It is important to mention one of the most popular communication procedures in decentralised setup – *the gossip protocol* [45–47]. This approach uses a certain matrix W. Local vectors during communications are "weighted" by multiplication by W. The convergence of decentralized algorithms is determined by the properties of this matrix. Therefore, we introduce its definition:

Definition 1. *We call a $M \times M$ matrix W a gossip matrix if it satisfies the following conditions: 1) W is symmetric positive semi-definite, 2) the kernel of W is the set of constant vectors: $\ker(W) = \text{span}(\mathbf{1})$, 3) W is defined on the edges of the network: $W_{ij} \neq 0$ only if $i = j$ or $(i, j) \in \mathcal{E}$.*

Let $\lambda_1(W) \geq \ldots \geq \lambda_M(W) = 0$ the spectrum of W, and condition number $\chi = \chi(W) = \frac{\lambda_1(W)}{\lambda_{M-1}(W)}$. Note that in practical algorithms [20, 41, 46, 48] is used not the matrix W, but $\tilde{W} = I - \frac{W}{\lambda_1(W)}$. To describe the convergence, we introduce $\lambda_2(\tilde{W}) = 1 - \frac{\lambda_{M-1}(W)}{\lambda_1(W)} = 1 - \frac{1}{\chi(W)} = 1 - \frac{1}{\chi}$.

The next definition is necessary to describe a certain class of distributed algorithms, for which we will obtain lower bounds. We use a definition quite similar to [41, 42].

Definition 2. *Let introduce some procedure with two parameters T and K, which we call* `Black-Box Procedure`(T, K). *Each agent m has its own local memories $\mathcal{M}^x_m$ and $\mathcal{M}^y_m$ for the x- and y-variables, respectively–with initialization $\mathcal{M}^x_m = \mathcal{M}^y_m = \{0\}$. $\mathcal{M}^x_m$ and $\mathcal{M}^x_m$ are updated as follows.*

- *Local computation: At each local iteration device m can sample uniformly and independently random variable ξ_m and adds to its $\mathcal{M}^x_m$ and $\mathcal{M}^y_m$ a finite number of points*

x, y, satisfying

$$x \in \text{span}\{x', \nabla_x f_m(x'', y'', \xi_m)\}, \quad y \in \text{span}\{y', \nabla_y f_m(x'', y'', \xi_m)\}, \tag{9}$$

for given $x', x'' \in \mathcal{M}_m^x$ and $y', y'' \in \mathcal{M}_m^y$.

· *Communication: Based upon communication rounds among neighbouring nodes, $\mathcal{M}_m^x$ and $\mathcal{M}_m^y$ are updated according to*

$$\mathcal{M}_m^x := \text{span}\left\{ \bigcup_{(i,m)\in E} \mathcal{M}_i^x \right\}, \quad \mathcal{M}_m^y := \text{span}\left\{ \bigcup_{(i,m)\in E} \mathcal{M}_i^y \right\}. \tag{10}$$

· *Output: The final global output is calculated as:*

$$\hat{x} \in \text{span}\left\{ \bigcup_{m=1}^{M} \mathcal{M}_m^x \right\}, \quad \hat{y} \in \text{span}\left\{ \bigcup_{m=1}^{M} \mathcal{M}_m^y \right\}.$$

We assume that each node makes no more than T local iterations (for simplicity, that exactly T) during the operation of the algorithm. The number of communication rounds is also limited to a certain number of $K < T$.

3. Lower bounds

Following the classical results on obtaining lower bounds, it is sufficient to give an example of a «bad» function [49], and the «bad» partitioning of this function between nodes [41]. First, let us divide the original problem into two independent ones: deterministic and stochastic. Consider $f_m(x, y) = f_m^{deter}(x^{deter}, y) + f^{stoch}(x^{stoch})$, where the vectors x^{deter} and x^{stoch} together give the vector x. At the same time we have access to $F_m(x, y, \xi) = F_m^{deter}(x^{deter}, y) + \nabla f^{stoch}(x^{stoch}, \xi)$. It means that for f_m^{deter} we have a deterministic oracle and stochastic – for f^{stoch}. Such f_m helps to rewrite the original problem (1) as follows:

$$\min_{x^{deter}\in X^{deter}} \max_{y\in Y} \frac{1}{M} \sum_{m=1}^{M} f_m^{deter}(x^{deter}, y) + \min_{x^{stoch}\in X^{stoch}} f^{stoch}(x^{stoch}). \tag{11}$$

Therefore, we separately prove the estimates for each of the problems, and then combine.

3.1. Deterministic lower bounds

In this part, we provide lower bounds for the centralized (Theorem 1) and decentralized (Theorem 2) cases.

Theorem 1. *For any $L > \mu > 0$ and any connected graph with diameter Δ, there exists a distributed saddle point problem (satisfying Assumptions 3 and 5) on $X \times Y = \mathbb{R}^n \times \mathbb{R}^n$ (where n is sufficient large) with $x^*, y^* \neq 0$, such that for any output $\hat{x}, \hat{y}$ of any procedure (Definition 2), the following estimates hold:*

$$\|\hat{x} - x^*\|^2 + \|\hat{y} - y^*\|^2 = \Omega\left(\exp\left(-\frac{4\mu}{L-\mu}\cdot\frac{K}{\Delta}\right)\|y_0 - y^*\|^2\right).$$

Theorem 2. *For any $L > \mu > 0$ and any $\chi \geq 1$, there exists a decentralized distributed saddle point problem (satisfying Assumptions 3 and 5) on $X \times Y = \mathbb{R}^n \times \mathbb{R}^n$ (where n is sufficient large) with $x^*, y^* \neq 0$ over a fixed network (Definition 1) with a gossip matrix W and characteristic number χ, such that for any output $\hat{x}, \hat{y}$ of any procedure (Definition 2), the following estimates hold:*

$$\|\hat{x} - x^*\|^2 + \|\hat{y} - y^*\|^2 = \Omega\left(\exp\left(-\frac{32\mu}{L-\mu}\cdot\frac{K}{\sqrt{\chi}}\right)\|y_0 - y^*\|^2\right).$$

Regularization and convex-concave case. Note that in the convex-concave case the problem is usually considered on a bounded set (Assumption 4), moreover, the convergence criterion for algorithms is a gap function:

$$\text{gap}(x, y) := \max_{y' \in Y} f(x, y') - \min_{x' \in X} f(x', y). \tag{12}$$

Therefore the lower bounds are also needed in terms of (12). Following the inequality 6 of [37], we can rewrite the estimates from Theorems 1 and 2 as follows

$$\text{gap}(x, y) \geq \frac{\mu}{2}\|x - x^*\|^2 + \frac{\mu}{2}\|y - y^*\|^2.$$

Next, to obtain bounds for the convex-concave case from bounds for the strongly convex-strongly concave case, we use a regularization trick:

$$f_{\text{reg}}(x, y) = f(x, y) + \frac{\varepsilon}{8\Omega_z^2}\|x - x_0\|^2 - \frac{\varepsilon}{8\Omega_z^2}\|y - y_0\|^2,$$

where Ω_z is an Euclidean diameter of the set $X \times Y$. It turns out that if $f(x, y)$ is a convex-concave function, then $f_{\text{reg}}(x, y)$ is $\frac{\varepsilon}{4\Omega_z^2}$ is strongly convex-strongly concave. The new problem is solved with an accuracy of $\varepsilon/2$, then we find a solution to the original problem with an accuracy of ε.

3.2. Stochastic lower bounds

Due to our choice of f_m from (11), one can note that in order to obtain lower stochastic bounds, we need to consider the minimization problem rather than the SPP, moreover, the non-distributed minimization problem. Therefore, these bounds depend on the total number of local calls of the oracle, and this number is equal to MT. Let us formulate two theorems for the convex and strongly convex cases of f^{stoch}.

Theorem 3. *For any $L > \mu > 0$ and any $M, T \in \mathbb{N}$, there exists a stochastic minimization problem with L-smooth and μ-strongly convex function such that for any output $\hat{x}$ of any* BBP(T, K) *(Definition 2) with M workers one can obtain the following estimate:*

$$\|\hat{x} - x^*\|^2 = \Omega\left(\frac{\sigma^2}{MT\mu^2}\right).$$

Theorem 4. *For any $L > 0$ and any $M, T \in \mathbb{N}$, there exists a stochastic minimization problem with L-smooth and convex function such that for any output $\hat{x}$ of any* BBP(T, K) *(Definition 2) with M workers one can obtain the following estimate:*

$$\mathbb{E}\left[f(\hat{x}) - f(x^*)\right] = \Omega\left(\frac{\sigma\Omega_x}{\sqrt{MT}}\right).$$

3.3. Connection of lower bounds

The connecting of deterministic and stochastic bounds follows from (11). The results for the centralized and decentralized cases are shown in Table 1. See Appendix 9 for complete proof of this part. To verify the tightness of our lower bounds, the next section designs algorithms that reach such bounds.

4. Optimal algorithms

This section focuses on theoretically near-optimal algorithms. It is easy to check that our algorithms satisfy the BBP definition.

4.1. Centralized case

We design our algorithm based on MiniBatch SGD and Extra Step. For this algorithm we introduce r as a maximum distance from nodes to server.

Algorithm 1 Centralized Extra Step Method

Parameters: Stepsize $\gamma \le \frac{1}{4L}$; communication rounds K, number of local steps T.

Initialization: Choose $(x^0, y^0) = z^0 \in \mathcal{Z}$, $k = \lfloor \frac{K}{r} \rfloor$ and batch size $b = \lfloor \frac{T}{2k} \rfloor$.

for $t = 0, 1, 2, \dots, k-1$ **do**

Generate batch ξ_m^t on each machine independently

Each machine m computes $g_m^t = \frac{1}{b} \sum\limits_{i=1}^{b} F_m(z^t, \xi_m^{t,i})$ and sends g_m^t to server

Server computes $z^{t+1/2} = \text{proj}_{\mathcal{Z}}(z^t - \frac{\gamma}{M} \sum\limits_{m=1}^{M} g_m^t)$ and then sends $z^{t+1/2}$ to machines

Generate batch $\xi_m^{t+1/2}$ on each machine independently

Each machine m computes $g_m^{t+1/2} = \frac{1}{b} \sum\limits_{i=1}^{b} F_m(z^{t+1/2}, \xi_m^{t+1/2,i})$ and sends $g_m^{t+1/2}$ to server

Server computes $z^{t+1} = \text{proj}_{\mathcal{Z}}(z^t - \frac{\gamma}{M} \sum\limits_{m=1}^{M} g_m^{t+1/2})$ and then sends z^{t+1} to machines

end for

Theorem 5. *Let $\{z^t\}_{t\ge 0}$ denote the iterates of Algorithm 1 for solving problem* (1)*. Let Assumptions 1(g), 3 be satisfied. Then, if $\gamma \le \frac{1}{4L}$, we have the following estimates in*

- *μ-strongly convex-strongly concave case (Assumption 2(sc)):*

$$\mathbb{E}[\|z^k - z^*\|^2] = \tilde{\mathcal{O}}\left(\|z^0 - z^*\|^2 \exp\left(-\frac{\mu K}{4L\Delta}\right) + \frac{\sigma^2}{\mu^2 M T}\right),$$

- *convex–concave case (Assumptions 2(c) and 4):*

$$\mathbb{E}[\text{gap}(z^k_{avg})] = \mathcal{O}\left(\frac{L\Omega_z^2 \Delta}{K} + \frac{\sigma \Omega_z}{\sqrt{MT}}\right),$$

- *non-convex-non-concave case (Assumptions 2(nc) and 4):*

$$\mathbb{E}\left[\frac{1}{k}\sum_{t=0}^{k-1} \|F(z^t)\|^2\right] = \mathcal{O}\left(\frac{L^2\Omega_z^2\Delta}{K} + \frac{\sigma^2 K}{MT\Delta}\right),$$

where $z^k_{avg} = \frac{1}{k}\sum\limits_{t=0}^{k-1} z^{t+1/2}$.

4.2. Decentralized case

The idea of Algorithm 2 combines three things: Extra Step, accelerated consensus - `FastMix` (see Algorithm 4 in Appendix 10 or [20, 48]) and the right size of batches.

Theorem 6. *Let $\{z_m^t\}_{t\ge 0}$ denote the iterates of Algorithm 2 for solving problem* (1)*. Let Assumptions 1(g), 1(l), 3 be satisfied. Then, if $\gamma \le \frac{1}{4L}$ and $P = \mathcal{O}\left(\sqrt{\chi} \log \frac{1}{\varepsilon}\right)$, we have the following estimates in*

Algorithm 2 Decentralized Extra Step Method

Parameters: Stepsize $\gamma \leq \frac{1}{4L}$; communication rounds K, number of local calls T, number of `FastMix` steps P.

Initialization: Choose $(x^0, y^0) = z^0 \in \mathcal{Z}$, $z_m^0 = z^0$, $k = \lfloor \frac{K}{P} \rfloor$ and batch size $b = \lfloor \frac{T}{2k} \rfloor$.

for $t = 0, 1, 2, \ldots, k-1$ **do**

Generate batch ξ_m^t on each machine independently

Each machine m compute $\hat{z}_m^{t+1/2} = z_m^t - \gamma \cdot \frac{1}{b} \sum_{i=1}^{b} F_m(z_m^t, \xi_m^{t,i})$

Communication: $\tilde{z}_1^{t+1/2}, \ldots, \tilde{z}_M^{t+1/2} =$ `FastMix`$(\hat{z}_1^{t+1/2}, \ldots, \hat{z}_M^{t+1/2}, P)$

Each machine m compute $z_m^{t+1/2} = \text{proj}_{\mathcal{Z}}(\tilde{z}_m^{t+1/2})$

Generate batch $\xi_m^{t+1/2}$ on each machine independently

Each machine m compute $\hat{z}_m^{t+1} = z_m^t - \gamma \cdot \frac{1}{b} \sum_{i=1}^{b} F_m(z_m^{t+1/2}, \xi_m^{t+1/2,i})$

Communication: $\tilde{z}_1^{t+1}, \ldots, \tilde{z}_M^{t+1} =$ `FastMix`$(\hat{z}_1^{t+1}, \ldots, \hat{z}_M^{t+1}, P)$

Each machine m compute $z_m^{t+1} = \text{proj}_{\mathcal{Z}}(\tilde{z}_m^{t+1})$

end for

- *μ-strongly convex-strongly concave case (Assumption 2(sc)):*

$$\mathbb{E}[\|\bar{z}^k - z^*\|^2] = \tilde{\mathcal{O}}\left(\|z^0 - z^*\|^2 \exp\left(-\frac{\mu K}{8L\sqrt{\chi}}\right) + \frac{\sigma^2}{\mu^2 MT}\right),$$

- *convex-concave case (Assumptions 2(c) and 4):*

$$\mathbb{E}[\text{gap}(\bar{z}^k_{avg})] = \tilde{\mathcal{O}}\left(\frac{L\Omega_z^2\sqrt{\chi}}{K} + \frac{\sigma\Omega_z}{\sqrt{MT}}\right),$$

- *non-convex-non-concave case (Assumptions 2(nc) and 4):*

$$\mathbb{E}\left[\frac{1}{k}\sum_{t=0}^{k-1}\|F(\bar{z}^t)\|^2\right] = \tilde{\mathcal{O}}\left(\frac{L^2\Omega_z^2\sqrt{\chi}}{K} + \frac{\sigma^2 K}{MT\sqrt{\chi}}\right),$$

where $\bar{z}^t = \frac{1}{M}\sum_{m=1}^{M} z_m^t$ *and* $\bar{z}^{k+1}_{avg} = \frac{1}{Mk}\sum_{t=0}^{k-1}\sum_{m=1}^{M} z_m^{t+1/2}$.

Discussions Let us make some comments on our Algorithms:

- It is easy to see that our Algorithms are near-optimal – see Table 1 for details. However, there are several practical drawbacks of these Algorithms. The first is related to the fact that in Algorithm 2 we need to take multi consensus steps at each iteration. This approach does not always pay off in practice. On the other hand, the optimal decentralized algorithms for minimization problems also use `FastMix` – see literature review in [50]. Secondly, if $T \gg K$, at each iteration we collect a very large batch, in practice such batches do not make sense. Therefore, an idea arises to use these local computations of gradients more efficiently, for example, doing local steps. This brings us to Section 5.

· It can be noted that in the non-convex-non-concave case, we do not guarantee the convergence, when $T \approx K$. But the method converge sublinearly if $\sigma = 0$. In this case, we cover the deterministic results from [30]. In the stochastic case ($\sigma > 0$), convergence is also not guaranteed in [31, 51]. Therefore, we cover and even overlap their analysis, since they consider only the homogeneous case ($f_m = f$).

5. Local algorithm

In this section, we work on sets $X = R^{n_x}$ and $Y = R^{n_y}$. Additionally, we introduce the following assumption:

Assumption 5. *The values of the local operator are considered sufficiently close to the value of the mean operator, i.e. for all $z \in Z$*

$$\|F_m(z) - F(z)\| \leq D. \tag{13}$$

This assumption is often called D - heterogeneity.

Our algorithm is a combination of Local SGD and Extra Step. One can note that such an algorithm is BBP(T, K).

Theorem 7. *Let $\{z_m^t\}_{t\geq 0}$ denote the iterates of Algorithm 3 for solving problem* (1). *Let Assumptions 1(l), 3 and 5 be satisfied. Also let $H = \max_p |k_{p+1} - k_p|$ is a maximum distance between moments of communication ($k_p \in I$). Then we have the following estimates in*

· *μ-strongly convex-strongly concave case (Assumption 2(sc)) with $\gamma \leq \frac{1}{21HL_{\max}}$:*

$$\mathbb{E}[\|\bar{z}^T - z^*\|^2] = \tilde{O}\left(\|z^0 - z^*\|^2 \cdot \exp\left(-\frac{\mu T}{42HL_{\max}}\right) + \frac{\sigma^2}{\mu^2 MT} + \frac{L_{\max}^2 H}{\mu^4 T^2}\left(HD^2 + \sigma^2\right)\right),$$

· *non-convex–non-concave case (Assumption 2(nc) and with assumption that for all t, $\|\bar{z}^t\| \leq \Omega$) with $\gamma \leq \frac{1}{4L_{\max}}$:*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|F(\bar{z}^t)\|^2\right] = O\left(\frac{L_{\max}^2\Omega^2}{T} + \frac{\left[HL_{\max}\Omega\left(HD^2+\sigma^2\right)\right]^{2/3}}{T^{1/3}} + \frac{\sigma^2}{M} + L_{\max}\Omega\sqrt{H\left(HD^2+\sigma^2\right)}\right),$$

where $\bar{z}^t = \frac{1}{M}\sum\limits_{m=1}^{M} z_m^t$.

Algorithm 3 Extra Step Local SGD

Parameters: stepsize $\gamma \leq \frac{1}{21HL_{\max}}$; number of local steps T, sets I of communications steps ($|I| = K$).

Initialization: Choose $(x^0, y^0) = z^0 \in Z$, for all m, $z_m^0 = z^0$ and $\hat{z} = z^0$.

for $t = 0, 1, 2, \ldots, T-1$ **do**

  Generate random variable ξ_m^t on each machine independently

  Each machine m computes $z_m^{t+1/2} = z_m^t - \gamma F_m(z_m^t, \xi_m^t)$

  Generate random variable $\xi_m^{t+1/2}$ on each machine independently

  Each machine m computes $z_m^{t+1} = z_m^t - \gamma F_m(z_m^{t+1/2}, \xi_m^{t+1/2})$

  if $t \in I$ **do**

    Each machine sends z_m^{t+1} on server

    Server computes $\hat{z} = \frac{1}{M}\sum_{m=1}^{M} z_m^{t+1}$, sends $\hat{z}$ to machines

    Each machine gets $\hat{z}$ and sets $z_m^{t+1} = \hat{z}$

end for

Output: $\hat{z}$.

Discussions

· Compared to Algorithm 1, Algorithm 3 gives worse convergence guarantees. Why then Algorithm 3 is needed? For practical purposes. Local SGD or FedAvg is a fairly well-known and popular federated learning concept. We extend this concept to min-max problems, including non-convex-non-concave ones. In particular, the theory states that for Algorithm 1 step $\gamma \leq \frac{1}{L_{\max}}$, and for Algorithm 1 $\gamma \leq \frac{1}{HL_{\max}}$, but in practice one can use the same steps (learning rates) for both Algorithms. It seems natural that Algorithm 3 can outperform Algorithm 1 in some regimes, simply because it takes more steps (see Section 6).

· As noted in Section 1.1, there are two more methods of the Local SGD type for SPPs [25, 26]. But these methods use Descent-Ascent instead of Extra Step as a base. Also, the stepsize of these methods is confusing, even in the strongly convex-strongly concave case, it is proposed to take $\gamma = \frac{\mu}{HL_{\max}^2}$, which in practice is a very small number and provide a very slow convergence of the methods.

6. Experiments

6.1. Bilinear problem

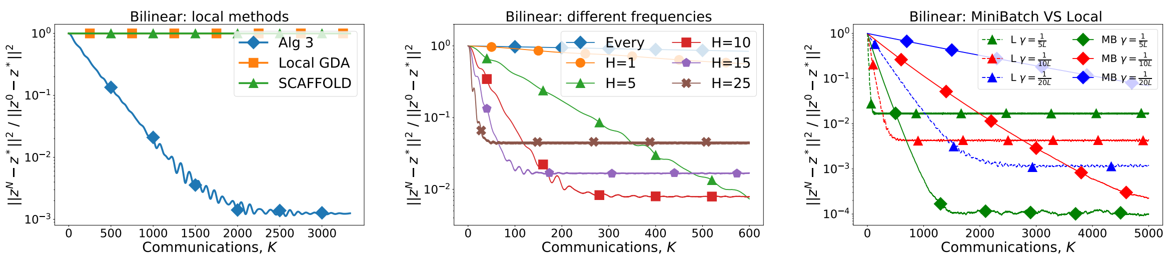
Let us start with an experiment on the bilinear problem:

$$\min_{x,y 2[-1;1]^n} \max \frac{1}{M} \sum_{m=1}^{M} \left(x^T A_m y + b_m^T x + c_m^T y \right), \tag{14}$$

where $n = 100$, $M = 100$, matrices A_m 0 are randomly generated with $\lambda_{\max} = 1000$ (then $L = 1000$). Coordinates b_m, c_m are generated uniformly on $[-1000; 1000]$. Moreover, we add noise with $\sigma^2 = 10000$ to the gradients. Starting point is zero.

The purpose of the first experiment is to compare our local method (Algorithm 3) with the local approaches from papers [25, 26]. For all methods $H = 3$, and the step is chosen for best convergence. See Figure 1 (a) for the results. Note that our Algorithm 3 outperforms the competitors. Moreover, methods from papers [25, 26] do not converge at all with any steps γ. As noted above (Section 1.1), this is due to the fact that these methods are based on Descent-Ascent.

The next experiment is aimed at comparing Algorithm 3 with different communication frequencies H. We take $\gamma = \frac{1}{15L}$. From the point of view of communications (Figure 1



(a) (b) (c)

Figure 1: (a) Comparison of Algorithm 3 and [25, 26] with $H = 3$ and tuned steps; (b) Comparison of Algorithm 3 with different communication frequencies H, as well as Algorithm 1 with batch size 1 (blue line – "Every") for (14); (c) Comparison of Algorithm 3 (L) with communication frequencies $H = 3$ and Algorithm 1 (MB) with batch size 6 for (14)

(b)), we get a standard result for local methods: less often communications, the faster convergence (in communications), but worse solution accuracy. This is due to fluctuations during local iterations, which lead away from the solution of the global solution.

In the third experiment, we want to vary the step and compare Algorithm 3 with a frequency of 3 and Algorithm 1 with batch 6 (such parameters give that there are 6 local calls for one communication for both Algorithms). This problem statement is interesting because Algorithm 1 is optimal, but Algorithm 3 is not, but it can be better in practice. We see (Figure 1 (c)) that the local method wins in rate, but loses in extreme accuracy.

6.2. Federated GANs

Model, data, optimizer A very popular enhancement of GANs is Conditional GAN, originally proposed in [52]. It allows to direct the generation process by introducing class labels. We use a more complex Deep Convolutional GAN [53] with adjustments allowing to condition the output by class labels. We consider the CIFAR-10 [54] and split the dataset into 4 parts. For each part, we select 2 majors class that forms 30% of the data, while the rest of the data split is filled uniformly by the other classes. As optimizers we use Algorithm 3 and Local Adam [55] - a variation of Algorithm 3, but where the local gradient steps are replaced with Adam updates.

Setting Here we would like to consider the experiment of Federated Learning. Communication is a strong bottleneck of federated setting, since data is the local data of users on their devices, and they may simply not be online for transmitting information. Therefore, the reducing communications is our goal, this is what local methods are needed for. Then we want to compare how our optimizers work with a different number of local steps. In particular, we try to communicate once in an epoch, once in 5 epochs and once in 10 epochs. It is interesting to check how the frequency of communication will affect the quality of training.

Results Based on the results of experiments on bilinear problems (Section 6.1), it was expected that methods which connect to the server less frequently (but do the same number of local epochs) would outperform their competitors in terms of communication budget. This trend is observed in Figures 2 and 3 – methods making fewer communications do not lose in terms of FID and IS. Meanwhile the strongly increasing distance between the communications can affect the quality of training considerably, especially in the last epochs. Therefore, we recommend using local methods with long gap between communications only in the initial stages of training, then it is worth communicating more and more frequently.

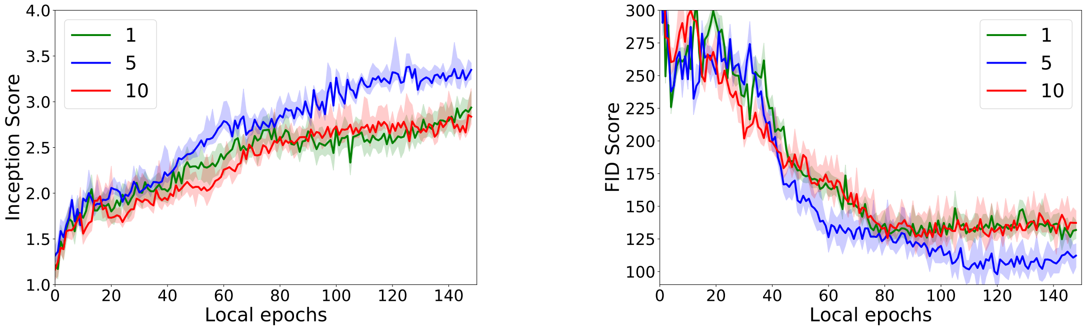
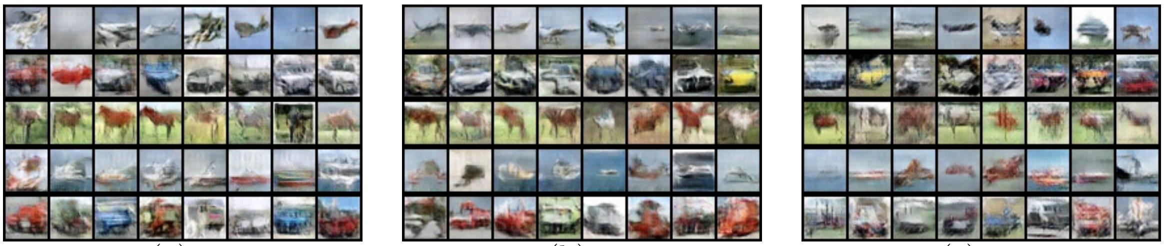


Figure 2: Comparison of three distance between communications in Local Adam in DCGAN distributed decentralized learning on CIFAR-10. We compare the FID Score and the Inception Score in terms of the local epochs number. The experiment was repeated 3 times on different data random splitting - the maximum and minimum deviations are on the plots.



(a) 1 (b) 5 (c) 10

Figure 3: Pictures generated by DSGAN trained distributed on different distance between communications: (a) 1, (b) 5, (c) 10 epochs.

7. Conclusion and future work

The paper derives lower bounds for deterministic and stochastic saddle point problems in distributed centralized and decentralized (fixed networks) setups. An interesting issue is the lower estimates for decentralized problems on time-varying networks.

We also give near-optimal algorithms that achieve lower bounds up to logarithmic factors. For future research, the question of obtaining optimal algorithms without additional logarithmic factors is important.

We present a centralized method with local steps, which is not optimal in theory but is more robust in practice. It is an interesting task to create a decentralized method with local steps.

Bibliography

1. Shalev-Shwartz S., Ben-David S. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.
2. McDonald R., Hall K., Mann G. Distributed training strategies for the structured perceptron // Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics. 2010. P. 456–464.
3. McMahan B., Moore E., Ramage D. et al. Communication-efficient learning of deep networks from decentralized data // Artificial Intelligence and Statistics / PMLR. 2017. P. 1273–1282.
4. Facchinei F., Pang J. Finite-Dimensional Variational Inequalities and Complementarity Problems. Springer Series in Operations Research and Financial Engineering. Springer New York, 2007. ISBN: 9780387218151. URL: https://books.google.ru/books?id=lX_7Rce3_Q0C.
5. Chambolle A., Pock T. A first-order primal-dual algorithm for convex problems with applications to imaging // Journal of mathematical imaging and vision. 2011. Vol. 40, no. 1. P. 120–145.
6. Esser E., Zhang X., Chan T. F. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science // SIAM Journal on Imaging Sciences. 2010. Vol. 3, no. 4. P. 1015–1046.
7. Abadeh S., Esfahani P., Kuhn D. Distributionally robust logistic regression // Advances in Neural Information Processing Systems (NeurIPS)). 2015. P. 1576–1584.
8. Jin Y., Sidford A. Efficiently Solving MDPs with Stochastic Mirror Descent // Proceedings of the 37th International Conference on Machine Learning / Ed. by H. D. III, A. Singh. Vol. 119 of Proceedings of Machine Learning Research. PMLR, 2020. "— 13–18 Jul. P. 4890–4900.
9. Madry A., Makelov A., Schmidt L. et al. Towards Deep Learning Models Resistant to Adversarial Attacks // International Conference on Learning Representations (ICLR). 2018.
10. Goodfellow I. J., Pouget-Abadie J., Mirza M. et al. Generative Adversarial Networks. 2014. arXiv:stat.ML/1406.2661.
11. Daskalakis C., Ilyas A., Syrgkanis V., Zeng H. Training gans with optimism // arXiv

preprint arXiv:1711.00141. 2017.

12. Gidel G., Berard H., Vignoud G. et al. A variational inequality perspective on generative adversarial networks // arXiv preprint arXiv:1802.10551. 2018.
13. Mertikopoulos P., Lecouat B., Zenati H. et al. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile // arXiv preprint arXiv:1807.02629. 2018.
14. Chavdarova T., Gidel G., Fleuret F., Lacoste-Julien S. Reducing noise in gan training with variance reduced extragradient // arXiv preprint arXiv:1904.08598. 2019.
15. Liang T., Stokes J. Interaction Matters: A Note on Non-asymptotic Local Convergence of Generative Adversarial Networks // Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics / Ed. by K. Chaudhuri, M. Sugiyama. Vol. 89 of Proceedings of Machine Learning Research. PMLR, 2019. "— 16–18 Apr. P. 907–915. URL: https://proceedings.mlr.press/v89/liang19b.html.
16. Peng W., Dai Y.-H., Zhang H., Cheng L. Training GANs with centripetal acceleration // Optimization Methods and Software. 2020. Vol. 35, no. 5. P. 955–973.
17. Gallager R. G., Humblet P. A., Spira P. M. A distributed algorithm for minimum-weight spanning trees // ACM Transactions on Programming Languages and systems (TOPLAS). 1983. Vol. 5, no. 1. P. 66–77.
18. Korpelevich G. M. The extragradient method for finding saddle points and other problems. 1976.
19. Juditsky A., Nemirovskii A. S., Tauvel C. Solving variational inequalities with Stochastic Mirror-Prox algorithm. 2008. arXiv:math.OC/0809.0815.
20. Liu J., Morse A. S. Accelerated linear iterations for distributed averaging // Annual Reviews in Control. 2011. Vol. 35, no. 2. P. 160–165.
21. McDonald R., Hall K., Mann G. Distributed Training Strategies for the Structured Perceptron // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, California: Association for Computational Linguistics, 2010. "— . P. 456–464. URL: https://aclanthology.org/N10-1069.
22. Stich S. U. Local SGD converges fast and communicates little // arXiv preprint arXiv:1805.09767. 2018.
23. Konečnỳ J., McMahan H. B., Yu F. X. et al. Federated learning: Strategies for improving communication efficiency // arXiv preprint arXiv:1610.05492. 2016.

24. Kairouz P., McMahan H. B., Avent B. et al. Advances and open problems in federated learning // arXiv preprint arXiv:1912.04977. 2019.

25. Deng Y., Mahdavi M. Local Stochastic Gradient Descent Ascent: Convergence Analysis and Communication Efficiency // International Conference on Artificial Intelligence and Statistics / PMLR. 2021. P. 1387–1395.

26. Hou C., Thekumparampil K. K., Fanti G., Oh S. Efficient Algorithms for Federated Saddle Point Optimization // arXiv preprint arXiv:2102.06333. 2021.

27. Goodfellow I. Nips 2016 tutorial: Generative adversarial networks // arXiv preprint arXiv:1701.00160. 2016.

28. Minty G. J. Monotone (nonlinear) operators in Hilbert space // Duke Mathematical Journal. 1962. Vol. 29, no. 3. P. 341 – 346. URL: https://doi.org/10.1215/S0012-7094-62-02933-2.

29. Diakonikolas J., Daskalakis C., Jordan M. Efficient methods for structured nonconvex-nonconcave min-max optimization // International Conference on Artificial Intelligence and Statistics (AISTATS) / PMLR. 2021. P. 2746–2754.

30. Liu W., Mokhtari A., Ozdaglar A. et al. A decentralized proximal point-type method for saddle point problems // arXiv preprint arXiv:1910.14380. 2019.

31. Liu M., Zhang W., Mroueh Y. et al. A Decentralized Parallel Algorithm for Training Generative Adversarial Nets // arXiv preprint arXiv:1910.12999. 2019.

32. Ben-Tal A., Nemirovski A. Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications. 2019.

33. Nemirovski A. Prox-Method with Rate of Convergence O(1/t) for Variational Inequalities with Lipschitz Continuous Monotone Operators and Smooth Convex-Concave Saddle Point Problems // SIAM Journal on Optimization. 2004. "— 01. Vol. 15. P. 229–251.

34. Nesterov Y. Dual extrapolation and its applications to solving variational inequalities and related problems // Mathematical Programming. 2007. Vol. 109, no. 2. P. 319–344.

35. Hsieh Y.-G., Iutzeler F., Malick J., Mertikopoulos P. On the convergence of single-call stochastic extra-gradient methods. 2019. arXiv:math.OC/1908.08465.

36. Tseng P. A Modified Forward-Backward Splitting Method for Maximal Monotone Mappings // SIAM Journal on Control and Optimization. 2000. Vol. 38, no. 2. P. 431–446. https://doi.org/10.1137/S0363012998338806. URL: https://doi.org/

10.1137/S0363012998338806.

37. Zhang J., Hong M., Zhang S. On lower iteration complexity bounds for the saddle point problems // arXiv preprint arXiv:1912.07481. 2019.
38. Ouyang Y., Xu Y. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems // Mathematical Programming. 2019. P. 1–35.
39. Nemirovskij A. S., Yudin D. B. Problem complexity and method efficiency in optimization. 1983.
40. Foster D. J., Sekhari A., Shamir O. et al. The complexity of making the gradient small in stochastic convex optimization // Conference on Learning Theory / PMLR. 2019. P. 1319–1345.
41. Scaman K., Bach F., Bubeck S. et al. Optimal algorithms for smooth and strongly convex distributed optimization in networks // arXiv preprint arXiv:1702.08704. 2017.
42. Arjevani Y., Shamir O. Communication complexity of distributed convex learning and optimization // arXiv preprint arXiv:1506.01900. 2015.
43. Mukherjee S., Chakraborty M. A decentralized algorithm for large scale min-max problems // 2020 59th IEEE Conference on Decision and Control (CDC). 2020. P. 2967–2972.
44. Rogozin A., Beznosikov A., Dvinskikh D. et al. Decentralized Distributed Optimization for Saddle Point Problems // arXiv preprint arXiv:2102.07758. 2021.
45. Kempe D., Dobra A., Gehrke J. Gossip-based computation of aggregate information // 44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings. / IEEE. 2003. P. 482–491.
46. Boyd S., Ghosh A., Prabhakar B., Shah D. Randomized gossip algorithms // IEEE transactions on information theory. 2006. Vol. 52, no. 6. P. 2508–2530.
47. Nedic A., Ozdaglar A. Distributed subgradient methods for multi-agent optimization // IEEE Transactions on Automatic Control. 2009. Vol. 54, no. 1. P. 48–61.
48. Ye H., Luo L., Zhou Z., Zhang T. Multi-consensus decentralized accelerated gradient descent // arXiv preprint arXiv:2005.00797. 2020.
49. Nesterov Y. Introductory lectures on convex optimization: A basic course. Springer Science & Business Media, 2013. Vol. 87.
50. Song Z., Shi L., Pu S., Yan M. Optimal Gradient Tracking for Decentralized Optimization // arXiv preprint arXiv:2110.05282. 2021.

51. Barazandeh B., Tarzanagh D. A., Michailidis G. Solving a class of non-convex min-max games using adaptive momentum methods // ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) / IEEE. 2021. P. 3625–3629.
52. Mirza M., Osindero S. Conditional generative adversarial nets // arXiv preprint arXiv:1411.1784. 2014.
53. Radford A., Metz L., Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks // arXiv preprint arXiv:1511.06434. 2015.
54. Krizhevsky A., Nair V., Hinton G. CIFAR-10 (Canadian Institute for Advanced Research). URL: `http://www.cs.toronto.edu/~kriz/cifar.html`.
55. Kingma D. P., Ba J. Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980. 2014.
56. Stich S. U. Unified optimal analysis of the (stochastic) gradient method // arXiv preprint arXiv:1907.04232. 2019.

8. General facts and technical lemmas

Lemma 1. *For arbitrary integer $n \geq 1$ and arbitrary set of positive numbers $a_1, \dots, a_n$ we have*

$$\left(\sum_{i=1}^{m} a_i\right)^2 \leq m \sum_{i=1}^{m} a_i^2. \tag{15}$$

Lemma 2. *Suppose given a convex closed set Z, then the operator of the Euclidean projection onto this set is non-expansive, i.e. for all $z, z' \in Z$,*

$$\|proj_Z(z) - proj_Z(z')\| \leq \|z - z'\|. \tag{16}$$

9. Proof of Theorems from Section 3

As mentioned in the main part of the paper we consider the following model of functions:

$$f_m(x, y) = f_m^{deter}(x^{deter}, y) + f^{stoch}(x^{stoch}). \tag{17}$$

Note that the function f_m^{deter} uses the vector x^{deter}, and the function f^{stoch} uses another vector x^{stoch}. The variables in the vectors x^{deter} and x^{stoch} do not i ntersect, but together x^{deter} and x^{stoch} form a complete vector x, for example, according to the following rule: $x_{2k-1} = x_k^{deter}$ and $x_{2k} = x_k^{stoch}$ for $k = 1, 2 \dots$. At the same time, for f_m^{deter}, we have access to $\nabla_x f_m^{deter}(x, y)$, $\nabla_y f_m^{deter}(x^{deter}, y)$, and for f^{stoch}, to stochastic realizations $\nabla_x f_m^{stoch}(x^{stoch}, \xi)$ that satisfy Assumption 3. Moreover, f_m^{deter} are different for each device, but f^{stoch} is the same.

We take "bad" functions with even $n_{x^{stoch}} = n_{x^{deter}} = n_y = n$. Moreover, n must be taken large enough, as stated in the Theorems.

9.1. Deterministic lower bounds

We begin with deterministic lower bounds. Our example builds on a splitting of the "bad" function for the non-distributed case from [37]. Next, we give an example of functions $f_m^{deter}(x^{deter}, y)$ and their location on the nodes. For simplicity of notation, in this subsection we use $f_m(x, y)$ instead of $f_m^{deter}(x^{deter}, y)$.

We introduce some auxiliary arrangement of functions on the nodes, prove some facts for it, and then present the final "bad" examples and prove the lower bounds.

Let $B \subset V$ – subset of nodes of G. For $d \in \mathbb{N}$ we define $B_d = \{v \in V : d(B,v) \geq d\}$, where $d(B,v)$ – distance between set B and node v. Then we construct the following arrangement of bilinearly functions on nodes:

$$f_m(x,y) = \begin{cases} f_1(x,y) = \frac{M}{2|B_d|} \cdot \frac{L}{2} x^T A_1 y + \frac{\mu}{2}\|x\|^2 - \frac{\mu}{2}\|y\|^2 + \frac{M}{2|B_d|} \cdot \frac{L^2}{2\mu} e_1^T y, & m \in B_d \\ f_2(x,y) = \frac{M}{2|B|} \cdot \frac{L}{2} x^T A_2 y + \frac{\mu}{2}\|x\|^2 - \frac{\mu}{2}\|y\|^2, & m \in B \\ f_3(x,y) = \frac{\mu}{2}\|x\|^2 - \frac{\mu}{2}\|y\|^2, & \text{otherwise} \end{cases}. \quad (18)$$

where $e_1 = (1, 0 \ldots, 0)$ and

$$A_1 = \begin{pmatrix} 1 & 0 & & & & & & \\ & 1 & -2 & & & & & \\ & & 1 & 0 & & & & \\ & & & 1 & -2 & & & \\ & & & & \ldots & \ldots & & \\ & & & & & 1 & -2 & \\ & & & & & & 1 & 0 \\ & & & & & & & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & -2 & & & & & & \\ & 1 & 0 & & & & & \\ & & 1 & -2 & & & & \\ & & & 1 & 0 & & & \\ & & & & \ldots & \ldots & & \\ & & & & & 1 & 0 & \\ & & & & & & 1 & -2 \\ & & & & & & & 1 \end{pmatrix}.$$

In most cases, we want the simplest case with $|B| = |B_d| = 1$.

Lemma 3. *Let Problem (18) be solved by any method that satisfies Definition 2. Then after K communication rounds, only the first $\left\lfloor \frac{K}{d} \right\rfloor$ coordinates of the global output can be non-zero while the rest of the $n - \left\lfloor \frac{K}{d} \right\rfloor$ coordinates are strictly equal to zero.*

Proof: We begin introducing some notation for our proof. Let

$$E_0 := \{0\}, \quad E_K := \text{span}\{e_1, \ldots, e_K\}.$$

Note that, the initialization from Definition (2) gives $\mathcal{M}_m^x = E_0$, $\mathcal{M}_m^y = E_0$.

Suppose that, for some m, $\mathcal{M}_m^x = E_K$ and $\mathcal{M}_m^y = E_K$, at some given time. Let us analyze how $\mathcal{M}_m^x$, $\mathcal{M}_m^y$ can change by performing only local computations.

Firstly, we consider the case when K odd. After one local update, we have the following:

· For machines m which own f_1, it holds

$$\begin{aligned} x &\in \operatorname{span}\{e_1\ ,\ x'\ ,\ A_1 y'\} = E_K, \\ y &\in \operatorname{span}\{e_1\ ,\ y'\ ,\ A_1^T x'\} = E_K, \end{aligned} \tag{19}$$

for given $x', x'' \in \mathcal{M}_m^x$ and $y', y'' \in \mathcal{M}_m^y$. Since A_1 has a block diagonal structure, after one local computation, we have $\mathcal{M}_m^x = E_K$ and $\mathcal{M}_m^y = E_K$. The situation does not change, no matter how many local computations one does.

· For machines m which own f_2, it holds

$$\begin{aligned} x &\in \operatorname{span}\{x'\ ,\ A_2 y'\} = E_{K+1}, \\ y &\in \operatorname{span}\{y'\ ,\ A_2^T x'\} = E_{K+1}, \end{aligned}$$

for given $x' \in \mathcal{M}_m^x$ and $y' \in \mathcal{M}_m^y$. It means that, after local computations (at least one local computation), one has $\mathcal{M}_m^x = E_{K+1}$ and $\mathcal{M}_m^y = E_{K+1}$. Therefore, machines with function f_2 can progress by one new non-zero coordinate.

This means that we constantly have to transfer progress from the group of machines with f_1 to the group of machines with f_2 and back. Initially, all devices have zero coordinates. Further, after at least one local computation, machines with f_1 can receive the first nonzero coordinate (but only the first, the second is not), and the rest of the devices are left with all zeros. Next, we pass the first non-zero coordinate to machines with f_2. To do this, d communication rounds are needed. By doing so, they can make the second coordinate non-zero, and then transfer this progress to the machines with f_1. Then the process continues in the same way. It remains to note that for this update in the number of non-zero coordinates, we need at least one local calculation for each non-zero coordinate. Note that the local computation budget is sufficient ($T > K$ – see Definition 2). This completes the proof.

Consider the problem with the global objective function:

$$\begin{aligned} f(x,y) &:= \frac{1}{M}\sum_{m=1}^{M} f_m(x,y) = \frac{1}{M}\left(|B_d|\cdot f_1(x,y) + |B|\cdot f_2(x,y) + (M - |B_d| - |B|)\cdot f_3(x,y)\right) \\ &= \frac{L}{2}x^T A y + \frac{\mu}{2}\|x\|^2 - \frac{\mu}{2}\|y\|^2 + \frac{L^2}{4\mu}e_1^T y, \quad \text{with} \quad A = \frac{1}{2}(A_1 + A_2) \end{aligned} \tag{20}$$

With the fact that $\|A\| \leq 2$, one can easy verify that (20) satisfies Assumptions 1(g) and 2(sc).

The previous lemma gives an idea of what the solution obtained using procedures that satisfy Definition 2. The next lemma is already to the approximate solution of the problem (20) and how it is closed to the real solution.

Lemma 4 (Lemma 3.3 from [37]) *Let* $\alpha = \frac{4\mu^2}{L^2}$ *and* $q = \frac{1}{2}\left(2 + \alpha - \sqrt{\alpha^2 + 4\alpha}\right) \in (0;1)$ – *the smallest root of* $q^2 - (2+\alpha)q + 1 = 0$, *and let introduce approximation* $\bar{y}^*$

$$\bar{y}^*_i = \frac{q^i}{1-q}. \tag{21}$$

Then error between approximation and real solution of (20) *can be bounded*

$$\|\bar{y}^* - y^*\| \leq \frac{q^{n+1}}{\alpha(1-q)}.$$

Proof: Let us write the dual function for (20):

$$g(y) = -\frac{1}{2}y^T\left(\frac{L^2}{4\mu}A^TA + \mu I\right)y + \frac{L^2}{4\mu}e_1^T y.$$

where one can easy found

$$AA^T = \begin{pmatrix} 1 & -1 & & & & & & \\ -1 & 2 & -1 & & & & & \\ & -1 & 2 & -1 & & & & \\ & & -1 & 2 & -1 & & & \\ & & & -1 & 2 & -1 & & \\ & & & & & \ldots & & \\ & & & & & -1 & 2 & -1 \\ & & & & & & -1 & 2 \end{pmatrix}.$$

The optimality of the dual problem $\nabla g(y^*) = 0$ gives

$$\left(\frac{L^2}{4\mu}A^TA + \mu I\right)y^* = \frac{L^2}{4\mu}e_1,$$

or

$$\left(A^T A + \alpha I\right) y^* = e_1.$$

Let us write in the form of a set of equations:

$$\begin{cases} (1+\alpha)y_1^* - y_2^* = 1 \\ -y_1^* + (2+\alpha)y_2^* - y_3^* = 0 \\ \dots \\ -y_{n-2}^* + (2+\alpha)y_{n-1}^* - y_n^* = 0 \\ -y_{n-1}^* + (2+\alpha)y_n^* = 0 \end{cases}$$

Note that the approximation (21) satisfies the following set of equations:

$$\begin{cases} (1+\alpha)\bar{y}_1^* - \bar{y}_2^* = 1 \\ -\bar{y}_1^* + (2+\alpha)\bar{y}_2^* - \bar{y}_3^* = 0 \\ \dots \\ -\bar{y}_{n-2}^* + (2+\alpha)\bar{y}_{n-1}^* - \bar{y}_n^* = 0 \\ -\bar{y}_{n-1}^* + (2+\alpha)\bar{y}_n^* = \frac{q^{n+1}}{1-q} \end{cases}$$

or in the short form:

$$\left(A^T A + \alpha I\right) \bar{y}^* = e_1 + \frac{q^{n+1}}{1-q} e_n.$$

Then the difference between the approximation and the true solution is

$$\bar{y}^* - y^* = \left(A^T A + \alpha I\right)^{-1} \frac{q^{n+1}}{1-q} e_n,$$

With the fact that $\alpha^{-1} I \quad \left(A^T A + \alpha I\right)^{-1} \quad 0$, it implies the statement of Lemma.

Now we formulate a key lemma (similar to Lemma 3.4 from [37]).

Lemma 5. *Consider a distributed saddle point problem in form* (18),(20) *with* $B_d \neq$?. *For any pairs* T, K *(*$T > K$*) one can found size of the problem* $n \geq \max\left\{2\log_q\left(\frac{\alpha}{4\sqrt{2}}\right), 2K\right\}$, *where* $\alpha = \frac{4\mu^2}{L^2}$ *and* $q = \frac{1}{2}\left(2 + \alpha - \sqrt{\alpha^2 + 4\alpha}\right) \in (0;1)$. *Then, any output* $\hat{x}, \hat{y}$ *produced by any* $\mathrm{BBP}(T, K)$ *satisfying Definition 2 after* K *communications rounds and* T *local computations, is such that*

$$\|\hat{x} - x^*\|^2 + \|\hat{y} - y^*\|^2 \geq q^{\frac{2K}{d}} \frac{\|y_0 - y^*\|^2}{16}.$$

Proof: Lemma 3 states that after K $(K < T)$ communications only $k = \left\lfloor \frac{K}{d} \right\rfloor$ coordinates in the output $\hat{y}$ can be non-zero. Therefore, by definition of $\bar{y}^*$ from (21), by $k \leq K \leq \frac{n}{2}$ and with $q < 1$, we have

$$\|\hat{y} - \bar{y}^*\|^2 \geq \sqrt{\sum_{j=k+1}^{n} (\bar{y}_j^*)^2} = \frac{q^k}{1-q}\sqrt{q^2 + q^4 + \ldots + q^{2(n-k)}}$$

$$\geq \frac{q^k}{\sqrt{2}(1-q)}\sqrt{q^2 + q^4 + \ldots + q^{2n}} = \frac{q^k}{\sqrt{2}}\|\bar{y}^*\|^2 = \frac{q^k}{\sqrt{2}}\|y_0 - \bar{y}^*\|^2.$$

For $n \geq 2\log_q\left(\frac{\alpha}{4\sqrt{2}}\right)$ we can guarantee that $\bar{y}^* \approx y^*$ (for more detailed see [37]) and

$$\|\hat{x} - x^*\|^2 + \|\hat{y} - y^*\|^2 \geq \|\hat{y} - y^*\|^2 \geq \frac{q^{2k}}{16}\|y_0 - y^*\|^2 = q^{2\lfloor \frac{K}{d} \rfloor}\frac{\|y_0 - y^*\|^2}{16} \geq q^{\frac{2K}{d}} \cdot \frac{\|y_0 - y^*\|^2}{16}.$$

Building on the above preliminary results, we are now ready to prove our complexity lower bound as stated in Theorems 1 and 2.

Centralized case

Theorem 8 (Theorem 1) *For any* $L > \mu > 0$, *any* Δ *and any* $T, K \in \mathbb{N}$ *with* $T > K$, *there exists a distributed saddle point problem of* $\Delta+1$ *functions with centralized architecture. For which the following statements are true:*

- *the diameter of graph* G *is equal to* Δ,
- $f = \frac{1}{M}\sum_{m=1}^{M} f_m : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *is* L*-smooth,* μ *– strongly convex-strongly concave,*
- *size* $n \geq \max\left\{2\log_q\left(\frac{\alpha}{4\sqrt{2}}\right), 2K\right\}$, *where* $\alpha = \frac{4\mu^2}{L^2}$ *and* $q = \frac{1}{2}\left(2 + \alpha - \sqrt{\alpha^2 + 4\alpha}\right) \in (0;1)$.

Then for any output $\hat{x}, \hat{y}$ *of any* $\mathrm{BBP}(T, K)$ *(Definition 2), the following estimates hold:*

$$\|\hat{x} - x^*\|^2 + \|\hat{y} - y^*\|^2 = \Omega\left(\exp\left(-\frac{4\mu}{L-\mu} \cdot \frac{K}{\Delta}\right)\|y_0 - y^*\|^2\right).$$

Proof: It suffices to consider a linear graph on $\Delta + 1$ vertices $\{v_1, \ldots, v_{\Delta+1}\}$ and apply Lemma 5 for problem (18),(20) with $B = \{v_1\}$ and $d = \Delta$. Then

$$\left(\frac{1}{q}\right)^{\frac{2K}{\Delta}} \geq \frac{\|y_0 - y^*\|^2}{16(\|\hat{x} - x^*\|^2 + \|\hat{y} - y^*\|^2)}.$$

Taking the logarithm of the two parts of the inequality, we get

$$\frac{2K}{\Delta} \geq \ln\left(\frac{\|y_0 - y^*\|^2}{16(\|\hat{x} - x^*\|^2 + \|\hat{y} - y^*\|^2)}\right)\frac{1}{\ln(q^{-1})}.$$

Next, we work with

$$\frac{1}{\ln(q^{-1})} = \frac{1}{\ln(1 + (1-q)/q))} \geq \frac{q}{1-q} = \frac{1 + \frac{2\mu^2}{L^2} - 2\sqrt{\frac{\mu^2}{L^2} + \left(\frac{\mu^2}{L^2}\right)^2}}{2\sqrt{\frac{\mu^2}{L^2} + \left(\frac{\mu^2}{L^2}\right)^2} - \frac{2\mu^2}{L^2}}$$

$$= \frac{2\sqrt{\frac{\mu^2}{L^2} + \left(\frac{\mu^2}{L^2}\right)^2} - \frac{2\mu^2}{L^2}}{\frac{4\mu^2}{L^2}}$$

$$= \frac{1}{2}\sqrt{\frac{L^2}{\mu^2} + 1} - \frac{1}{2}.$$

Finally, one can obtain

$$\frac{2K}{\Delta} \geq \ln\left(\frac{\|y_0 - y^*\|^2}{16(\|\hat{x} - x^*\|^2 + \|\hat{y} - y^*\|^2)}\right) \cdot \frac{1}{2}\left(\frac{L}{\mu} - 1\right),$$

and

$$\exp\left(\frac{4\mu}{L-\mu}\frac{K}{\Delta}\right) \geq \frac{\|y_0 - y^*\|^2}{16(\|\hat{x} - x^*\|^2 + \|\hat{y} - y^*\|^2)},$$

which completes the proof.

Decentralized case

Theorem 9. *For any $L > \mu > 0$, any Δ and any $T, K \in \mathbb{N}$ with $T > K$, there exists a distributed saddle point problem with decentralized architecture and a gossip matrix W. For which the following statements are true:*

- *a gossip matrix W have $\chi(W) = \chi$,*
- *$f = \frac{1}{M}\sum\limits_{m=1}^{M} f_m : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is L-smooth, μ – strongly convex-strongly concave,*

- *size* $n \geq \max\left\{2\log_q\left(\frac{\alpha}{4\sqrt{2}}\right), 2K\right\}$, *where* $\alpha = \frac{4\mu^2}{L^2}$ *and* $q = \frac{1}{2}\left(2+\alpha-\sqrt{\alpha^2+4\alpha}\right)$.

Then for any output $\hat{x}, \hat{y}$ *of any* BBP(T, K) *(Definition 2), the following estimates hold:*

$$\|\hat{x}-x^*\|^2+\|\hat{y}-y^*\|^2 = \Omega\left(\exp\left(-\frac{32\mu}{L-\mu}\cdot\frac{K}{\sqrt{\chi}}\right)\|y_0-y^*\|^2\right).$$

Proof: The proof follow similar steps as in the proof of [41, Theorem 2]. Let $\gamma_M = \frac{1-\cos\frac{\pi}{M}}{1+\cos\frac{\pi}{M}}$ be a decreasing sequence of positive numbers. Since $\gamma_2 = 1$ and $\lim_m \gamma_M = 0$, there exists $M \geq 2$ such that $\gamma_M \geq \frac{1}{\chi} > \gamma_{M+1}$.

· If $M \geq 3$, let us consider linear graph of size M with vertexes $v_1, \ldots v_M$, and weighted with $w_{1,2} = 1-a$ and $w_{i,i+1} = 1$ for $i \geq 2$. We apply Lemma 5 for problem (18),(20) with $B = \{v_1\}$ and $d = M-1$, then we have $B_d = \{v_M\}$. Hence,

$$\|\hat{x}-x^*\|^2+\|\hat{y}-y^*\|^2 \geq q^{\frac{2K}{d}}\frac{\|y_0-y^*\|^2}{16}.$$

If W_a is the Laplacian of the weighted graph G, one can note that with $a=0$, $\frac{1}{\chi(W_a)} = \gamma_M$, with $a=1$, we have $\frac{1}{\chi(W_a)} = 0$. Hence, there exists $a \in (0;1]$ such that $\frac{1}{\chi(W_a)} = \chi$. Then $\frac{1}{\chi} \geq \gamma_{M+1} \geq \frac{2}{(M+1)^2}$, and $M \geq \sqrt{2\chi}-1 \geq \frac{\sqrt{\chi}}{4}$. Finally, since $M \geq 3$, we get $d = M-1 \geq \frac{M}{2} \geq \frac{\sqrt{\chi}}{8}$. Hence,

$$\|\hat{x}-x^*\|^2+\|\hat{y}-y^*\|^2 \geq q^{\frac{16K}{\sqrt{\chi}}}\frac{\|y_0-y^*\|^2}{16}.$$

Similarly to the proof of the previous theorem

$$\exp\left(\frac{32\mu}{L-\mu}\frac{K}{\sqrt{\chi}}\right) \geq \frac{\|y_0-y^*\|^2}{16(\|\hat{x}-x^*\|^2+\|\hat{y}-y^*\|^2)}. \tag{22}$$

· If $M = 2$, we construct a fully connected network with 3 nodes with weight $w_{1,3} = a \in [0;1]$. Let W_a is the Laplacian. If $a = 0$, then the network is a linear graph and $\rho(W_a) = \gamma_3 = \frac{1}{3}$. Hence, there exists $a \in [0;1]$ such that $\chi(W_a) = \chi$. Finally, $B = \{v_1\}$, $B_d = \{v_3\}$ and $d \geq 1 \geq \frac{\sqrt{\chi}}{2}$. Whence, it follows that in this case (22) is also valid.

9.2. Stochastic lower bounds

Strongly-convex case

We consider the following simple problem with function $f : \mathbb{R} \to \mathbb{R}$:

$$\min_{x\in\mathbb{R}} f(x) = \frac{\mu}{2}(x-x_0)^2, \tag{23}$$

where we do not know the constant $x_0 \neq 0$. $f(x)$ is a μ-strongly-convex and μ-smooth function. We minimize this function by using stochastic first order oracle

$$\nabla f(x, \xi) = \mu(x + \xi - x_0), \text{ where } \xi \in \mathcal{N}\left(0, \frac{\sigma^2}{\mu^2}\right).$$

One can note that $\mathsf{E}[\nabla f(x,\xi)] = \mu(x - x_0) = \nabla f(x)$, and $\mathsf{E}\left[|\nabla f(x,\xi) - \nabla f(x)|^2\right] = \mathsf{E}\left[\mu^2|\xi|^2\right] = \sigma^2$. We use some $\mathsf{BBP}(T, K)$ (Definition 2), which calls the stochastic oracle $N = MT$ times in some set of points $\{x_i\}_{i=1}^N$, for these points oracle returns $y_i = \mu(x_i - x_0 + \xi_i)$, where all $\xi_i \in \mathcal{N}(0, \sigma^2/\mu^2)$ and independent. Using x_i, y_i, one can compute point $z_i = x_i - y_i/\mu = x_0 - \xi_i \in \mathcal{N}(x_0, \sigma^2/\mu^2)$ and independent. Hence, the original problem (23) and the working of any BBP are easy to reformulate in the following way: after N calls of the oracle we have set of pairs $\{(x_i, z_i)\}_{i=1}^N$, where $z_i \in \mathcal{N}(x_0, \sigma^2/\mu^2)$ and independent. By these pairs we need to estimate the unknown constant x_0. One can do it by MLE:

$$x_N^{\mathsf{MLE}} = \frac{1}{N}\sum_{i=1}^{N} z_i, \quad x_N^{\mathsf{MLE}} \in \mathcal{N}\left(x_0, \frac{\sigma^2}{N\mu^2}\right).$$

Then

$$\mathsf{E}\left[\|x_N^{\mathsf{MLE}} - x^*\|^2\right] = \mathsf{E}\left[|x_N^{\mathsf{MLE}} - x_0|^2\right] = \mathrm{Var}\left[x_N^{\mathsf{MLE}}\right] = \frac{\sigma^2}{N\mu^2},$$

or

$$\mathsf{E}\left[f(x_N^{\mathsf{MLE}}) - f(x^*)\right] = \frac{\mu}{2}\mathsf{E}\left[|x_N^{\mathsf{MLE}} - x_0|^2\right] = \frac{\mu}{2}\mathrm{Var}\left[x_N^{\mathsf{MLE}}\right] = \frac{\sigma^2}{2N\mu}.$$

We need to show that the estimate obtained with the MLE is the best in terms of N, for this we use the Cramer–Rao bound:

Lemma 6. *Suppose x_0 is an unknown parameter which is to be estimated from $\{z_i\}_{i=1}^N$ independent observations of z, each distributed according to some probability density function $f_{x_0}(z)$. Consider an estimator $\hat{x}_0 = \hat{x}_0(z_1, \ldots, z_N)$ with bias $b(x_0) = \mathsf{E}[\hat{x}_0 - x_0]$. Thus, the estimator $\hat{x}_0$ satisfies*

$$\mathsf{E}\left[|\hat{x}_0 - x_0|^2\right] \geq \frac{[1 + b'(x_0)]^2}{NI(x_0)} + b^2(x_0),$$

where $I(x_0) = \mathsf{E}\left[\left(\frac{\partial \ln f_{x_0}(z)}{\partial x_0}\right)^2\right]$ – Fisher information.

For normal distribution $I(x_0) = \frac{\mu^2}{\sigma^2}$. Then

$$\mathsf{E}\left[|\hat{x}_0 - x_0|^2\right] \geq \frac{\sigma^2}{N\mu^2}[1 + b'(x_0)]^2 + b^2(x_0)$$

Suppose that there is such estimate that it is better than the MLE in terms of N. Hence,

$$[1 + b'(x_0)]^2 \quad \frac{1}{N^{2\alpha}}, \text{ where } \alpha > 0.$$

It means $-b'(x_0) \quad \frac{N^\alpha - 1}{N^\alpha}$ or $\frac{N^\alpha + 1}{N^\alpha}$. With enough big N we get that $b'(x_0) \quad -1$ in terms of N. Then $b^2(x_0) \quad x_0^2$. We arrive at a contradiction in the existence of an estimate that asymptotically (in N) better than the MLE. Then we have the following theorem:

Theorem 10 (Theorem 3) *For any $L > \mu > 0$ and any $M, T \in \mathbb{N}$, there exists a stochastic minimization problem with L-smooth and μ-strongly convex function such that for any output $\hat{x}$ of any* BBP(T, K) *(Definition 2) with M workers one can obtain the following estimate:*

$$\|\hat{x} - x^*\|^2 = \Omega\left(\frac{\sigma^2}{MT\mu^2}\right).$$

Convex case

For convex case, we work with

$$\min_{x \in [-\frac{\Omega_x}{2}, \frac{\Omega_x}{2}]} \frac{\tilde{\varepsilon}}{\Omega_x} x, \tag{24}$$

where $\tilde{\varepsilon}$ can only take two values ε or $-\varepsilon$ with some positive ε. Of course, we do not know which of the two values $\tilde{\varepsilon}$ takes. For example, we can assume that at the very beginning $\tilde{\varepsilon}$ is chosen randomly and equally probable. It is easy to verify that (24) is convex and L-smooth for any L and ε. The first order stochastic oracle returns $\nabla f(x, \xi) = \xi \in \mathcal{N}(\tilde{\varepsilon}/\Omega_x, \sigma^2)$. One can note that $\mathbb{E}[\nabla f(x, \xi)] = \tilde{\varepsilon}/\Omega_x = \nabla f(x)$, and $\mathbb{E}\left[|\nabla f(x, \xi) - \nabla f(x)|^2\right] = \sigma^2$. We use some procedure BBP(T, K) (Definition 2), which calls oracle $N = MT$ times in some set of points $\{x_i\}_{i=1}^N$. For these points oracle returns ξ_i, where all $\xi_i \in \mathcal{N}(\tilde{\varepsilon}, \sigma^2)$ and independent. Note that we can say in advance that $x^* = \frac{\Omega_x}{2}$ if $\tilde{\varepsilon} = -\varepsilon$ and $x^* = -\frac{\Omega_x}{2}$ if $\tilde{\varepsilon} = \varepsilon$. We have a rather simple task, from independent samples $\{\xi_i\}_{i=1}^N \in \mathcal{N}(\tilde{\varepsilon}/\Omega_x, \sigma^2)$, we need to determine $\tilde{\varepsilon}$ from two equally probable hypothesis $H_1 : \tilde{\varepsilon} = \varepsilon$ or $H_2 : \tilde{\varepsilon} = -\varepsilon$. For these problem one can use likelihood-ratio criterion:

$$\delta(\xi_1, \ldots, \xi_N) = \begin{cases} H_1, & T(\xi_1, \ldots, \xi_N) < c \\ H_2, & T(\xi_1, \ldots, \xi_N) \geq c \end{cases}, \quad T(\xi_1, \ldots, \xi_N) = \frac{f_{H_2}(\xi_1, \ldots, \xi_N)}{f_{H_1}(\xi_1, \ldots, \xi_N)}, \tag{25}$$

where f_H is a density function of a random vector $\xi_1, \ldots, \xi_N$ with distribution from the hypothesis H. The Neyman–Pearson lemma gives

Lemma 7. *There is a constant c for which the likelihood-ratio criterion (25) is*

- *minmax criterion. The number c should be chosen so that the type I error and the type II error were the same;*
- *Bayesian criterion for given prior probabilities r and s. The number c is chosen equal to the ratio r/s.*

Due to the symmetry of the hypotheses with respect to zero, as well as the prior probabilities can be considered equal to $1/2$, we have that $c = 1$ for minmax and Bayesian criterions. By simple transformations we can rewrite (25):

$$\delta(\xi_1, \ldots, \xi_N) = \begin{cases} H_1, & \sum\limits_{i=1}^N \xi_i > 0 \\ H_2, & \sum\limits_{i=1}^N \xi_i \leq 0 \end{cases}, \qquad \hat{x}_N = \begin{cases} -\frac{\Omega_x}{2}, & \sum\limits_{i=1}^N \xi_i > 0 \\ \frac{\Omega_x}{2}, & \sum\limits_{i=1}^N \xi_i \leq 0 \end{cases}.$$

This criterion is more than natural. Neyman–Pearson lemma says it is optimal. Next we analyse error of this criterion (we will consider only case with $\tilde{\varepsilon} = \varepsilon$, the other case one can parse similarly):

$$\mathsf{E}\left[f(\hat{x}_N) - f(x^*)\right] = \mathsf{E}\left[\frac{\varepsilon}{\Omega_x}\hat{x}_N + \frac{\varepsilon}{2}\right] = \varepsilon \cdot \mathsf{P}\left\{\sum_{i=1}^N \xi_i \leq 0\right\} = \varepsilon \cdot \mathsf{P}\left\{S_N \leq 0\right\},$$

where $S_N = \sum\limits_{i=1}^N \xi_i \in \mathcal{N}(\varepsilon N/\Omega_x, \sigma^2 N)$, then $\frac{S_N - \varepsilon N/\Omega_x}{\sigma\sqrt{N}} \in \mathcal{N}(0,1)$. Finally, we get

$$\mathsf{E}\left[f(\hat{x}_N) - f(x^*)\right] = \varepsilon \mathsf{P}\left\{\frac{S_N - \varepsilon N/\Omega_x}{\sigma\sqrt{N}} \leq -\frac{\varepsilon\sqrt{N}}{\Omega_x\sigma}\right\} = \varepsilon \mathsf{P}\left\{S_N \geq \frac{\varepsilon\sqrt{N}}{\Omega_x\sigma}\right\}$$
$$\geq \varepsilon \cdot \frac{1}{3t}\exp\left(-\frac{t^2}{2}\right)\cdot\left(1 - \frac{1}{t^2}\right).$$

In last inequality we define $t = \frac{\varepsilon\sqrt{N}}{\Omega_x\sigma}$ and use lower bound for tail of standard normal distribution. With $\varepsilon = \frac{2\Omega_x\sigma}{\sqrt{N}}$, we have $t = 2$ and then

$$\mathsf{E}\left[f(\hat{x}_N) - f(x^*)\right] \geq \frac{\varepsilon}{4t}\exp(-2) \geq \frac{1}{4\exp(2)} \cdot \frac{\sigma\Omega_x}{\sqrt{N}}.$$

Hence, we get the next theorem:

Theorem 11 (Theorem 4) *For any $L > 0$ and any $M, T \in \mathbb{N}$, there exists a stochastic minimization problem with L-smooth and convex function such that for any output $\hat{x}$ of any $\mathsf{BBP}(T, K)$ (Definition 2) with M workers one can obtain the following estimate:*

$$\mathsf{E}\left[f(\hat{x}) - f(x^*)\right] = \Omega\left(\frac{\sigma\Omega_x}{\sqrt{MT}}\right).$$

10. Proof of Theorems from Section 4

10.1. Centralized case

We start our proof with the following lemma:

Lemma 8. *Let $z, y \in \mathbb{R}^n$ and $Z \subseteq \mathbb{R}^n$ is convex compact set. We set $z^+ = \text{proj}_Z(z - y)$, then for all $u \in Z$:*

$$\|z^+ - u\|^2 \leq \|z - u\|^2 - 2\langle y, z^+ - u\rangle - \|z^+ - z\|^2.$$

Proof: For all $u \in Z$ we have $\langle z^+ - (z - y), z^+ - u\rangle \leq 0$. Then

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

Before proof the main theorems, we add the following notation:

$$\bar{g}^t = \frac{1}{M}\sum_{m=1}^{M} g_m^t, \qquad \bar{g}^{t+1/2} = \frac{1}{M}\sum_{m=1}^{M} g_m^{t+1/2}.$$

Strongly convex-strongly concave problems

Theorem 12 (Theorem 5) *Let $\{z^t\}_{t\geq 0}$ denote the iterates of Algorithm 1 for solving problem (1). Let Assumptions 1(g), 2(sc) and 3 be satisfied. Then, if $\gamma \leq \frac{1}{4L}$, we have the following estimate for the distance to the solution z^*:*

$$\mathbb{E}\left[\|z^k - z^*\|^2\right] = \mathcal{O}\left(\|z^0 - z^*\|^2 \exp\left(-\frac{\mu}{4L}\cdot\frac{K}{\Delta}\right) + \frac{\sigma^2}{\mu^2 MT}\right).$$

Proof: Applying the previous Lemma with $z^+ = z^{t+1}$, $z = z^t$, $u = z$ and $y = \gamma\bar{g}^{t+1/2}$, we get

$$\|z^{t+1} - z\|^2 \leq \|z^t - z\|^2 - 2\gamma\langle\bar{g}^{t+1/2}, z^{t+1} - z\rangle - \|z^{t+1} - z^t\|^2,$$

and with $z^+ = z^{t+1/2}$, $z = z^t$, $u = z^{t+1}$, $y = \gamma\bar{g}^t$:

$$\|z^{t+1/2} - z^{t+1}\|^2 \leq \|z^t - z^{t+1}\|^2 - 2\gamma\langle\bar{g}^t, z^{t+1/2} - z^{t+1}\rangle - \|z^{t+1/2} - z^t\|^2.$$

Next, we sum up the two previous equalities

$$\|z^{t+1} - z\|^2 + \|z^{t+1/2} - z^{t+1}\|^2 \leq \|z^t - z\|^2 - \|z^{t+1/2} - z^t\|^2 \\ -2\gamma\langle \bar{g}^{t+1/2}, z^{t+1} - z\rangle - 2\gamma\langle \bar{g}^t, z^{t+1/2} - z^{t+1}\rangle.$$

A small rearrangement gives

$$\begin{aligned}\|z^{t+1} - z\|^2 + \|z^{t+1/2} - z^{t+1}\|^2 &\leq \|z^t - z\|^2 - \|z^{t+1/2} - z^t\|^2 \\ &\quad - 2\gamma\langle \bar{g}^{t+1/2}, z^{t+1/2} - z\rangle + 2\gamma\langle \bar{g}^{t+1/2} - \bar{g}^t, z^{t+1/2} - z^{t+1}\rangle \\ &\leq \|z^t - z\|^2 - \|z^{t+1/2} - z^t\|^2 \\ &\quad - 2\gamma\langle \bar{g}^{t+1/2}, z^{t+1/2} - z\rangle + \gamma^2\|\bar{g}^{t+1/2} - \bar{g}^t\|^2 + \|z^{t+1/2} - z^{t+1}\|^2. \end{aligned} \tag{26}$$

Then we substitute $z = z^*$ and take the total expectation of both sides of the equation

$$\mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] \leq \mathbb{E}\left[\|z^t - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] \\ -2\gamma\mathbb{E}\left[\langle \bar{g}^{t+1/2}, z^{t+1/2} - z^*\rangle\right] + \gamma^2\mathbb{E}\left[\|\bar{g}^{t+1/2} - \bar{g}^t\|^2\right]. \tag{27}$$

Let work with $\mathbb{E}\left[\|\bar{g}^{t+1/2} - \bar{g}^t\|^2\right]$:

$$\begin{aligned}\mathbb{E}\left[\|\bar{g}^{t+1/2} - \bar{g}^t\|^2\right] &= \mathbb{E}\left[\|\bar{g}^{t+1/2} - F(z^{t+1/2}) + F(z^t) - \bar{g}^t + F(z^{t+1/2}) - F(z^t)\|^2\right] \\ &\overset{(15)}{\leq} 3\mathbb{E}\left[\|\bar{g}^{t+1/2} - F(z^{t+1/2})\|^2\right] + 3\mathbb{E}\left[\|F(z^t) - \bar{g}^t\|^2\right] \\ &\quad + 3\mathbb{E}\left[\|F(z^{t+1/2}) - F(z^t)\|^2\right] \\ &\overset{(3)}{\leq} 3\mathbb{E}\left[\left\|\frac{1}{bM}\sum_{m=1}^{M}\sum_{i=1}^{b}(F_m(z^{t+1/2}, \xi_m^{t+1/2,i}) - F_m(z^{t+1/2}))\right\|^2\right] \\ &\quad + 3\mathbb{E}\left[\left\|\frac{1}{bM}\sum_{m=1}^{M}\sum_{i=1}^{b}(F_m(z^t, \xi_m^{t,i}) - F_m(z^t))\right\|^2\right] \\ &\quad + 3L^2\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] \\ &= \frac{3}{(bM)^2}\mathbb{E}\left[\left\|\sum_{m=1}^{M}\sum_{i=1}^{b}(F_m(z^{t+1/2}, \xi_m^{t+1/2,i}) - F_m(z^{t+1/2}))\right\|^2\right] \\ &\quad + \frac{3}{(bM)^2}\mathbb{E}\left[\left\|\sum_{m=1}^{M}\sum_{i=1}^{b}(F_m(z^t, \xi_m^{t,i}) - F_m(z^t))\right\|^2\right] \\ &\quad + 3L^2\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right]. \end{aligned}$$

Using that all $\{\xi_m^{t,i}\}_{i=1,m=1}^{b,M}$ and $\{\xi_m^{t+1/2,i}\}_{i=1,m=1}^{b,M}$ are independent, we get

$$\mathbb{E}\left[\|\bar{g}^{t+1/2} - \bar{g}^t\|^2\right] \leq \frac{3}{(bM)^2}\sum_{m=1}^{M}\sum_{i=1}^{b}\mathbb{E}\left[\left\|F_m(z^{t+1/2}, \xi_m^{t+1/2,i}) - F_m(z^{t+1/2})\right\|^2\right]$$

$$
\begin{aligned}
&+\frac{3}{(bM)^2}\sum_{m=1}^{M}\sum_{i=1}^{b}\mathbb{E}\left[\left\|F_m(z^t,\xi_m^{t,i}) - F_m(z^t)\right\|^2\right] \\
&+3L^2\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] \\
\overset{(7)}{\leq}\; & 3L^2\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] + \frac{6\sigma^2}{bM}. \qquad (28)
\end{aligned}
$$

Next we estimate $\mathbb{E}\left[\langle \bar{g}^{t+1/2}, z^{t+1/2} - z^*\rangle\right]$. To begin with, we use the independence of all ξ, as well as the unbiasedness of $\bar{g}^{t+1/2}$ with respect to the conditional m.o. by random variables $\{\xi_m^{t+1/2,i}\}_{i=1,m=1}^{b,M}$:

$$
\begin{aligned}
\mathbb{E}\left[\langle \bar{g}^{t+1/2}, z^{t+1/2} - z^*\rangle\right] &= \mathbb{E}\left[\mathbb{E}_{\{\xi_m^{t+1/2,i}\}_{i=1,m=1}^{b,M}}\left[\langle \bar{g}^{t+1/2}, z^{t+1/2} - z^*\rangle\right]\right] \\
&= \mathbb{E}\left[\langle \mathbb{E}_{\{\xi_m^{t+1/2,i}\}_{i=1,m=1}^{b,M}}\left[\bar{g}^{t+1/2}\right], z^{t+1/2} - z^*\rangle\right] \\
&= \mathbb{E}\left[\langle F(z^{t+1/2}), z^{t+1/2} - z^*\rangle\right]. \qquad (29)
\end{aligned}
$$

By property of z^*, we get

$$
\mathbb{E}\left[\langle \bar{g}^{t+1/2}, z^{t+1/2} - z^*\rangle\right] \;\geq\; \mathbb{E}\left[\langle F(z^{t+1/2}) - F(z^*), z^{t+1/2} - z\rangle\right] \geq \mu\mathbb{E}\left[\|z^{t+1/2} - z^*\|^2\right].
$$

Let use a simple fact $\|z^{t+1/2} - z^*\|^2 \geq \frac{1}{2}\|z^t - z^*\|^2 - \|z^{t+1/2} - z^t\|^2$, then

$$
\mathbb{E}\left[\langle \bar{g}^{t+1/2}, z^{t+1/2} - z^*\rangle\right] \geq \frac{\mu}{2}\mathbb{E}\left[\|z^t - z^*\|^2\right] - \mu\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right]. \qquad (30)
$$

Combining 3 inequalities (27) with $z = z^*$, (28), (30):

$$
\mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] \leq (1-\mu\gamma)\mathbb{E}\left[\|z^t - z^*\|^2\right] + (2\mu\gamma + 3\gamma^2L^2 - 1)\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] + \frac{6\sigma^2\gamma^2}{bM}.
$$

In Algorithm 1 the step $\gamma \leq \frac{1}{4L}$, then

$$
\mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] \;\leq\; (1-\mu\gamma)\mathbb{E}\left[\|z^t - z^*\|^2\right] + \frac{6\sigma^2\gamma^2}{bM}.
$$

Let us run the recursion from 0 to $k-1$:

$$
\mathbb{E}\left[\|z^k - z^*\|^2\right] \;\leq\; (1-\mu\gamma)^k\mathbb{E}\left[\|z^0 - z^*\|^2\right] + \frac{6\sigma^2\gamma}{\mu bM}.
$$

Then we carefully choose $\gamma = \min\left\{\frac{1}{4L}; \frac{\ln(\max\{2; bM\mu^2\|z^0 - z^*\|^2 k/6\sigma^2\})}{\mu k}\right\}$ and get (for more details one can see [56])

$$
\mathbb{E}\left[\|z^{k+1} - z^*\|^2\right] = \tilde{O}\left(\|z^0 - z^*\|^2\exp\left(-\frac{\mu k}{4L}\right) + \frac{\sigma^2}{\mu^2 bMk}\right).
$$

Substitute the batch size b and the number of iterations k from the description of the Algorithm 1:

$$
\mathbb{E}\left[\|z^{k+1} - z^*\|^2\right] = \tilde{O}\left(\|z^0 - z^*\|^2\exp\left(-\frac{\mu}{4L}\cdot\frac{K}{r}\right) + \frac{\sigma^2}{\mu^2 MT}\right).
$$

Finally, we remember that $r \leq \Delta$ and finish the proof.

Convex-concave problems

Theorem 13 (Theorem 5) *Let $\{z^t\}_{t\geq 0}$ denote the iterates of Algorithm 1 for solving problem* (1). *Let Assumptions 1(g), 2(c), 3 and 4 be satisfied. Then, if $\gamma \leq \frac{1}{4L}$, we have the following estimate:*

$$\mathbb{E}[gap(z^k_{avg})] = O\left(\frac{L\Omega_z^2 \Delta}{K} + \frac{\sigma\Omega_z}{\sqrt{MT}}\right).$$

Proof: We have already shown some of the necessary estimates, namely, we need to use (26) with some small rearrangement

$$\begin{aligned} 2\gamma\langle F(z^{t+1/2}), z^{t+1/2} - z\rangle \quad &\leq \quad \|z^t - z\|^2 - \|z^{t+1} - z\|^2 - \|z^{t+1/2} - z^t\|^2 \\ &\quad +2\gamma\langle F(z^{t+1/2}) - \bar{g}^{t+1/2}, z^{t+1/2} - z\rangle + \gamma^2\|\bar{g}^{t+1/2} - \bar{g}^t\|^2. \end{aligned}$$

Next, we sum over all t from 0 to $k - 1$

$$\begin{aligned} \frac{1}{k}\sum_{t=0}^{k-1}\langle F(z^{t+1/2}), z^{t+1/2} - z\rangle \leq \; & \frac{\|z^0 - z\|^2 - \|z^{k+1} - z\|^2}{2\gamma k} \\ & + \frac{1}{k}\sum_{t=0}^{k-1}\langle F(z^{t+1/2}) - \bar{g}^{t+1/2}, z^{t+1/2} - z\rangle \\ & + \frac{1}{2\gamma k}\sum_{t=0}^{k-1}\gamma^2\|\bar{g}^{t+1/2} - \bar{g}^t\|^2 - \|z^{t+1/2} - z^t\|^2. \end{aligned} \tag{31}$$

Then, by $x^k_{avg} = \frac{1}{k}\sum_{t=0}^{k-1} x^{t+1/2}$ and $y^k_{avg} = \frac{1}{k}\sum_{t=0}^{k-1} y^{t+1/2}$, Jensen's inequality and convexity-concavity of f:

$$\begin{aligned} \text{gap}(z^k_{avg}) \quad &\leq \quad \max_{y^0\in Y} f\left(\frac{1}{k}\left(\sum_{t=0}^{k-1} x^{t+1/2}\right), y^0\right) - \min_{x^0\in X} f\left(x^0, \frac{1}{k}\left(\sum_{t=0}^{k-1} y^{t+1/2}\right)\right) \\ &\leq \quad \max_{y^0\in Y}\frac{1}{k}\sum_{t=0}^{k-1} f(x^{t+1/2}, y^0) - \min_{x^0\in X}\frac{1}{k}\sum_{t=0}^{k-1} f(x^0, y^{t+1/2}). \end{aligned}$$

Given the fact of linear independence of x^0 and y^0:

$$\text{gap}(z^k_{avg}) \quad \leq \quad \max_{(x^0,y^0)\in Z}\frac{1}{k}\sum_{t=0}^{k-1}\left(f(x^{t+1/2}, y^0) - f(x^0, y^{t+1/2})\right).$$

Using convexity and concavity of the function f:

$$\begin{aligned} \text{gap}(z^k_{avg}) &\leq \max_{(x^0,y^0)\in Z}\frac{1}{k}\sum_{t=0}^{k-1}\left(f(x^{t+1/2}, y^0) - f(x^0, y^{t+1/2})\right) \\ &= \max_{(x^0,y^0)\in Z}\frac{1}{k}\sum_{t=0}^{k-1}\Big(f(x^{t+1/2}, y^0) - f(x^{t+1/2}, y^{t+1/2}) + f(x^{t+1/2}, y^{t+1/2}) - f(x^0, y^{t+1/2})\Big) \end{aligned}$$

$$\leq \max_{(x^0,y^0)\in Z} \frac{1}{k}\sum_{t=0}^{k-1}\left(\langle \nabla_y f(x^{t+1/2},y^{t+1/2}), y^0 - y^{t+1/2}\rangle + \langle \nabla_x f(x^{t+1/2},y^{t+1/2}), x^{t+1/2} - x^0\rangle\right)$$

$$\leq \max_{z\in Z} \frac{1}{k}\sum_{t=0}^{k-1}\langle F(z^{t+1/2}), z^{t+1/2} - z\rangle. \tag{32}$$

Together with (32), (31) gives (additionally, we take a full expectation)

$$\begin{aligned}
\mathbb{E}[\mathrm{gap}(z^k_{avg})] \quad &\leq \quad \mathbb{E}\left[\max_{z\in Z}\frac{\|z^0 - z\|^2 - \|z^k - z\|^2}{2\gamma k}\right] \\
&\quad + \frac{1}{k}\mathbb{E}\left[\max_{z\in Z}\sum_{t=0}^{k-1}\langle F(z^{t+1/2}) - \bar{g}^{t+1/2}, z^{t+1/2} - z\rangle\right] \\
&\quad + \frac{1}{2\gamma k}\mathbb{E}\left[\sum_{t=0}^{k-1}\gamma^2\|\bar{g}^{t+1/2} - \bar{g}^t\|^2 - \|z^{t+1/2} - z^t\|^2\right] \\
&\overset{(8),(28)}{\leq} \quad \frac{\Omega_z^2}{2\gamma k} + \frac{1}{k}\mathbb{E}\left[\max_{z\in Z}\sum_{t=0}^{k-1}\langle F(z^{t+1/2}) - \bar{g}^{t+1/2}, z^{t+1/2} - z\rangle\right] \\
&\quad + \frac{1}{2\gamma k}\mathbb{E}\left[\sum_{t=0}^{k-1}3\gamma^2L^2\|z^{t+1/2} - z^t\|^2 + \frac{6\gamma^2\sigma^2}{bM} - \|z^{t+1/2} - z^t\|^2\right].
\end{aligned}$$

With $\gamma = \frac{1}{4L}$ we get

$$\mathbb{E}[\mathrm{gap}(z^k_{avg})] \leq \frac{\Omega_z^2}{2\gamma k} + \frac{1}{k}\mathbb{E}\left[\max_{z\in Z}\sum_{t=0}^{k-1}\langle F(z^{t+1/2}) - \bar{g}^{t+1/2}, z^{t+1/2} - z\rangle\right] + \frac{3\gamma\sigma^2}{bM}. \tag{33}$$

To finish the proof we need to estimate $\mathbb{E}\left[\max_{z\in Z}\sum_{t=0}^{k-1}\langle F(z^{t+1/2}) - \bar{g}^{t+1/2}, z^{t+1/2} - z\rangle\right]$. Let define sequence v: $v^0 = z^{1/2}$, $v^{t+1} = \mathrm{proj}_Z(v^t - \gamma\delta_t)$ with $\delta^t = F(z^{t+1/2}) - \bar{g}^{t+1/2}$. Then we have

$$\sum_{t=0}^{k-1}\langle\delta^t, z^{t+1/2} - z\rangle = \sum_{t=0}^{k-1}\langle\delta^t, z^{t+1/2} - v^t\rangle + \sum_{t=0}^{k-1}\langle\delta^t, v^t - z\rangle. \tag{34}$$

By the definition of v^{t+1}, we have for all $z \in Z$

$$\langle v^{t+1} - v^t + \gamma\delta^t, z - v^{t+1}\rangle \geq 0.$$

Rewriting this inequality, we get

$$\begin{aligned}
\langle\gamma\delta^t, v^t - z\rangle &\leq \langle\gamma\delta^t, v^t - v^{t+1}\rangle + \langle v^{t+1} - v^t, z - v^{t+1}\rangle \\
&\leq \langle\gamma\delta^t, v^t - v^{t+1}\rangle + \frac{1}{2}\|v^t - z\|^2 - \frac{1}{2}\|v^{t+1} - z\|^2 - \frac{1}{2}\|v^t - v^{t+1}\|^2 \\
&\leq \frac{\gamma^2}{2}\|\delta^t\|^2 + \frac{1}{2}\|v^t - v^{t+1}\|^2 + \frac{1}{2}\|v^t - z\|^2 - \frac{1}{2}\|v^{t+1} - z\|^2 - \frac{1}{2}\|v^t - v^{t+1}\|^2 \\
&= \frac{\gamma^2}{2}\|\delta^t\|^2 + \frac{1}{2}\|v^t - z\|^2 - \frac{1}{2}\|v^{t+1} - z\|^2.
\end{aligned}$$

With (34) it gives

$$\begin{aligned}\sum_{t=0}^{k-1}\langle \delta^t, z^{t+1/2} - z\rangle &\le \sum_{t=0}^{k-1}\langle \delta^t, z^{t+1/2} - v^t\rangle + \frac{1}{\gamma}\sum_{t=0}^{k-1}\left(\frac{\gamma^2}{2}\|\delta^t\|^2 + \frac{1}{2}\|v^t - z\|^2 - \frac{1}{2}\|v^{t+1} - z\|^2\right)\\ &\le \sum_{t=0}^{k-1}\langle \delta^t, z^{t+1/2} - v^t\rangle + \frac{\gamma}{2}\sum_{t=0}^{k-1}\|\delta^t\|^2 + \frac{1}{2\gamma}\|v^0 - z\|^2\\ &\le \sum_{t=0}^{k-1}\langle \delta^t, z^{t+1/2} - v^t\rangle + \frac{\gamma}{2}\sum_{t=0}^{k-1}\|\delta^t\|^2 + \frac{\Omega_z^2}{2\gamma}.\end{aligned}$$

The right side is independent of z, then

$$\max_{z\in Z}\sum_{t=0}^{k-1}\langle \delta^t, z^{t+1/2} - z\rangle \le \sum_{t=0}^{k-1}\langle \delta^t, z^{t+1/2} - v^t\rangle + \frac{\gamma}{2}\sum_{t=0}^{k-1}\|F(z^{t+1/2}) - \bar{g}^{t+1/2}\|^2 + \frac{\Omega_z^2}{2\gamma}. \quad (35)$$

Taking the full expectation and using independence $v^t - z^{t+1/2}$, $\{\xi_m^{t+1/2,i}\}_{i=1,m=1}^{b,M}$, we get

$$\begin{aligned}\mathbb{E}\left[\max_{z\in Z}\sum_{t=0}^{k-1}\langle \delta^t, z^{t+1/2} - z\rangle\right] &\le \mathbb{E}\left[\sum_{t=0}^{k-1}\langle \delta^t, z^{t+1/2} - v^t\rangle\right]\\ &\quad+\frac{\gamma}{2}\sum_{t=0}^{k-1}\mathbb{E}\left[\|F(z^{t+1/2}) - \bar{g}^{t+1/2}\|^2\right] + \frac{\Omega_z^2}{2\gamma}\\ &= \mathbb{E}\left[\sum_{t=0}^{k-1}\langle \mathbb{E}_{\{\xi_m^{t+1/2,i}\}_{i=1,m=1}^{b,M}}\left[F(z^{t+1/2}) - \bar{g}^{t+1/2}\right], z^{t+1/2} - v^t\rangle\right]\\ &\quad+\frac{\gamma}{2}\sum_{t=0}^{k-1}\mathbb{E}\left[\|F(z^{t+1/2}) - \bar{g}^{t+1/2}\|^2\right] + \frac{\Omega_z^2}{2\gamma}\\ &= \frac{\gamma}{2}\sum_{t=0}^{k-1}\mathbb{E}\left[\|F(z^{t+1/2}) - \bar{g}^{t+1/2}\|^2\right] + \frac{\Omega_z^2}{2\gamma}\\ &\overset{(28)}{\le} \frac{\gamma k}{2}\cdot\frac{3\sigma^2}{bM} + \frac{\Omega_z^2}{2\gamma}.\end{aligned}$$

Then we can finish (33) and get

$$\mathbb{E}[\mathrm{gap}(z^k_{avg})] \le \frac{\Omega_z^2}{\gamma k} + \frac{\gamma}{2}\cdot\frac{5\sigma^2}{bM}.$$

Let $\gamma = \min\left\{\frac{1}{4L}; \Omega_z\sqrt{\frac{2bM}{5k\sigma^2}}\right\}$ then

$$\mathbb{E}[\mathrm{gap}(z^k_{avg})] = \mathcal{O}\left(\frac{L\Omega_z^2}{k} + \frac{\sigma\Omega_z}{\sqrt{bMk}}\right).$$

Substitute the batch size b and the number of iterations k from the description of the Algorithm 1 with $r \le \Delta$:

$$\mathbb{E}[\mathrm{gap}(z^k_{avg})] = \mathcal{O}\left(\frac{L\Omega_z^2\Delta}{K} + \frac{\sigma\Omega_z}{\sqrt{MT}}\right).$$

Non-convex-non-concave problems

Theorem 14 (Theorem 5) *Let $\{z^t\}_{t\geq 0}$ denote the iterates of Algorithm 1 for solving problem* (1)*. Let Assumptions 1(g), 2(m), 3, 4 be satisfied. Then, if $\gamma \leq \frac{1}{4L}$, we have the following estimates:*

$$\mathbb{E}\left[\frac{1}{k}\sum_{t=0}^{k-1}\|F(z^t)\|^2\right] = \mathcal{O}\left(\frac{L^2\Omega_z^2\Delta}{K} + \frac{\sigma^2 K}{MT\Delta}\right).$$

Proof: We start proof with combining (27), (28) and (29)

$$\begin{aligned}\mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] &\leq \mathbb{E}\left[\|z^t - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] \\ &\quad -2\gamma\mathbb{E}\left[\langle F(z^{t+1/2}), z^{t+1/2} - z^*\rangle\right] + 3\gamma^2L^2\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] + \frac{6\gamma^2\sigma^2}{bM}.\end{aligned}$$

Using minty assumption (6), we obtain

$$\begin{aligned}\mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] &\leq \mathbb{E}\left[\|z^t - z^*\|^2\right] - (1 - 3\gamma^2L^2)\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] + \frac{6\gamma^2\sigma^2}{bM} \\ &= \mathbb{E}\left[\|z^t - z^*\|^2\right] - \gamma^2(1 - 3\gamma^2L^2)\mathbb{E}\left[\left\|\bar{g}^t\right\|^2\right] + \frac{6\gamma^2\sigma^2}{bM}.\end{aligned}$$

With $\gamma \leq \frac{1}{4L}$

$$\begin{aligned}\mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] &\leq \mathbb{E}\left[\|z^t - z^*\|^2\right] - (1 - 3\gamma^2L^2)\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] + \frac{6\gamma^2\sigma^2}{bM} \\ &= \mathbb{E}\left[\|z^t - z^*\|^2\right] - \frac{3\gamma^2}{4}\mathbb{E}\left[\left\|\bar{g}^t\right\|^2\right] + \frac{6\gamma^2\sigma^2}{bM}.\end{aligned}$$

The fact: $-\|\bar{g}^t\|^2 \leq -\frac{2}{3}\|F(z^t)\|^2 + 2\|\bar{g}^t - F(z^t)\|^2$, gives

$$\mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] \leq \mathbb{E}\left[\|z^t - z^*\|^2\right] - \frac{\gamma^2}{2}\mathbb{E}\left[\left\|F(z^t)\right\|^2\right] + 2\gamma^2\|\bar{g}^t - F(z^t)\|^2 + \frac{6\gamma^2\sigma^2}{bM}.$$

The term $\|\bar{g}^t - F(z^t)\|^2$ was estimated, when we deduced (28). Then

$$\frac{\gamma^2}{2}\mathbb{E}\left[\left\|F(z^t)\right\|^2\right] \leq \mathbb{E}\left[\|z^t - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] + \frac{8\gamma^2\sigma^2}{bM}.$$

Summing over all t from 0 to $k - 1$:

$$\mathbb{E}\left[\frac{1}{k}\sum_{t=0}^{k-1}\left\|F(z^t)\right\|^2\right] \leq \frac{2\mathbb{E}\left[\|z^0 - z^*\|^2\right]}{\gamma^2 k} + \frac{16\sigma^2}{bM}.$$

Next we substitute $\gamma = \frac{1}{4L}$, k, b and finish the proof.

10.2. Decentralized case

First of all, we present the missing Algorithm 4:

Algorithm 4 FastMix

Parameters: Vectors $z_1, ..., z_M$, communic. rounds P.

Initialization: Construct matrix $\mathbf{z}$ with rows $z_1^T, ..., z_M^T$,

choose $\mathbf{z}^{-1} = \mathbf{z}$, $\mathbf{z}^0 = \mathbf{z}$, $\eta = \frac{1-\sqrt{1-\lambda_2^2(\tilde{W})}}{1+\sqrt{1-\lambda_2^2(\tilde{W})}}$.

for $h = 0, 1, 2, \ldots, P-1$ **do**

$\mathbf{z}^{h+1} = (1+\eta)\tilde{W}\mathbf{z}^h - \eta \mathbf{z}^{h-1}$,

end for

Output: rows $z_1, ..., z_M$ of $\mathbf{z}^P$.

We introduce the following notation

$$z^t = \frac{1}{M}\sum_{m=1}^M z_m^t, \qquad z^{t+1/2} = \frac{1}{M}\sum_{m=1}^M z_m^{t+1/2}, \qquad g^t = \frac{1}{M}\sum_{m=1}^M g_m^t, \qquad g^{t+1/2} = \frac{1}{M}\sum_{m=1}^M g_m^{t+1/2},$$

$$\hat{z}^t = \frac{1}{M}\sum_{m=1}^M \hat{z}_m^t, \qquad \hat{z}^{t+1/2} = \frac{1}{M}\sum_{m=1}^M \hat{z}_m^{t+1/2}, \qquad \tilde{z}^t = \frac{1}{M}\sum_{m=1}^M \tilde{z}_m^t, \qquad \tilde{z}^{t+1/2} = \frac{1}{M}\sum_{m=1}^M \tilde{z}_m^{t+1/2}.$$

Next, we introduce the convergence of FastMix [20, 48]:

Lemma 9. *Assume that $\{\tilde{z}_m^{t+1}\}_{m=1}^M$ are output of Algorithm 4 with input $\{\hat{z}_m^{t+1}\}_{m=1}^M$. Then it holds that*

$$\frac{1}{M}\sum_{m=1}^M \|\tilde{z}_m^{t+1} - \tilde{z}^{t+1}\|^2 \leq \left(1 - \frac{1}{\sqrt{\chi}}\right)^{2P} \left(\frac{1}{M}\sum_{m=1}^M \|\hat{z}_m^{t+1} - \hat{z}^{t+1}\|^2\right) \quad \textit{and} \quad \hat{z}^t = \tilde{z}^t.$$

Let after H iteration we get ε_0 accuracy of consensus, i.e.

$$\tilde{z}_m^t - \tilde{z}^t = \delta_m^t, \qquad \|\delta_m^t\| \leq \varepsilon_0, \qquad \tilde{z}_m^{t+1/2} - \tilde{z}^{t+1/2} = \delta_m^{t+1/2}, \qquad \|\delta_m^{t+1/2}\| \leq \varepsilon_0. \tag{36}$$

Then let us estimate the number of iterations H to achieve such ε_0 (how to choose this parameter we will talk later) accuracy:

Corollary 1. *To achieve accuracy ε_0 in terms of* (36) *we need to take P:*

- *in convex-concave (Assumptions 2(c) and 4) and non-convex-non-concave (Assumptions 2(nc) and 4) cases*

$$P = O\left(\sqrt{\chi}\log\left(1 + \frac{\Omega_z^2 + \frac{Q^2+\sigma^2/b}{L_{\max}^2}}{\varepsilon_0^2}\right)\right),$$

· *in strongly convex-strongly concave case (Assumption 2(sc))*

$$P = O\left(\sqrt{\chi} \log\left(1 + \frac{\|z^0 - z^*\|^2 + \frac{Q^2+\sigma^2/b}{L_{\max}^2}}{\varepsilon_0^2}\right)\right),$$

where $Q^2 = \frac{1}{M}\sum\limits_{m=1}^{M} \|F_m(z^*)\|^2$.

Proof: The proof is in a rough estimate of $\frac{1}{M}\sum\limits_{m=1}^{M} \|\hat{z}_m^{t+1} - \hat{z}^{t+1}\|^2$:

$$\begin{aligned}
\frac{1}{M}\sum_{m=1}^{M} \|\hat{z}_m^{t+1} - \hat{z}^{t+1}\|^2 &= \frac{1}{M}\sum_{m=1}^{M} \|z_m^t - \gamma g_m^{t+1/2} - z^t + \gamma g^{t+1/2}\|^2 \\
&\leq \frac{2}{M}\sum_{m=1}^{M} \|z_m^t - z^t\|^2 + \frac{2\gamma^2}{M}\sum_{m=1}^{M} \|g_m^{t+1/2} - g^{t+1/2}\|^2 \\
&\leq \frac{2}{M}\sum_{m=1}^{M} \|\mathrm{proj}_{\mathcal{Z}}(\tilde{z}_m^t) - \frac{1}{M}\sum_{i=1}^{M} \mathrm{proj}_{\mathcal{Z}}(\tilde{z}_i^t)\|^2 + \frac{2\gamma^2}{M}\sum_{m=1}^{M} \|g_m^{t+1/2}\|^2
\end{aligned}$$

In last inequality we use property: $\frac{1}{M}\sum\limits_{m=1}^{M} \|g_m^{t+1/2} - g^{t+1/2}\|^2 \leq \frac{1}{M}\sum\limits_{m=1}^{M} \|g_m^{t+1/2}\|^2$.

$$\begin{aligned}
\frac{1}{M}\sum_{m=1}^{M} \|\hat{z}_m^t - \hat{z}^t\|^2 &\overset{(15)}{\leq} \frac{4}{M}\sum_{m=1}^{M} \|\mathrm{proj}_{\mathcal{Z}}(\tilde{z}_m^t) - \mathrm{proj}_{\mathcal{Z}}(\tilde{z}^t)\|^2 \\
&\quad + 4\|\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^t) - \frac{1}{M}\sum_{i=1}^{M} \mathrm{proj}_{\mathcal{Z}}(\tilde{z}_i^t)\|^2 \\
&\quad + \frac{2\gamma^2}{M}\sum_{m=1}^{M} \|g_m^{t+1/2}\|^2 \\
&\overset{(16),(15)}{\leq} \frac{8}{M}\sum_{m=1}^{M} \|\tilde{z}_m^t - \tilde{z}^t\|^2 + \frac{2\gamma^2}{M}\sum_{m=1}^{M} \|g_m^{t+1/2}\|^2 \\
&\overset{(36),(13)}{\leq} 8\varepsilon_0^2 + \frac{4\gamma^2}{M}\sum_{m=1}^{M} \|F_m(z^{t+1/2})\|^2 + \frac{4\gamma^2\sigma^2}{b} \\
&\leq 8\varepsilon_0^2 + \frac{8\gamma^2}{M}\sum_{m=1}^{M} \|F_m(z^{t+1/2}) - F_m(z^*)\|^2 \\
&\quad + \frac{8\gamma^2}{M}\sum_{m=1}^{M} \|F_m(z^*)\|^2 + \frac{4\gamma^2\sigma^2}{b} \\
&\overset{(4)}{\leq} 8\varepsilon_0^2 + 8\gamma^2 L_{\max}^2 \|z^{t+1/2} - z^*\|^2 + \frac{8\gamma^2}{M}\sum_{m=1}^{M} \|F_m(z^*)\|^2 + \frac{4\gamma^2\sigma^2}{b}.
\end{aligned}$$

The proof of the theorem follows from $\gamma \leq \frac{1}{4L_{\max}}$ and the fact that in the convex-concave and non-convex-non-concave cases we can bounded $\|z^{t+1/2} - z^*\| \leq \Omega_z$, in strongly convex-strongly concave – $\|z^{t+1/2} - z^*\| \leq \|z^0 - z^*\|$.

We are now ready to prove the main theorems. Note we can rewrite one step of the algorithm as follows:

$$
\begin{aligned}
z^{t+1/2} &= \frac{1}{M}\sum_{m=1}^{M} z_m^{t+1/2} = \frac{1}{M}\sum_{m=1}^{M} \mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2} + \delta_m^{t+1/2}) \\
&= \mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2}) + \frac{1}{M}\sum_{m=1}^{M} \mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2} + \delta_m^{t+1/2}) - \mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2}) \\
&= \mathrm{proj}_{\mathcal{Z}}\left(\frac{1}{M}\sum_{m=1}^{M} z_m^t - \gamma g_m^t\right) + \Delta^t = \mathrm{proj}_{\mathcal{Z}}\left(z^t - \gamma g^t\right) + \Delta^t
\end{aligned}
$$

Here we add one more notation: $\frac{1}{M}\sum\limits_{m=1}^{M} \mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2} + \delta_m^{t+1/2}) - \mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2}) = \Delta^{t+1/2}$. It is easy to see $\|\Delta^{t+1/2}\| \leq \varepsilon_0$. We see that the step of the algorithm is very similar to step of Algorithm 1, but with imprecise projection onto a set. Let us prove the following lemma:

Lemma 10. *Let $z, y, \Delta \in \mathbb{R}^n$ and $\mathcal{Z} \subseteq \mathbb{R}^n$ is convex compact set. We set $z^+ = \mathrm{proj}_{\mathcal{Z}}(z - y) + \Delta$, then for all $u \in \mathcal{Z}$:*

$$
\|z^+ - u\|^2 \leq \|z - u\|^2 - 2\langle y, z^+ - u\rangle - \|z^+ - z\|^2
$$

Proof: Let $r = \mathrm{proj}_{\mathcal{Z}}(z - y)$. For all $u \in \mathcal{Z}$ we have $\langle r - (z - y), r - u\rangle \leq 0$. Then

$$
\begin{aligned}
\|z^+ - u\|^2 &= \|z^+ - z + z - u\|^2 \\
&= \|z - u\|^2 + 2\langle z^+ - z, z - u\rangle + \|z^+ - z\|^2 \\
&= \|z - u\|^2 + 2\langle z^+ - z, z^+ - u\rangle - \|z^+ - z\|^2 \\
&= \|z - u\|^2 + 2\langle z^+ - (z - y), z^+ - u\rangle - 2\langle y, z^+ - u\rangle - \|z^+ - z\|^2 \\
&= \|z - u\|^2 + 2\langle r - (z - y), r - u\rangle + 2\langle \Delta, r - u\rangle + 2\langle z^+ - (z - y), \Delta\rangle \\
&\quad - 2\langle y, z^+ - u\rangle - \|z^+ - z\|^2 \\
&\leq \|z - u\|^2 + 2\langle \Delta, z^+ - u\rangle + 2\langle \Delta, r - (z - y)\rangle - 2\langle y, z^+ - u\rangle - \|z^+ - z\|^2 \\
&\leq \|z - u\|^2 + 2\|\Delta\| \cdot \|z^+ - u\| + 2\|\Delta\| \cdot \|\mathrm{proj}_{\mathcal{Z}}(z - y) - \mathrm{proj}_{\mathcal{Z}}(z)\| \\
&\quad + 2\|\Delta\| \cdot \|y\| - 2\langle y, z^+ - u\rangle - \|z^+ - z\|^2 \\
&\leq \|z - u\|^2 + 2\|\Delta\| \cdot \|z^+ - u\| + 4\|\Delta\| \cdot \|y\| - 2\langle y, z^+ - u\rangle - \|z^+ - z\|^2
\end{aligned}
$$

Convex-concave problems

Theorem 15 (Theorem 6) *Let $\{z_m^t\}_{t\geq 0}$ denote the iterates of Algorithm 2 for solving problem* (1)*. Let Assumptions 1(g), 1(l), 2(nc), 3 and 4 be satisfied. Then, if $\gamma \leq \frac{1}{4L}$ and $P = \mathcal{O}\left(\sqrt{\chi}\log\frac{1}{\varepsilon}\right)$, we have the following estimates in*

$$\mathbb{E}[gap(\bar{z}^k_{avg})] = \tilde{\mathcal{O}}\left(\frac{L\Omega_z^2\sqrt{\chi}}{K} + \frac{\sigma\Omega_z}{\sqrt{MT}}\right).$$

Proof: The same way as in Theorem 12 one can get

$$\begin{aligned}
\|z^{t+1} - z\|^2 \leq& \|z^t - z\|^2 - \|z^{t+1/2} - z^t\|^2 - 2\gamma\langle g^{t+1/2}, z^{t+1/2} - z\rangle \\
&+ \gamma^2\|g^{t+1/2} - g^t\|^2 + 4\|\Delta^{t+1/2}\|\cdot\|z^{t+1} - z\| + 4\mathbb{E}\left[\|\Delta^{t+1/2}\|\cdot\|\gamma g^{t+1/2}\|\right] \\
&+ 4\|\Delta^t\|\cdot\|z^{t+1/2} - z^{t+1}\| + 4\|\Delta^t\|\cdot\|\gamma g^t\| \\
\leq& \|z^t - z\|^2 - \|z^{t+1/2} - z^t\|^2 \\
&- 2\gamma\langle g^{t+1/2}, z^{t+1/2} - z\rangle + \gamma^2\|g^{t+1/2} - g^t\|^2 \\
&+ 4\varepsilon_0\|z^{t+1} - z\| + 4\varepsilon_0\gamma\|g^{t+1/2}\| + 4\varepsilon_0\|z^{t+1/2} - z^{t+1}\| + 4\varepsilon_0\gamma\|g^t\|. \qquad (37)
\end{aligned}$$

Here we use $\|\Delta^t\|, \|\Delta^{t+1/2}\| \leq \varepsilon_0$ and the triangle inequality. Next we use estimate on *gap* (32) and taking full expectation:

$$\begin{aligned}
2\gamma k\cdot\mathbb{E}[\text{gap}(\bar{z}^k_{avg})] \leq& 2\gamma\mathbb{E}\left[\max_{z\in\mathcal{Z}}\sum_{t=0}^{k-1}\langle F(z^{t+1/2}), z^{t+1/2} - z\rangle\right] \\
\leq& \Omega_z^2 - \sum_{t=0}^{k-1}\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] \\
&+ 2\gamma\mathbb{E}\left[\max_{z\in\mathcal{Z}}\sum_{t=0}^{k-1}\langle F(z^{t+1/2}) - g^{t+1/2}, z^{t+1/2} - z\rangle\right] \\
&+ \gamma^2\sum_{t=0}^{k-1}\mathbb{E}\left[\|g^{t+1/2} - g^t\|^2\right] + 4\varepsilon_0\sum_{t=0}^{k-1}\mathbb{E}\left[\max_{z\in\mathcal{Z}}\|z^{t+1} - z\|\right] \\
&+ 4\varepsilon_0\gamma\sum_{t=0}^{k-1}\mathbb{E}\left[\|g^{t+1/2}\|\right] + 4\varepsilon_0\sum_{t=0}^{k-1}\mathbb{E}\left[\|z^{t+1/2} - z^{t+1}\|\right] \\
&+ 4\varepsilon_0\gamma\sum_{t=0}^{k-1}\mathbb{E}\left[\|g^t\|\right]. \qquad (38)
\end{aligned}$$

Let work with $\mathbb{E}\left[\|\bar{g}^{t+1/2} - \bar{g}^t\|^2\right]$:

$$\mathbb{E}\left[\|g^{t+1/2} - g^t\|^2\right] = \mathbb{E}\Bigg[\Bigg\|g^{t+1/2} - \frac{1}{M}\sum_{m=1}^M F_m(z_m^{t+1/2}) + \frac{1}{M}\sum_{m=1}^M F_m(z_m^{t+1/2})$$

$$
\begin{aligned}
& -F(z^{t+1/2}) + F(z^t) - \frac{1}{M}\sum_{m=1}^{M} F_m(z^t_m) + \frac{1}{M}\sum_{m=1}^{M} F_m(z^t_m) - g^t \\
& +F(z^{t+1/2}) - F(z^t) k^2 \Bigg] \\
\overset{(15)}{\leq}\; & 5\mathsf{E}\left[\left\|\frac{1}{bM}\sum_{m=1}^{M}\sum_{i=1}^{b}(F_m(z_m^{t+1/2}, \xi_m^{t+1/2,i}) - F_m(z_m^{t+1/2}))\right\|^2\right] \\
& +5\mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} F_m(z_m^{t+1/2}) - F(z^{t+1/2})\right\|^2\right] \\
& +5\mathsf{E}\left[\left\|F(z^t) - \frac{1}{M}\sum_{m=1}^{M} F_m(z_m^t)\right\|^2\right] \\
& +5\mathsf{E}\left[\left\|\frac{1}{bM}\sum_{m=1}^{M}\sum_{i=1}^{b}(F_m(z_m^t, \xi_m^{t,i}) - F_m(z_m^t))\right\|^2\right] \\
& +5\mathsf{E}\left[\left\|F(z^{t+1/2}) - F(z^t)\right\|^2\right].
\end{aligned}
$$

Using that all $f\xi_m^{t,i} g_{i=1,m=1}^{b,M}$ and $f\xi_m^{t+1/2,i} g_{i=1,m=1}^{b,M}$ are independent, we get

$$
\begin{aligned}
\mathsf{E}\left[k g^{t+1/2} - g^t k^2\right] \overset{(3)}{\leq}\; & \frac{5}{(bM)^2}\sum_{m=1}^{M}\sum_{i=1}^{b}\mathsf{E}\left[\left\|F_m(z_m^{t+1/2}, \xi_m^{t+1/2,i}) - F_m(z_m^{t+1/2})\right\|^2\right] \\
& + \frac{5}{(bM)^2}\sum_{m=1}^{M}\sum_{i=1}^{b}\mathsf{E}\left[\left\|F_m(z_m^t, \xi_m^{t,i}) - F_m(z_m^t)\right\|^2\right] + 5L^2\mathsf{E}\left[k z^{t+1/2} - z^t k^2\right] \\
& + 5\mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}(F_m(z_m^{t+1/2}) - F_m(z^{t+1/2}))\right\|^2\right] \\
& + 5\mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}(F_m(z_m^t) - F_m(z^t))\right\|^2\right] \\
\overset{(4),(7),(15)}{\leq}\; & 5L^2\mathsf{E}\left[k z^{t+1/2} - z^t k^2\right] + \frac{10\sigma^2}{bM} \\
& + 5L^2_{\max}\mathsf{E}\left[\frac{1}{M}\sum_{m=1}^{M}\left\|\mathrm{proj}_Z(\tilde{z}^{t+1/2} + \delta_m^{t+1/2}) - \frac{1}{M}\sum_{j=1}^{M}\mathrm{proj}_Z(\tilde{z}^{t+1/2} + \delta_j^{t+1/2})\right\|^2\right] \\
& + 5L^2_{\max}\mathsf{E}\left[\frac{1}{M}\sum_{m=1}^{M}\left\|\mathrm{proj}_Z(\tilde{z}^t + \delta_m^t) - \frac{1}{M}\sum_{j=1}^{M}\mathrm{proj}_Z(\tilde{z}^t + \delta_j^t)\right\|^2\right]
\end{aligned}
$$

$$
\mathsf{E}\left[k g^{t+1/2} - g^t k^2\right] \leq 5L^2\mathsf{E}\left[k z^{t+1/2} - z^t k^2\right] + \frac{10\sigma^2}{bM}
$$

$$+ 10L_{\max}^2 \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M}\left\|\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2}+\delta_m^{t+1/2}) - \mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2})\right\|^2\right]$$
$$+ 10L_{\max}^2 \mathbb{E}\left[\frac{1}{M}\sum_{j=1}^{M}\left\|(\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2}+\delta_j^{t+1/2}) - \mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2}))\right\|^2\right]$$
$$+ 10L_{\max}^2 \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M}\left\|\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t}+\delta_m^{t}) - \mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t})\right\|^2\right]$$
$$+ 10L_{\max}^2 \mathbb{E}\left[\frac{1}{M}\sum_{j=1}^{M}\left\|(\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t}+\delta_j^{t}) - \mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t}))\right\|^2\right]$$
$$\overset{(36)}{\leq} 5L^2\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] + \frac{10\sigma^2}{bM} + 40L_{\max}^2\varepsilon_0^2. \qquad (39)$$

Next we estimate $\mathbb{E}\left[\max_{z\in\mathcal{Z}}\sum_{t=0}^{k-1}\langle F(z^{t+1/2}) - g^{t+1/2}, z^{t+1/2} - z\rangle\right]$. To begin with, we use the same approach as in (34), (35) with sequence v: $v^0 = z^{1/2}$, $v^{t+1} = \mathrm{proj}_{\mathcal{Z}}(v^t - \gamma(F(z^{t+1/2}) - g^{t+1/2}))$ and get

$$\max_{z\in\mathcal{Z}}\sum_{t=0}^{k-1}\langle F(z^{t+1/2}) - g^{t+1/2}, z^{t+1/2} - z\rangle \leq \sum_{t=0}^{k-1}\langle F(z^{t+1/2}) - g^{t+1/2}, z^{t+1/2} - v^t\rangle + \frac{\gamma}{2}\sum_{t=0}^{k-1}\|F(z^{t+1/2}) - g^{t+1/2}\|^2 + \frac{\Omega_z^2}{2\gamma}.$$

To begin with, we use the independence of all ξ, as well as the unbiasedness of $g^{t+1/2}$ with respect to the conditional m.o. by random variables $\{\xi_m^{t+1/2,i}\}_{i=1,m=1}^{b,M}$:

$$\mathbb{E}\left[\max_{z\in\mathcal{Z}}\sum_{t=0}^{k-1}\langle F(z^{t+1/2}) - g^{t+1/2}, z^{t+1/2} - z\rangle\right]$$
$$\leq \sum_{t=0}^{k-1}\mathbb{E}\left[\mathbb{E}_{\{\xi_m^{t+1/2,i}\}_{i=1,m=1}^{b,M}}\left[\langle F(z^{t+1/2}) - g^{t+1/2}, z^{t+1/2} - v^t\rangle\right]\right] + \frac{\gamma}{2}\sum_{t=0}^{k-1}\mathbb{E}\left[\|F(z^{t+1/2}) - g^{t+1/2}\|^2\right] + \frac{\Omega_z^2}{2\gamma}$$
$$= \sum_{t=0}^{k-1}\mathbb{E}\left[\langle\mathbb{E}_{\{\xi_m^{t+1/2,i}\}_{i=1,m=1}^{b,M}}\left[F(z^{t+1/2}) - g^{t+1/2}\right], z^{t+1/2} - v^t\rangle\right] + \frac{\gamma}{2}\sum_{t=0}^{k-1}\mathbb{E}\left[\|F(z^{t+1/2}) - g^{t+1/2}\|^2\right] + \frac{\Omega_z^2}{2\gamma}$$
$$= \sum_{t=0}^{k-1}\mathbb{E}\left[\langle\frac{1}{M}\sum_{m=1}^{M}(F_m(z^{t+1/2}) - F_m(z_m^{t+1/2})), z^{t+1/2} - v^t\rangle\right] + \frac{\gamma}{2}\sum_{t=0}^{k-1}\mathbb{E}\left[\|F(z^{t+1/2}) - g^{t+1/2}\|^2\right] + \frac{\Omega_z^2}{2\gamma}$$

$$
\begin{aligned}
&\mathbb{E}\left[\max_{z\in\mathcal{Z}}\sum_{t=0}^{k-1}\langle F(z^{t+1/2})-g^{t+1/2}, z^{t+1/2}-z\rangle\right] \\
&\le \sum_{t=0}^{k-1}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}(F_m(z_m^{t+1/2})-F_m(z^{t+1/2}))\right\|\cdot\|z^{t+1/2}-v^t\|\right] \\
&\quad+\frac{\gamma}{2}\sum_{t=0}^{k-1}\mathbb{E}\left[\|F(z^{t+1/2})-g^{t+1/2}\|^2\right]+\frac{\Omega_z^2}{2\gamma} \\
&\overset{(4)}{\le} \sum_{t=0}^{k-1}\mathbb{E}\left[\left(\frac{L_{\max}}{M}\sum_{m=1}^{M}\left\|z_m^{t+1/2}-z^{t+1/2}\right\|\right)\cdot\|z^{t+1/2}-v^t\|\right] \\
&\quad+\frac{\gamma}{2}\sum_{t=0}^{k-1}\mathbb{E}\left[\|F(z^{t+1/2})-g^{t+1/2}\|^2\right]+\frac{\Omega_z^2}{2\gamma} \\
&\le \sum_{t=0}^{k-1}\mathbb{E}\left[\left(\frac{L_{\max}}{M}\sum_{m=1}^{M}\left\|\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2}+\delta_m^{t+1/2})-\frac{1}{M}\sum_{j=1}^{M}\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2}+\delta_j^{t+1/2})\right\|\right)\cdot\|z^{t+1/2}-v^t\|\right] \\
&\quad+\frac{\gamma}{2}\sum_{t=0}^{k-1}\mathbb{E}\left[\|F(z^{t+1/2})-g^{t+1/2}\|^2\right]+\frac{\Omega_z^2}{2\gamma} \\
&\le \sum_{t=0}^{k-1}\mathbb{E}\left[\left(\frac{L_{\max}}{M}\sum_{m=1}^{M}\left\|\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2}+\delta_m^{t+1/2})-\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2})\right\|\right)\cdot\|z^{t+1/2}-v^t\|\right] \\
&\quad+\mathbb{E}\left[\left(\frac{L_{\max}}{M}\sum_{j=1}^{M}\left\|(\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2}+\delta_j^{t+1/2})-\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2}))\right\|\right)\cdot\|z^{t+1/2}-v^t\|\right] \\
&\quad+\frac{\gamma}{2}\sum_{t=0}^{k-1}\mathbb{E}\left[\|F(z^{t+1/2})-g^{t+1/2}\|^2\right]+\frac{\Omega_z^2}{2\gamma} \\
&\overset{(36)}{\le} 2L_{\max}\varepsilon_0\sum_{t=0}^{k-1}\mathbb{E}\left[\|z^{t+1/2}-v^t\|\right]+\frac{\gamma}{2}\sum_{t=0}^{k-1}\mathbb{E}\left[\|F(z^{t+1/2})-g^{t+1/2}\|^2\right]+\frac{\Omega_z^2}{2\gamma} \\
&\le 2L_{\max}\varepsilon_0 k\Omega_z+\frac{\gamma}{2}\sum_{t=0}^{k-1}\mathbb{E}\left[\|F(z^{t+1/2})-g^{t+1/2}\|^2\right]+\frac{\Omega_z^2}{2\gamma}. \qquad (40)
\end{aligned}
$$

Next we combine (38), (39), (40)

$$
\begin{aligned}
2\gamma k\mathbb{E}[\mathrm{gap}(\bar{z}_{avg}^k)] \le{}& 2\Omega_z^2+(5L^2\gamma^2-1)\sum_{t=0}^{k-1}\mathbb{E}\left[\|z^{t+1/2}-z^t\|^2\right] \\
&+4\gamma L_{\max}\varepsilon_0 k\Omega_z+\gamma^2\sum_{t=0}^{k-1}\mathbb{E}\left[\|F(z^{t+1/2})-g^{t+1/2}\|^2\right]+\gamma^2\frac{10k\sigma^2}{bM} \\
&+40\gamma^2 kL_{\max}^2\varepsilon_0^2+4\varepsilon_0\sum_{t=0}^{k-1}\mathbb{E}\left[\max_{z\in\mathcal{Z}}\|z^{t+1}-z\|\right]+4\varepsilon_0\gamma\sum_{t=0}^{k-1}\mathbb{E}\left[\|g^{t+1/2}\|\right] \\
&+4\varepsilon_0\sum_{t=0}^{k-1}\mathbb{E}\left[\|z^{t+1/2}-z^{t+1}\|\right]+4\varepsilon_0\gamma\sum_{t=0}^{k-1}\mathbb{E}\left[\|g^t\|\right].
\end{aligned}
$$

Then we use $\gamma \leq \frac{1}{4L}$ and Assumption 4:

$$
\begin{aligned}
2\gamma k \mathbb{E}[\mathrm{gap}(\bar{z}^k_{avg})] \leq 2\Omega_z^2 + \gamma^2 \sum_{t=0}^{k-1} \mathbb{E}\left[\|F(z^{t+1/2}) - g^{t+1/2}\|^2\right] \\
+ \gamma^2 \frac{10k\sigma^2}{bM} + 40\gamma^2 k L_{\max}^2 \varepsilon_0^2 + 8(1+\gamma L_{\max})\varepsilon_0 k \Omega_z \\
+ 4\varepsilon_0 \gamma \sum_{t=0}^{k-1} \mathbb{E}\left[\|g^{t+1/2}\|\right] + 4\varepsilon_0 \gamma \sum_{t=0}^{k-1} \mathbb{E}\left[\|g^t\|\right]. \qquad (41)
\end{aligned}
$$

It remains to estimate $\mathbb{E}\left[\|g^{t+1/2}\| + \|g^t\|\right]$:

$$
\begin{aligned}
\mathbb{E}\left[\|g^t\|\right] &= \mathbb{E}\left[\left\|F(z^*) - F(z^*) + F(z^t) - F(z^t) + \frac{1}{M}\sum_{m=1}^M F_m(z^t_m) - \frac{1}{M}\sum_{m=1}^M F_m(z^t_m) + g^t\right\|\right] \\
&\leq \|F(z^*)\| + \mathbb{E}\left[\|F(z^t) - F(z^*)\|\right] + \mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M F_m(z^t_m) - F(z^t)\right\|\right] \\
&\quad + \mathbb{E}\left[\left\|\frac{1}{bM}\sum_{m=1}^M \sum_{i=1}^b (F_m(z^t_m, \xi^{t,i}_m) - F_m(z^t_m))\right\|\right].
\end{aligned}
$$

From (39) we have that $\mathbb{E}\left[\left\|\frac{1}{bM}\sum\limits_{m=1}^M \sum\limits_{i=1}^b (F_m(z^t_m, \xi^{t,i}_m) - F_m(z^t_m))\right\|^2\right] \leq \frac{\sigma^2}{bM}$ and from (40) $\mathbb{E}\left[\left\|\frac{1}{M}\sum\limits_{m=1}^M F_m(z^t_m) - F(z^t)\right\|\right] \leq 2L_{\max}\varepsilon_0$, then

$$
\begin{aligned}
\mathbb{E}\left[\|g^t\|\right] &\leq \|F(z^*)\| + \mathbb{E}\left[\|F(z^t) - F(z^*)\|\right] + 2L_{\max}\varepsilon_0 + \frac{\sigma}{\sqrt{bM}} \\
&\leq Q + L\Omega_z + 2L_{\max}\varepsilon_0 + \frac{\sigma}{\sqrt{bM}},
\end{aligned}
$$

where $Q^2 = \frac{1}{M}\sum\limits_{m=1}^M \|F_m(z^*)\|^2$. Hence, we can rewrite (41):

$$
\begin{aligned}
\mathbb{E}[\mathrm{gap}(\bar{z}^k_{avg})] &\leq \frac{\Omega_z^2}{2\gamma k} + \frac{\gamma}{2k}\sum_{t=0}^{k-1} \mathbb{E}\left[\|F(z^{t+1/2}) - g^{t+1/2}\|^2\right] + \frac{5\sigma^2\gamma}{bM} \\
&\quad + 20\gamma L_{\max}^2 \varepsilon_0^2 + 4\left(\frac{1}{\gamma} + L_{\max}\right)\varepsilon_0\Omega_z + 4\varepsilon_0\left(Q + L\Omega_z + 2L_{\max}\varepsilon_0 + \frac{\sigma}{\sqrt{bM}}\right).
\end{aligned}
$$

The same way as (39), one can estimate $E\left[\|F(z^{t+1/2}) - g^{t+1/2}\|^2\right]$:

$$
\begin{aligned}
\mathbb{E}[\mathrm{gap}(\bar{z}^k_{avg})] &\leq \frac{\Omega_z^2}{2\gamma k} + \frac{6\sigma^2\gamma}{bM} \\
&\quad + 24\gamma L_{\max}^2 \varepsilon_0^2 + 4\left(\frac{1}{\gamma} + L_{\max}\right)\varepsilon_0\Omega_z + 4\varepsilon_0\left(Q + L\Omega_z + 2L_{\max}\varepsilon_0 + \frac{\sigma}{\sqrt{bM}}\right).
\end{aligned}
$$

Let $\gamma = \min\left\{\frac{1}{4L}; \Omega_z\sqrt{\frac{bM}{12k\sigma^2}}\right\}$ and $\varepsilon_0 = \mathcal{O}\left(\frac{\varepsilon}{\Omega_z L_{\max} + Q}\right)$, where $\varepsilon = \max\left(\frac{L\Omega_z^2}{k}; \frac{\sigma\Omega_z}{\sqrt{bMk}}\right)$.Then for the output of the Algorithm 4 it is true

$$
\mathbb{E}[\mathrm{gap}(\bar{z}^k_{avg})] = \mathcal{O}\left(\frac{L\Omega_z^2}{k} + \frac{\sigma\Omega_z}{\sqrt{bMk}}\right).
$$

Substitute the batch size b and the number of iterations k from the description of the Algorithm 2 and Corollary 1:

$$\mathbb{E}[\text{gap}(\bar{z}^k_{avg})] = \tilde{\mathcal{O}}\left(\frac{L\Omega_z^2\sqrt{\chi}}{K} + \frac{\sigma\Omega_z}{\sqrt{MT}}\right).$$

Strongly convex-strongly concave problems

Theorem 16 (Theorem 6) *Let $\{z^t_m\}_{t\geq 0}$ denote the iterates of Algorithm 2 for solving problem* (1). *Let Assumptions 1(g), 1(l), 2(sc) and 3 be satisfied. Then, if $\gamma \leq \frac{1}{4L}$ and $P = \mathcal{O}\left(\sqrt{\chi}\log\frac{1}{\varepsilon}\right)$, we have the following estimate for the distance to the solution z^*:*

$$\mathbb{E}\left[\|\bar{z}^k - z^*\|^2\right] = \tilde{\mathcal{O}}\left(\|z^0 - z^*\|^2 \exp\left(-\frac{\mu}{8L}\cdot\frac{K}{\sqrt{\chi}}\right) + \frac{\sigma^2}{\mu^2 MT}\right).$$

Proof: We start with substituting $z = z^*$ in (37) and taking full expectation. Then we use (39) and get

$$\begin{aligned}
2\gamma\mathbb{E}\left[\langle F(z^{t+1/2}), z^{t+1/2} - z^*\rangle\right] &\leq \mathbb{E}\left[\|z^t - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] \\
&+ 5L^2\gamma^2\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] + \frac{10\sigma^2\gamma^2}{bM} \\
&+ 4\varepsilon_0\mathbb{E}\left[\|z^{t+1} - z^*\|\right] + 4\varepsilon_0\gamma\mathbb{E}\left[\|g^{t+1/2}\|\right] + 4\varepsilon_0\mathbb{E}\left[\|z^{t+1/2} - z^{t+1}\|\right] \\
&+ 4\varepsilon_0\gamma\mathbb{E}\left[\|g^t\|\right] + 2\gamma\mathbb{E}\left[\langle F(z^{t+1/2}) - g^{t+1/2}, z^{t+1/2} - z^*\rangle\right] + 40\gamma^2 L_{\max}^2\varepsilon_0^2.
\end{aligned}$$

The same way as (40), one can get

$$\begin{aligned}
&\mathbb{E}\left[\langle F(z^{t+1/2}) - g^{t+1/2}, z^{t+1/2} - z^*\rangle\right] \\
&\quad\leq \mathbb{E}\left[\mathbb{E}_{\{\xi^{t+1/2,i}_m\}_{i=1,m=1}^{b,M}}\left[\langle F(z^{t+1/2}) - g^{t+1/2}, z^{t+1/2} - z^*\rangle\right]\right] \\
&\quad= \mathbb{E}\left[\langle\mathbb{E}_{\{\xi^{t+1/2,i}_m\}_{i=1,m=1}^{b,M}}\left[F(z^{t+1/2}) - g^{t+1/2}\right], z^{t+1/2} - z^*\rangle\right] \\
&\quad= \mathbb{E}\left[\langle\frac{1}{M}\sum_{m=1}^{M}(F_m(z^{t+1/2}) - F_m(z^{t+1/2}_m)), z^{t+1/2} - z^*\rangle\right] \\
&\quad\leq \mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}(F_m(z^{t+1/2}_m) - F_m(z^{t+1/2}))\right\|\cdot\|z^{t+1/2} - z^*\|\right] \\
&\quad\overset{(4)}{\leq} \mathbb{E}\left[\left(\frac{L_{\max}}{M}\sum_{m=1}^{M}\left\|z^{t+1/2}_m - z^{t+1/2}\right\|\right)\cdot\|z^{t+1/2} - z^*\|\right] \\
&\quad\leq \mathbb{E}\left[\left(\frac{L_{\max}}{M}\sum_{m=1}^{M}\left\|\text{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2} + \delta^{t+1/2}_m) - \frac{1}{M}\sum_{j=1}^{M}\text{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2} + \delta^{t+1/2}_j)\right\|\right)\cdot\|z^{t+1/2} - z^*\|\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\left(\frac{L_{\max}}{M}\sum_{m=1}^{M}\left\|\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2}+\delta_m^{t+1/2})-\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2})\right\|\right)\cdot\|z^{t+1/2}-z^*\|\right] \\
&\quad+\mathbb{E}\left[\left(\frac{L_{\max}}{M}\sum_{j=1}^{M}\left\|(\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2}+\delta_j^{t+1/2})-\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2}))\right\|\right)\cdot\|z^{t+1/2}-z^*\|\right] \\
&\overset{(36)}{\leq} 2L_{\max}\varepsilon_0\mathbb{E}\left[\|z^{t+1/2}-z^*\|\right].
\end{aligned}
$$

and then

$$
\begin{aligned}
&2\gamma\mathbb{E}\left[\langle F(z^{t+1/2}),z^{t+1/2}-z^*\rangle\right] \\
&\quad\leq \mathbb{E}\left[\|z^t-z^*\|^2\right]-\mathbb{E}\left[\|z^{t+1}-z^*\|^2\right]-\mathbb{E}\left[\|z^{t+1/2}-z^t\|^2\right] \\
&\qquad+5L^2\gamma^2\mathbb{E}\left[\|z^{t+1/2}-z^t\|^2\right]+\frac{10\sigma^2\gamma^2}{bM} \\
&\qquad+4\varepsilon_0\mathbb{E}\left[\|z^{t+1}-z^*\|\right]+4\varepsilon_0\gamma\mathbb{E}\left[\|g^{t+1/2}\|\right]+4\varepsilon_0\mathbb{E}\left[\|z^{t+1/2}-z^{t+1}\|\right]+4\varepsilon_0\gamma\mathbb{E}\left[\|g^t\|\right] \\
&\qquad+4\gamma L_{\max}\varepsilon_0\mathbb{E}\left[\|z^{t+1/2}-z^*\|\right]+40\gamma^2L_{\max}^2\varepsilon_0^2.
\end{aligned}
$$

Next, we work with

$$
\begin{aligned}
z^{t+1} &= \frac{1}{M}\sum_{m=1}^{M}z_m^{t+1}=\frac{1}{M}\sum_{m=1}^{M}\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1/2}+\delta_m^{t+1}) \\
&= \mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1})+\frac{1}{M}\sum_{m=1}^{M}\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1}+\delta_m^{t+1})-\mathrm{proj}_{\mathcal{Z}}(\tilde{z}^{t+1}) \\
&= \mathrm{proj}_{\mathcal{Z}}\left(\frac{1}{M}\sum_{m=1}^{M}z_m^t-\gamma g_m^{t+1/2}\right)+\Delta^{t+1/2}=\mathrm{proj}_{\mathcal{Z}}\left(z^t-\gamma g^{t+1/2}\right)+\Delta^{t+1/2},
\end{aligned}
$$

and get

$$
\begin{aligned}
&2\gamma\mathbb{E}\left[\langle F(z^{t+1/2}),z^{t+1/2}-z^*\rangle\right] \\
&\quad\leq \mathbb{E}\left[\|z^t-z^*\|^2\right]-\mathbb{E}\left[\|z^{t+1}-z^*\|^2\right]-\mathbb{E}\left[\|z^{t+1/2}-z^t\|^2\right] \\
&\qquad+5L^2\gamma^2\mathbb{E}\left[\|z^{t+1/2}-z^t\|^2\right]+\frac{10\sigma^2\gamma^2}{bM} \\
&\qquad+8\varepsilon_0\mathbb{E}\left[\|z^{t+1}-z^t\|\right]+4\varepsilon_0\gamma\mathbb{E}\left[\|g^{t+1/2}\|\right]+4\varepsilon_0\gamma\mathbb{E}\left[\|g^t\|\right] \\
&\qquad+4\varepsilon_0(1+\gamma L_{\max})\mathbb{E}\left[\|z^{t+1/2}-z^t\|\right]+4\varepsilon_0(1+\gamma L_{\max})\mathbb{E}\left[\|z^t-z^*\|\right]+40\gamma^2L_{\max}^2\varepsilon_0^2 \\
&\quad\leq \mathbb{E}\left[\|z^t-z^*\|^2\right]-\mathbb{E}\left[\|z^{t+1}-z^*\|^2\right]-\mathbb{E}\left[\|z^{t+1/2}-z^t\|^2\right] \\
&\qquad+5L^2\gamma^2\mathbb{E}\left[\|z^{t+1/2}-z^t\|^2\right]+\frac{10\sigma^2\gamma^2}{bM} \\
&\qquad+8\varepsilon_0\mathbb{E}\left[\|\mathrm{proj}_{\mathcal{Z}}\left(z^t-\gamma g^{t+1/2}\right)+\Delta^{t+1/2}-z^t\|\right]+4\varepsilon_0\gamma\mathbb{E}\left[\|g^{t+1/2}\|\right]+4\varepsilon_0\gamma\mathbb{E}\left[\|g^t\|\right] \\
&\qquad+4\varepsilon_0(1+\gamma L_{\max})\mathbb{E}\left[\|z^{t+1/2}-z^t\|\right]+4\varepsilon_0(1+\gamma L_{\max})\mathbb{E}\left[\|z^t-z^*\|\right]+40\gamma^2L_{\max}^2\varepsilon_0^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\|z^t - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] \\
&\quad + 5L^2\gamma^2\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] + \frac{10\sigma^2\gamma^2}{bM} \\
&\quad + 8\varepsilon_0\mathbb{E}\left[\|\mathrm{proj}_{\mathcal{Z}}\left(z^t - \gamma g^{t+1/2}\right) - \mathrm{proj}_{\mathcal{Z}}(z^t)\|\right] + 8\varepsilon_0^2 \\
&\quad + 4\varepsilon_0(1+\gamma L_{\max})\mathbb{E}\left[\|z^t - z\|\right] + 4\varepsilon_0\gamma\mathbb{E}\left[\|g^{t+1/2}\|\right] \\
&\quad + 4\varepsilon_0\gamma\mathbb{E}\left[\|g^t\|\right] + 4\varepsilon_0(1+\gamma L_{\max})\mathbb{E}\left[\|z^{t+1/2} - z^t\|\right] + 40\gamma^2 L_{\max}^2\varepsilon_0^2 \\
&\overset{(16)}{\leq} \mathbb{E}\left[\|z^t - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] \\
&\quad + 5L^2\gamma^2\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] + \frac{10\sigma^2\gamma^2}{bM} \\
&\quad + 8\varepsilon_0^2 + 4\varepsilon_0(1+\gamma L_{\max})\mathbb{E}\left[\|z^t - z^*\|\right] + 12\varepsilon_0\gamma\mathbb{E}\left[\|g^{t+1/2}\|\right] \\
&\quad + 4\varepsilon_0\gamma\mathbb{E}\left[\|g^t\|\right] + 4\varepsilon_0(1+\gamma L_{\max})\mathbb{E}\left[\|z^{t+1/2} - z^t\|\right] + 40\gamma^2 L_{\max}^2\varepsilon_0^2. \qquad (42)
\end{aligned}
$$

It remains to estimate $\mathbb{E}\left[\|g^{t+1/2}\| + \|g^t\|\right]$:

$$
\begin{aligned}
\mathbb{E}\left[\|g^t\|\right] &= \mathbb{E}\left[\|F(z^*) - F(z^*) + F(z^t) - F(z^t) + \frac{1}{M}\sum_{m=1}^{M} F_m(z_m^t) - \frac{1}{M}\sum_{m=1}^{M} F_m(z_m^t) + g^t\|\right] \\
&\leq \|F(z^*)\| + \mathbb{E}\left[\left\|F(z^t) - F(z^*)\right\|\right] + \mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} F_m(z_m^t) - F(z^t)\right\|\right] \\
&\quad + \mathbb{E}\left[\left\|\frac{1}{bM}\sum_{m=1}^{M}\sum_{i=1}^{b}(F_m(z_m^t, \xi_m^{t,i}) - F_m(z_m^t))\right\|\right].
\end{aligned}
$$

From (39) we have that $\mathbb{E}\left[\left\|\frac{1}{bM}\sum\limits_{m=1}^{M}\sum\limits_{i=1}^{b}(F_m(z_m^t, \xi_m^{t,i}) - F_m(z_m^t))\right\|^2\right] \leq \frac{\sigma^2}{bM}$ and from (40) $\mathbb{E}\left[\left\|\frac{1}{M}\sum\limits_{m=1}^{M} F_m(z_m^t) - F(z^t)\right\|\right] \leq 2L_{\max}\varepsilon_0$, then

$$
\begin{aligned}
\mathbb{E}\left[\|g^t\|\right] &\leq \|F(z^*)\| + \mathbb{E}\left[\left\|F(z^t) - F(z^*)\right\|\right] + 2L_{\max}\varepsilon_0 + \sqrt{\frac{\sigma}{bM}} \\
&\leq Q + L\mathbb{E}\left[\|z^t - z^*\|\right] + 2L_{\max}\varepsilon_0 + \sqrt{\frac{\sigma}{bM}}.
\end{aligned}
$$

Substituting in (42):

$$
\begin{aligned}
&2\gamma\mathbb{E}\left[\langle F(z^{t+1/2}), z^{t+1/2} - z^*\rangle\right] \\
&\quad \leq \mathbb{E}\left[\|z^t - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] \\
&\qquad + 5L^2\gamma^2\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] + \frac{10\sigma^2\gamma^2}{bM} + 8\varepsilon_0^2 + 4\varepsilon_0(1+\gamma L_{\max})\mathbb{E}\left[\|z^t - z^*\|\right] \\
&\qquad + 12\varepsilon_0\gamma\left(Q + L\mathbb{E}\left[\|z^{t+1/2} - z^*\|\right] + 2L_{\max}\varepsilon_0 + \sqrt{\frac{\sigma}{bM}}\right) \\
&\qquad + 4\varepsilon_0\gamma\left(Q + L\mathbb{E}\left[\|z^t - z^*\|\right] + 2L_{\max}\varepsilon_0 + \sqrt{\frac{\sigma}{bM}}\right)
\end{aligned}
$$

$$+ 4\varepsilon_0(1+\gamma L_{\max})\mathsf{E}\left[\|z^{t+1/2} - z^t\|\right] + 40\gamma^2 L_{\max}^2\varepsilon_0^2. \tag{43}$$

By simple fact $2ab \le a^2 + b^2$, we get

$$\begin{aligned}
&2\gamma\mathsf{E}\left[\langle F(z^{t+1/2}), z^{t+1/2} - z^*\rangle\right] \\
&\le \mathsf{E}\left[\|z^t - z^*\|^2\right] - \mathsf{E}\left[\|z^{t+1} - z^*\|^2\right] - \mathsf{E}\left[\|z^{t+1/2} - z^t\|^2\right] \\
&\quad + 5L^2\gamma^2\mathsf{E}\left[\|z^{t+1/2} - z^t\|^2\right] + \frac{10\sigma^2\gamma^2}{bM} \\
&\quad + 20\varepsilon_0(1+\gamma L_{\max})\mathsf{E}\left[\|z^t - z^*\|\right] + 16\varepsilon_0(1+\gamma L_{\max})\mathsf{E}\left[\|z^{t+1/2} - z^t\|\right] \\
&\quad + 16\varepsilon_0\gamma\left(Q + 2L_{\max}\varepsilon_0 + \frac{\sigma}{\sqrt{bM}}\right) + 40\gamma^2 L_{\max}^2\varepsilon_0^2 + 8\varepsilon_0^2 \\
&\le (1+10\varepsilon_0)\mathsf{E}\left[\|z^t - z^*\|^2\right] - \mathsf{E}\left[\|z^{t+1} - z^*\|^2\right] \\
&\quad + (5L^2\gamma^2 + 8\varepsilon_0 - 1)\mathsf{E}\left[\|z^{t+1/2} - z^t\|^2\right] + \frac{10\sigma^2\gamma^2}{bM} \\
&\quad + 20\varepsilon_0(1+\gamma L_{\max})^2 + 16\varepsilon_0\gamma\left(Q + 2L_{\max}\varepsilon_0 + \frac{\sigma}{\sqrt{bM}}\right) + 40\gamma^2 L_{\max}^2\varepsilon_0^2 + 8\varepsilon_0^2.
\end{aligned} \tag{44}$$

By property of z^*, we get

$$\mathsf{E}\left[\langle F(z^{t+1/2}), z^{t+1/2} - z^*\rangle\right] \ge \mathsf{E}\left[\langle F(z^{t+1/2}) - F(z^*), z^{t+1/2} - z\rangle\right] \ge \mu\mathsf{E}\left[\|z^{t+1/2} - z^*\|^2\right].$$

Let use a simple fact $\|z^{t+1/2} - z^*\|^2 \ge \frac{1}{2}\|z^t - z^*\|^2 - \|z^{t+1/2} - z^t\|^2$, then

$$\mathsf{E}\left[\langle F(z^{t+1/2}), z^{t+1/2} - z^*\rangle\right] \ge \frac{\mu}{2}\mathsf{E}\left[\|z^t - z^*\|^2\right] - \mu\mathsf{E}\left[\|z^{t+1/2} - z^t\|^2\right].$$

Then (44) gives

$$\begin{aligned}
\mathsf{E}\left[\|z^{t+1} - z^*\|^2\right] \le{}& (1 + 10\varepsilon_0 - \mu\gamma)\mathsf{E}\left[\|z^t - z^*\|^2\right] + \frac{10\sigma^2\gamma^2}{bM} \\
&+ (5L^2\gamma^2 + 2\gamma\mu + 8\varepsilon_0 - 1)\mathsf{E}\left[\|z^{t+1/2} - z^t\|^2\right] \\
&+ 20\varepsilon_0(1+\gamma L_{\max})^2 + 16\varepsilon_0\gamma\left(Q + 2L_{\max}\varepsilon_0 + \frac{\sigma}{\sqrt{bM}}\right) \\
&+ 40\gamma^2 L_{\max}^2\varepsilon_0^2 + 8\varepsilon_0^2.
\end{aligned}$$

With $\varepsilon_0 \le \min\left(\frac{1}{50}, \frac{\mu\gamma}{20}\right)$ and $\gamma \le \frac{1}{4L}$ we have

$$\begin{aligned}
\mathsf{E}\left[\|z^{t+1} - z^*\|^2\right] \le{}& \left(1 - \frac{\mu\gamma}{2}\right)\mathsf{E}\left[\|z^t - z^*\|^2\right] + \frac{10\sigma^2\gamma^2}{bM} \\
&+ 20\varepsilon_0(1+\gamma L_{\max})^2 + 16\varepsilon_0\gamma\left(Q + 2L_{\max}\varepsilon_0 + \frac{\sigma}{\sqrt{bM}}\right) \\
&+ 40\gamma^2 L_{\max}^2\varepsilon_0^2 + 8\varepsilon_0^2.
\end{aligned}$$

Let's run the recursion from 0 to $k-1$:

$$\mathbb{E}\left[\|z^k - z^*\|^2\right] \leq \left(1 - \frac{\mu\gamma}{2}\right)^k \mathbb{E}\left[\|z^0 - z^*\|^2\right] + \frac{20\sigma^2\gamma}{\mu bM} + \frac{2\varepsilon_0}{\mu\gamma}\left(20(1+\gamma L_{\max})^2 + 16\gamma\left(Q + 2L_{\max}\varepsilon_0 + \sqrt{\frac{\sigma}{bM}}\right) + 40\gamma^2 L_{\max}^2\varepsilon_0 + 8\varepsilon_0\right).$$

Then we carefully choose $\gamma = \min\left\{\frac{1}{4L}; \frac{2\ln(\max\{2; bM\mu^2\|z^0 - z^*\|^2 k/20\sigma^2\})}{\mu k}\right\}$ and $\varepsilon_0 = O(\varepsilon\mu\gamma(1 + Q + \gamma L_{\max})^2)$, where $\varepsilon = \max\left(\|z^0 - z^*\|^2 \exp\left(-\frac{\mu k}{8L}\right); \frac{\sigma^2}{\mu^2 bMk}\right)$. Then the output of the Algorithm 4 it is true

$$\mathbb{E}\left[\|\bar{z}^k - z^*\|^2\right] = \tilde{O}\left(\|z^0 - z^*\|^2 \exp\left(-\frac{\mu k}{8L}\right) + \frac{\sigma^2}{\mu^2 bMk}\right).$$

Substitute the batch size b and the number of iterations k from the description of the Algorithm 1:

$$\mathbb{E}\left[\|\bar{z}^k - z^*\|^2\right] = \tilde{O}\left(\|z^0 - z^*\|^2 \exp\left(-\frac{\mu}{8L} \cdot \frac{K}{P}\right) + \frac{\sigma^2}{\mu^2 MT}\right).$$

Corollary 1 ends the proof.

10.3. Non-convex-non-concave problems

Theorem 17 (Theorem 6) *Let $\{z_m^t\}_{t\geq 0}$ denote the iterates of Algorithm 2 for solving problem* (1)*. Let Assumptions 1(g), 1(l), 2(nc), 3 and 4 be satisfied. Then, if $\gamma \leq \frac{1}{4L}$ and $P = O\left(\sqrt{\chi}\log\frac{1}{\varepsilon}\right)$, we have the following estimate:*

$$\mathbb{E}\left[\frac{1}{k}\sum_{t=0}^{k-1}\|F(z^t)\|^2\right] = \tilde{O}\left(\frac{L^2\Omega_z^2\sqrt{\chi}}{K} + \frac{\sigma^2 K}{MT\sqrt{\chi}}\right).$$

Proof: We start from (43) with using diameter Ω_z:

$$2\gamma\mathbb{E}\left[\langle F(z^{t+1/2}), z^{t+1/2} - z^*\rangle\right] \leq \mathbb{E}\left[\|z^t - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] - (1 - 5L^2\gamma^2)\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] + \frac{10\sigma^2\gamma^2}{bM} + 16\varepsilon_0\gamma\left(L\Omega_z + 2L_{\max}\varepsilon_0 + \sqrt{\frac{\sigma}{bM}}\right) + 8\varepsilon_0(1+\gamma L_{\max})\Omega_z + 8(1+5\gamma^2 L_{\max}^2)\varepsilon_0^2.$$

With minty assumption it transforms to

$$0 \leq \mathbb{E}\left[\|z^t - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] - (1 - 5L^2\gamma^2)\mathbb{E}\left[\|z^{t+1/2} - z^t\|^2\right] + \frac{10\sigma^2\gamma^2}{bM}$$

$$+ 16\varepsilon_0\gamma\left(Q + L\Omega_z + 2L_{\max}\varepsilon_0 + \frac{\sigma}{\sqrt{bM}}\right) + 8\varepsilon_0(1+\gamma L_{\max})\Omega_z + 8(1+5\gamma^2L_{\max}^2)\varepsilon_0^2$$

$$= \mathbb{E}\left[\|z^t - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] - \gamma^2(1-5L^2\gamma^2)\mathbb{E}\left[\|g^t\|^2\right] + \frac{10\sigma^2\gamma^2}{bM}$$

$$+ 16\varepsilon_0\gamma\left(Q + L\Omega_z + 2L_{\max}\varepsilon_0 + \frac{\sigma}{\sqrt{bM}}\right) + 8\varepsilon_0(1+\gamma L_{\max})\Omega_z + 8(1+5\gamma^2L_{\max}^2)\varepsilon_0^2.$$

After the choice of $\gamma \leq \frac{1}{4L}$ we get

$$0 \leq \mathbb{E}\left[\|z^t - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] - \frac{\gamma^2}{2}\mathbb{E}\left[\|g^t\|^2\right] + \frac{10\sigma^2\gamma^2}{bM}$$

$$+ 16\varepsilon_0\gamma\left(Q + L\Omega_z + 2L_{\max}\varepsilon_0 + \frac{\sigma}{\sqrt{bM}}\right) + 8\varepsilon_0(1+\gamma L_{\max})\Omega_z + 8(1+5\gamma^2L_{\max}^2)\varepsilon_0^2.$$

The fact: $-\|g^t\|^2 \leq -\frac{1}{2}\|F(z^t)\|^2 + \|g^t - F(z^t)\|^2$, gives

$$0 \leq \mathbb{E}\left[\|z^t - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] - \frac{\gamma^2}{4}\mathbb{E}\left[\|F(z^t)\|^2\right] + \frac{\gamma^2}{2}\mathbb{E}\left[\|g^t - F(z^t)\|^2\right] + \frac{10\sigma^2\gamma^2}{bM}$$

$$+ 16\varepsilon_0\gamma\left(Q + L\Omega_z + 2L_{\max}\varepsilon_0 + \frac{\sigma}{\sqrt{bM}}\right) + 8\varepsilon_0(1+\gamma L_{\max})\Omega_z + 8(1+5\gamma^2L_{\max}^2)\varepsilon_0^2.$$

The term $\|\bar{g}^t - F(z^t)\|^2$ was estimated, when we deduced (39). Then

$$\frac{\gamma^2}{4}\mathbb{E}\left[\|F(z^t)\|^2\right] \leq \mathbb{E}\left[\|z^t - z^*\|^2\right] - \mathbb{E}\left[\|z^{t+1} - z^*\|^2\right] + \frac{11\sigma^2\gamma^2}{bM}$$

$$+ 16\varepsilon_0\gamma\left(Q + L\Omega_z + 2L_{\max}\varepsilon_0 + \frac{\sigma}{\sqrt{bM}}\right)$$

$$+ 8\varepsilon_0(1+\gamma L_{\max})\Omega_z + 8(1+6\gamma^2L_{\max}^2)\varepsilon_0^2.$$

Summing over all t from 0 to $k-1$:

$$\mathbb{E}\left[\frac{1}{k}\sum_{t=0}^{k-1}\left\|F(z^t)\right\|^2\right] \leq \frac{4\mathbb{E}\left[\|z^0 - z^*\|^2\right]}{\gamma^2 k} + \frac{44\sigma^2}{bM} + \frac{64\varepsilon_0}{\gamma}\left(Q + L\Omega_z + 2L_{\max}\varepsilon_0 + \frac{\sigma}{\sqrt{bM}}\right)$$

$$+ \frac{32\varepsilon_0}{\gamma^2}(1+\gamma L_{\max})\Omega_z + \frac{32}{\gamma^2}(1+6\gamma^2L_{\max}^2)\varepsilon_0^2.$$

Let $\gamma = \frac{1}{4L}$ and $\varepsilon_0 = \mathcal{O}\left(\frac{\varepsilon}{\Omega_z L_{\max} L}\right)$, where $\varepsilon = \max\left(\frac{L^2\Omega_z^2}{k}; \frac{\sigma^2}{bM}\right)$.Then for the output of the Algorithm 4 it holds

$$\mathbb{E}\left[\frac{1}{k}\sum_{t=0}^{k-1}\left\|F(z^t)\right\|^2\right] = \mathcal{O}\left(\frac{\mathbb{E}\left[L^2\|z^0 - z^*\|^2\right]}{k} + \frac{\sigma^2}{bM}\right).$$

Substitute the batch size b and the number of iterations k from the description of the Algorithm 2 and Corollary 1:

$$\mathbb{E}\left[\frac{1}{k}\sum_{t=0}^{k-1}\|F(z^t)\|^2\right] = \tilde{\mathcal{O}}\left(\frac{L^2\Omega_z^2\sqrt{\chi}}{K} + \frac{\sigma^2 K}{MT\sqrt{\chi}}\right).$$

11. Proof of Theorems from Section 5

Here we present a theoretical analysis of the proposed method. To begin with, we introduce auxiliary sequences that we need only in theoretical analysis (Algorithm 3 does not compute them):

$$\bar{z}^t = \frac{1}{M}\sum_{m=1}^{M} z_m^t, \qquad \bar{g}^t = \frac{1}{M}\sum_{m=1}^{M} F_m(z_m^t, \xi_m^t),$$
$$\bar{z}^{t+1/2} = \bar{z}^t - \gamma \bar{g}^t, \qquad \bar{z}^{t+1} = \bar{z}^t - \gamma \bar{g}^{t+1/2} \tag{45}$$

Such sequences are virtual, but at the communication moment $\bar{x}^t = x_m^t$ or $\bar{y}^t = y_m^t$.

11.1. Strongly convex-strongly concave problems

Theorem 18 (Theorem 7) *Let $\{z_m^t\}_{t\geq 0}$ denote the iterates of Algorithm 3 for solving problem* (1). *Let Assumptions 1(l), 2(sc), 3 and 5 be satisfied. Also let $H = \max_p |k_{p+1} - k_p|$ is a maximum distance between moments of communication ($k_p \in I$). Then, if $\gamma \leq \frac{1}{21HL_{\max}}$, we have the following estimate for the distance to the solution z^*:*

$$\mathbb{E}[\|\bar{z}^T - z^*\|^2] = \tilde{\mathcal{O}}\left(\exp\left(-\frac{\mu K}{42HL_{\max}}\right)\|z^0 - z^*\|^2 + \frac{\sigma^2}{\mu^2 MT} + \frac{(D^2H + \sigma^2)HL_{\max}^2}{\mu^4 T^2}\right).$$

Proof of Theorem: We start our proof with the following lemma.

Lemma 11. *Let $z, y \in \mathbb{R}^n$. We set $z^+ = z - y$, then for all $u \in \mathbb{R}^n$:*

$$\|z^+ - u\|^2 \leq \|z - u\|^2 - 2\langle y, z^+ - u\rangle - \|z^+ - z\|^2.$$

Proof: Simple manipulations give

$$\begin{aligned}
\|z^+ - u\|^2 &= \|z^+ - z + z - u\|^2 \\
&= \|z - u\|^2 + 2\langle z^+ - z, z - u\rangle + \|z^+ - z\|^2 \\
&= \|z - u\|^2 + 2\langle z^+ - z, z^+ - u\rangle - \|z^+ - z\|^2 \\
&= \|z - u\|^2 + 2\langle z^+ - (z - y), z^+ - u\rangle - 2\langle y, z^+ - u\rangle - \|z^+ - z\|^2 \\
&= \|z - u\|^2 - 2\langle y, z^+ - u\rangle - \|z^+ - z\|^2.
\end{aligned}$$

Applying this Lemma with $z = \bar{z}^{t+1}$, $z = \bar{z}^t$, $u = z^*$ and $y = \gamma \bar{g}^{t+1/2}$, we get

$$\|\bar{z}^{t+1} - z^*\|^2 = \|\bar{z}^t - z^*\|^2 - 2\gamma\langle \bar{g}^{t+1/2}, \bar{z}^{t+1} - z^*\rangle - \|\bar{z}^{t+1} - \bar{z}^t\|^2,$$

and with $z = \bar{z}^{t+1/2}$, $z = \bar{z}^t$, $u = z^{t+1}$, $y = \gamma \bar{g}^t$:

$$\|\bar{z}^{t+1/2} - \bar{z}^{t+1}\|^2 = \|\bar{z}^t - \bar{z}^{t+1}\|^2 - 2\gamma\langle \bar{g}^t, \bar{z}^{t+1/2} - \bar{z}^{t+1}\rangle - \|\bar{z}^{t+1/2} - \bar{z}^t\|^2.$$

Next, we sum up the two previous equalities

$$\begin{aligned}\|\bar{z}^{t+1} - z^*\|^2 + \|\bar{z}^{t+1/2} - \bar{z}^{t+1}\|^2 &= \|\bar{z}^t - z^*\|^2 - \|\bar{z}^{t+1/2} - \bar{z}^t\|^2 \\ &\quad -2\gamma\langle \bar{g}^{t+1/2}, \bar{z}^{t+1} - z^*\rangle - 2\gamma\langle \bar{g}^t, \bar{z}^{t+1/2} - \bar{z}^{t+1}\rangle.\end{aligned}$$

A small rearrangement gives

$$\begin{aligned}\|\bar{z}^{t+1} - z^*\|^2 + \|\bar{z}^{t+1/2} - \bar{z}^{t+1}\|^2 &= \|\bar{z}^t - z^*\|^2 - \|\bar{z}^{t+1/2} - \bar{z}^t\|^2 \\ &\quad -2\gamma\langle \bar{g}^{t+1/2}, \bar{z}^{t+1/2} - z^*\rangle + 2\gamma\langle \bar{g}^{t+1/2} - \bar{g}^t, \bar{z}^{t+1/2} - \bar{z}^{t+1}\rangle \\ &\leq \|\bar{z}^t - z^*\|^2 - \|\bar{z}^{t+1/2} - \bar{z}^t\|^2 \\ &\quad -2\gamma\langle \bar{g}^{t+1/2}, \bar{z}^{t+1/2} - z^*\rangle + \gamma^2\|\bar{g}^{t+1/2} - \bar{g}^t\|^2 \\ &\quad +\|\bar{z}^{t+1/2} - \bar{z}^{t+1}\|^2,\end{aligned}$$

Then we take the total expectation of both sides of the equation

$$\begin{aligned}\mathbb{E}\left[\|\bar{z}^{t+1} - z^*\|^2\right] &= \mathbb{E}\left[\|\bar{z}^t - z^*\|^2\right] - \mathbb{E}\left[\|\bar{z}^{t+1/2} - \bar{z}^t\|^2\right] \\ &\quad -2\gamma\mathbb{E}\left[\langle \bar{g}^{t+1/2}, \bar{z}^{t+1/2} - z^*\rangle\right] + \gamma^2\mathbb{E}\left[\|\bar{g}^{t+1/2} - \bar{g}^t\|^2\right]. \qquad (46)\end{aligned}$$

Further, we need to additionally estimate two terms $-2\gamma\langle \bar{g}^{t+1/2}, \bar{z}^{t+1/2} - z^*\rangle$ and $\|\bar{g}^{t+1/2} - \bar{g}^t\|^2$. For this we prove the following two lemmas, but before that we introduce the additional notation:

$$\mathrm{Err}(t) = \frac{1}{M}\sum_{m=1}^{M}\|\bar{z}^t - z_m^t\|^2. \qquad (47)$$

Lemma 12. *The following estimate is valid:*

$$-2\gamma\mathbb{E}\left[\langle \bar{g}^{t+1/2}, \bar{z}^{t+1/2} - z^*\rangle\right] \leq -\gamma\mu\mathbb{E}\left[\|\bar{z}^{t+1/2} - z^*\|^2\right] + \frac{\gamma L_{\max}^2}{\mu}\mathbb{E}\left[\mathrm{Err}(t + 1/2)\right]. \qquad (48)$$

Proof: We take into account the independence of all random vectors $\xi^i = (\xi_1^i, \ldots, \xi_m^i)$ and select only the conditional expectation $\mathbb{E}_{\xi^{t+1/2}}$ on vector $\xi^{t+1/2}$

$$-2\gamma\mathbb{E}\left[\langle \bar{g}^{t+1/2}, \bar{z}^{t+1/2} - z^*\rangle\right] = -2\gamma\mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\xi^{t+1/2}}[F_m(z_m^{t+1/2}, \xi_m^{t+1/2})], \bar{z}^{t+1/2} - z^*\right\rangle\right]$$

$$
\begin{aligned}
&\overset{(7)}{=} -2\gamma \mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M} F_m(z_m^{t+1/2}), \bar{z}^{t+1/2} - z^* \right\rangle\right] \\
&= -2\gamma \mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M} F_m(\bar{z}^{t+1/2}), \bar{z}^{t+1/2} - z^* \right\rangle\right] \\
&\quad +2\gamma \mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M} [F_m(\bar{z}^{t+1/2}) - F_m(z_m^{t+1/2})], \bar{z}^{t+1/2} - z^* \right\rangle\right] \\
&= -2\gamma \mathbb{E}\left[\left\langle F(\bar{z}^{t+1/2}), \bar{z}^{t+1/2} - z^* \right\rangle\right] \\
&\quad +2\gamma \mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M} [F_m(\bar{z}^{t+1/2}) - F_m(z_m^{t+1/2})], \bar{z}^{t+1/2} - z^* \right\rangle\right].
\end{aligned}
$$

Using property of z^*, we have:

$$
\begin{aligned}
-2\gamma \mathbb{E}\left[\langle \bar{g}^{t+1/2}, \bar{z}^{t+1/2} - z^* \rangle\right] &= -2\gamma \mathbb{E}\left[\left\langle F(\bar{z}^{t+1/2}) - F(z^*), \bar{z}^{t+1/2} - z^* \right\rangle\right] \\
&\quad +2\gamma \mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M} [F_m(\bar{z}^{t+1/2}) - F_m(z_m^{t+1/2})], \bar{z}^{t+1/2} - z^* \right\rangle\right] \\
&\overset{(5)}{\leq} -2\gamma\mu \mathbb{E}\left[\|\bar{z}^{t+1/2} - z^*\|^2\right] \\
&\quad +2\gamma \mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M} [F_m(\bar{z}^{t+1/2}) - F_m(z_m^{t+1/2})], \bar{z}^{t+1/2} - z^* \right\rangle\right].
\end{aligned}
$$

For $c > 0$ it is true that $2\langle a, b\rangle \leq \frac{1}{c}\|a\|^2 + c\|b\|^2$, then

$$
\begin{aligned}
-2\gamma \mathbb{E}\left[\langle \bar{g}^{t+1/2}, \bar{z}^{t+1/2} - z^* \rangle\right] &\leq -2\gamma\mu \mathbb{E}\left[\|\bar{z}^{t+1/2} - z^*\|^2\right] + \gamma\mu \mathbb{E}\left[\left\|\bar{z}^{t+1/2} - z^*\right\|^2\right] \\
&\quad + \frac{\gamma}{\mu}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}[F_m(\bar{z}^{t+1/2}) - F_m(z_m^{t+1/2})]\right\|^2\right] \\
&= -\gamma\mu \mathbb{E}\left[\|\bar{z}^{t+1/2} - z^*\|^2\right] \\
&\quad + \frac{\gamma}{\mu M^2}\mathbb{E}\left[\left\|\sum_{m=1}^{M}[F_m(\bar{z}^{t+1/2}) - F_m(z_m^{t+1/2})]\right\|^2\right] \\
&\overset{(15)}{\leq} -\gamma\mu \mathbb{E}\left[\|\bar{z}^{t+1/2} - z^*\|^2\right] \\
&\quad + \frac{\gamma}{\mu M}\mathbb{E}\left[\sum_{m=1}^{M}\left\|F_m(\bar{z}^{t+1/2}) - F_m(z_m^{t+1/2})\right\|^2\right] \\
&\overset{(4)}{\leq} -\gamma\mu \mathbb{E}\left[\|\bar{z}^{t+1/2} - z^*\|^2\right] + \frac{\gamma L_{\max}^2}{\mu M}\mathbb{E}\left[\sum_{m=1}^{M}\left\|\bar{z}^{t+1/2} - z_m^{t+1/2}\right\|^2\right].
\end{aligned}
$$

Definition (47) ends the proof.

Lemma 13. *The following estimate is valid:*

$$\mathsf{E}\left[\|\bar g^{t+1/2} - \bar g^t\|^2\right] \le 5L_{\max}^2 \mathsf{E}\left[\|\bar z^{t+1/2} - \bar z^t\|^2\right] + \frac{10\sigma^2}{M} + 5L_{\max}^2 \mathsf{E}\left[\mathrm{Err}(t+1/2)\right] + 5L_{\max}^2 \mathsf{E}\left[\mathrm{Err}(t)\right]. \tag{49}$$

Proof:

$$\begin{aligned}
&\mathsf{E}\left[\|\bar g^{t+1/2}-\bar g^t\|^2\right] \\
&= \mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M F_m(z_m^{t+1/2}, \xi_m^{t+1/2}) - \frac{1}{M}\sum_{m=1}^M F_m(z_m^t, \xi_m^t)\right\|^2\right] \\
&\overset{(15)}{\le} 5\mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^{t+1/2}, \xi_m^{t+1/2}) - F_m(z_m^{t+1/2})]\right\|^2\right] \\
&\quad + 5\mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^t, \xi_m^k) - F_m(z_m^t)]\right\|^2\right] \\
&\quad + 5\mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^{t+1/2}) - F_m(\bar z^{t+1/2})]\right\|^2\right] \\
&\quad + 5\mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^t) - F_m(\bar z^t)]\right\|^2\right] \\
&\quad + 5\mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(\bar z^{t+1/2}) - F_m(\bar z^t)]\right\|^2\right] \\
&\overset{(15)}{\le} 5\mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^{t+1/2}, \xi_m^{t+1/2}) - F_m(z_m^{t+1/2})]\right\|^2\right] \\
&\quad + 5\mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^t, \xi_m^k) - F_m(z_m^t)]\right\|^2\right] \\
&\quad + \frac{5}{M}\sum_{m=1}^M \mathsf{E}\left[\|F_m(z_m^{t+1/2}) - F_m(\bar z^{t+1/2})\|^2\right] \\
&\quad + \frac{5}{M}\sum_{m=1}^M \mathsf{E}\left[\|F_m(z_m^t) - F_m(\bar z^t)\|^2\right] + 5\mathsf{E}\left[\|F(\bar z^{t+1/2}) - F(\bar z^t)\|^2\right] \\
&\overset{(4),(47)}{\le} 5\mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^{t+1/2}, \xi_m^{t+1/2}) - F_m(z_m^{t+1/2})]\right\|^2\right] \\
&\quad + 5\mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^t, \xi_m^k) - F_m(z_m^t)]\right\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
& +5L_{\max}^2 \mathsf{E}\left[\mathrm{Err}(t+1/2)\right] + 5L_{\max}^2 \mathsf{E}\left[\mathrm{Err}(t)\right] + 5L_{\max}^2 \mathsf{E}\left[\|\bar{z}^{t+1/2} - \bar{z}^t\|^2\right] \\
= \quad & 5\mathsf{E}\left[\mathsf{E}_{\xi_{t+1/2}}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}[F_m(z_m^{t+1/2}, \xi_m^{t+1/2}) - F_m(z_m^{t+1/2})]\right\|^2\right]\right] \\
& +5\mathsf{E}\left[\mathsf{E}_{\xi_t}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}[F_m(z_m^t, \xi_m^t) - F_m(z_m^t)]\right\|^2\right]\right] \\
& +5L_{\max}^2 \mathsf{E}\left[\mathrm{Err}(t+1/2)\right] + 5L_{\max}^2 \mathsf{E}\left[\mathrm{Err}(t)\right] + 5L_{\max}^2 \mathsf{E}\left[\|\bar{z}^{t+1/2} - \bar{z}^t\|^2\right].
\end{aligned}
$$

Using independence of each machine and (7), we get:

$$
\begin{aligned}
\mathsf{E}\left[\|\bar{g}^{t+1/2} - \bar{g}^t\|^2\right] \quad \leq \quad & \frac{10\sigma^2}{M} + 5L_{\max}^2 \mathsf{E}\left[\mathrm{Err}(t+1/2)\right] \\
& +5L_{\max}^2 \mathsf{E}\left[\mathrm{Err}(t)\right] + 5L_{\max}^2 \mathsf{E}\left[\|\bar{z}^{t+1/2} - \bar{z}^t\|^2\right].
\end{aligned}
$$

We are now ready to combine (46), (48), (49) and get

$$
\begin{aligned}
\mathsf{E}\left[\|\bar{z}^{t+1} - z^*\|^2\right] \quad \leq \quad & \mathsf{E}\left[\|\bar{z}^t - z^*\|^2\right] - \mathsf{E}\left[\|\bar{z}^{t+1/2} - \bar{z}^t\|^2\right] \\
& -\gamma\mu \mathsf{E}\left[\|\bar{z}^{t+1/2} - z^*\|^2\right] + \frac{\gamma L_{\max}^2}{\mu}\mathsf{E}\left[\mathrm{Err}(t+1/2)\right] \qquad (50) \\
& +5\gamma^2 L_{\max}^2 \mathsf{E}\left[\|\bar{z}^{t+1/2} - \bar{z}^t\|^2\right] + \frac{10\gamma^2\sigma^2}{M} \\
& +5\gamma^2 L_{\max}^2 \mathsf{E}\left[\mathrm{Err}(t+1/2)\right] + 5\gamma^2 L_{\max}^2 \mathsf{E}\left[\mathrm{Err}(t)\right]. \qquad (51)
\end{aligned}
$$

Together with $-\|\bar{z}^{t+1/2} - z^*\|^2 \leq \|\bar{z}^{t+1/2} - \bar{z}^t\|^2 - 1/2\|\bar{z}^t - z^*\|^2$ it transforms to

$$
\begin{aligned}
\mathsf{E}\left[\|\bar{z}^{t+1} - z^*\|^2\right] \quad \leq \quad & \left(1 - \frac{\mu\gamma}{2}\right)\mathsf{E}\left[\|\bar{z}^t - z^*\|^2\right] + \frac{10\gamma^2\sigma^2}{M} \\
& +(\mu\gamma + 5\gamma^2 L_{\max}^2 - 1)\|\bar{z}^{t+1/2} - \bar{z}^t\|^2 \\
& +\frac{\gamma L_{\max}^2}{\mu}\mathsf{E}\left[\mathrm{Err}(t+1/2)\right] + 5\gamma^2 L_{\max}^2 \mathsf{E}\left[\mathrm{Err}(t+1/2)\right] \\
& +5\gamma^2 L_{\max}^2 \mathsf{E}\left[\mathrm{Err}(t)\right].
\end{aligned}
$$

Taking $\gamma \leq \frac{1}{6HL_{\max}}$ gives

$$
\begin{aligned}
\mathsf{E}\left[\|\bar{z}^{t+1} - z^*\|^2\right] \quad \leq \quad & \left(1 - \frac{\mu\gamma}{2}\right)\mathsf{E}\left[\|\bar{z}^t - z^*\|^2\right] + \frac{10\gamma^2\sigma^2}{M} \\
& +\frac{7\gamma L_{\max}^2}{\mu}\mathsf{E}\left[\mathrm{Err}(t+1/2)\right] + 5\gamma^2 L_{\max}^2 \mathsf{E}\left[\mathrm{Err}(t)\right]. \qquad (52)
\end{aligned}
$$

It remains to estimate $\mathsf{E}\left[\mathrm{Err}(t+1/2)\right]$ and $\mathsf{E}\left[\mathrm{Err}(t)\right]$.

Lemma 14. *For $t \in [t_p + 1, t_{p+1}]$ the following estimate is valid:*

$$\mathbb{E}\left[\mathrm{Err}(t+1/2)\right] \le 216(D^2H+\sigma^2)H\gamma^2. \tag{53}$$

Proof: First, let us look at the nearest past consensus point $t_p < t$, then $z_m^{t_p+1} = \bar{z}^{t_p+1}$:

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Err}(t+1/2)\right] &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|\bar{z}^{t+1/2} - z_m^{t+1/2}\|^2 \\
&= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|\bar{z}^{t+1/2} - \bar{z}^{t_p} + z_m^{t_p} - z_m^{t+1/2}\|^2 \\
&= \frac{\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\left\|F_m(z_m^t,\xi_m^t) - \bar{g}^t + \sum_{k=t_p+1}^{t-1}\left[F_m(z_m^{k+1/2},\xi_m^{k+1/2}) - \bar{g}^{k+1/2}\right]\right\|^2.
\end{aligned}$$

Only $\bar{g}^t$ and $F_m(z_m^k,\xi_m^k)$ depend on ξ^k, as well as the unbiasedness of $\bar{g}^t$ and $F_m(z_m^k,\xi_m^k)$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Err}(t+1/2)\right] = \frac{\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}&\left\|-\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^t) - \sum_{k=t_p+1}^{t-1}\bar{g}^{k+1/2} + F_m(z_m^t) + \sum_{k=t_p+1}^{t-1}F_m(z_m^{k+1/2},\xi_m^{k+1/2})\right\|^2 \\
&+ \frac{\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^t) - \bar{g}^t - F_m(z_m^t) + F_m(z_m^t,\xi_m^t)\right\|^2.
\end{aligned}$$

We want to continue the same way, but note that z_i^t depends on $\xi^{k-1+1/2}$, then let us make the estimate rougher than in previous case

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Err}(t+1/2)\right] \le (1+\beta_0)&\frac{\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\left\|-\sum_{k=t_p+1}^{t-1}\bar{g}^{k+1/2} + \sum_{k=t_p+1}^{t-1}F_m(z_m^{k+1/2},\xi_m^{k+1/2})\right\|^2 \\
&+ (1+\beta_0^{-1})\frac{\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\left\|-\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^t) + F_m(z_m^t)\right\|^2 \\
&+ \frac{\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^t) - \bar{g}^t - F_m(z_m^t) + F_m(z_m^t,\xi_m^t)\right\|^2.
\end{aligned}$$

Here β_0 is some positive constant, which we define later. Then

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Err}(t+1/2)\right] \\
&\le (1+\beta_0)\frac{\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\Bigg\| - \frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{t-1+1/2}) - \sum_{k=t_p+1}^{t-2}\bar{g}^{k+1/2} \\
&\qquad + F_m(z_m^{t-1+1/2},\xi_m^{t-1+1/2}) + \sum_{k=t_p+1}^{t-2}F_m(z_m^{k+1/2},\xi_m^{k+1/2})\Bigg\|^2 \\
&\quad + (1+\beta_0^{-1})\frac{\gamma^2}{M}\sum_{m=1}^{M}\mathbb{E}\left\|-\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^t) + F_m(z_m^t)\right\|^2
\end{aligned}$$

$$+ (1+\beta_0)\frac{\gamma^2}{M}\sum_{m=1}^{M}\mathsf{E}\left\|\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{t-1+1/2}) - \bar{g}^{t-1+1/2} - F_m(z_m^{t-1+1/2}) + F_m(z_m^{t-1+1/2},\xi_m^{t-1+1/2})\right\|^2$$
$$+ \frac{\gamma^2}{M}\sum_{m=1}^{M}\mathsf{E}\left\|\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{t}) - \bar{g}^{t} - F_m(z_m^{t}) + F_m(z_m^{t},\xi_m^{t})\right\|^2.$$

and

$$\begin{aligned}
&\mathsf{E}\left[\mathrm{Err}(t+1/2)\right]\\
&\leq (1+\beta_0)(1+\beta_1)\frac{\gamma^2}{M}\sum_{m=1}^{M}\mathsf{E}\left\| - \sum_{k=t_p+1}^{t-2}\bar{g}^{k+1/2} + \sum_{k=t_p+1}^{t-2}F_m(z_m^{k+1/2},\xi_m^{k+1/2})\right\|^2\\
&\quad + (1+\beta_0)(1+\beta_1^{-1})\frac{\gamma^2}{M}\sum_{m=1}^{M}\mathsf{E}\left\| - \frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{t-1+1/2}) + F_m(z_m^{t-1+1/2})\right\|^2\\
&\quad + (1+\beta_0^{-1})\frac{\gamma^2}{M}\sum_{m=1}^{M}\mathsf{E}\left\|-\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{t}) + F_m(z_m^{t})\right\|^2\\
&\quad + (1+\beta_0)\frac{\gamma^2}{M}\sum_{m=1}^{M}\mathsf{E}\left\|\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{t-1+1/2}) - \bar{g}^{t-1+1/2} - F_m(z_m^{t-1+1/2}) + F_m(z_m^{t-1+1/2},\xi_m^{t-1+1/2})\right\|^2\\
&\quad + \frac{\gamma^2}{M}\sum_{m=1}^{M}\mathsf{E}\left\|\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{t}) - \bar{g}^{t} - F_m(z_m^{t}) + F_m(z_m^{t},\xi_m^{t})\right\|^2.
\end{aligned}$$

One can continue this way for all terms, setting $\beta_i = \frac{1}{\alpha - i - 1}$, where $\alpha = 4H$. Then for all $i = 0, \ldots, (t - t_p - 1)$

$$(1+\beta_0)(1+\beta_1)(1+\beta_2)\ldots(1-\beta_i) = \frac{\alpha}{\alpha - i - 1}.$$

Note that $t - t_p \leq 2H$, hence for all $i = 0, \ldots, (t - t_p - 1)$

$$(1+\beta_0)(1+\beta_1)(1+\beta_2)\ldots(1+\beta_i) \leq (1+\beta_1)(1+\beta_2)\ldots(1+\beta_{t-t_p-1}) \leq \frac{\alpha}{\alpha - 2H} \leq 2.$$

Additionally, $1 + \beta_i^{-1} \leq \alpha$, then $(\alpha = 4H)$

$$\begin{aligned}
&\mathsf{E}\left[\mathrm{Err}(t+1/2)\right]\\
&\leq \frac{2\alpha\gamma^2}{M}\sum_{k=t_p+1}^{t-1}\sum_{m=1}^{M}\mathsf{E}\left\| - \frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{k+1/2}) + F_m(z_m^{k+1/2})\right\|^2\\
&\quad + \frac{2\alpha\gamma^2}{M}\sum_{m=1}^{M}\mathsf{E}\left\|-\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{t}) + F_m(z_m^{t})\right\|^2\\
&\quad + \frac{2\gamma^2}{M}\sum_{k=t_p+1}^{t-1}\sum_{m=1}^{M}\mathsf{E}\left\|\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{k+1/2}) - \bar{g}^{k+1/2} - F_m(z_m^{k+1/2}) + F_m(z_m^{k+1/2},\xi_m^{k+1/2})\right\|^2
\end{aligned}$$

$$+\frac{2\gamma^2}{M}\sum_{m=1}^{M}\mathsf{E}\left\|\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^t)-\bar{g}^t-F_m(z_m^t)+F_m(z_m^t,\xi_m^t)\right\|^2$$

$$=\frac{8\gamma^2 H}{M}\sum_{k=t_p+1}^{t-1}\sum_{m=1}^{M}\mathsf{E}\left\|-\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{k+1/2})+F_m(z_m^{k+1/2})\right\|^2$$

$$+\frac{8\gamma^2 H}{M}\sum_{m=1}^{M}\mathsf{E}\left\|-\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^t)+F_m(z_m^t)\right\|^2$$

$$+\frac{8\gamma^2}{M}\sum_{k=t_p+1}^{t-1}\sum_{m=1}^{M}\mathsf{E}\left\|\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{k+1/2})-\bar{g}^{k+1/2}-F_m(z_m^{k+1/2})+F_m(z_m^{k+1/2},\xi_m^{k+1/2})\right\|^2$$

$$+\frac{8\gamma^2}{M}\sum_{m=1}^{M}\mathsf{E}\left\|\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^t)-\bar{g}^t-F_m(z_m^t)+F_m(z_m^t,\xi_m^t)\right\|^2.$$

It remains to estimate

$$\frac{1}{M}\sum_{m=1}^{M}\mathsf{E}\left\|-\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{k+1/2})+F_m(z_m^{k+1/2})\right\|^2$$

$$\overset{(15)}{\leq}\frac{3}{M}\sum_{m=1}^{M}\mathsf{E}\left\|-\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{k+1/2})+\frac{1}{M}\sum_{i=1}^{M}F_i(\bar{z}^{k+1/2})\right\|^2$$

$$+\frac{3}{M}\sum_{m=1}^{M}\mathsf{E}\left\|-\frac{1}{M}\sum_{i=1}^{M}F_i(\bar{z}^{k+1/2})+F_m(\bar{z}^{k+1/2})\right\|^2$$

$$+\frac{3}{M}\sum_{m=1}^{M}\mathsf{E}\left\|-F_m(\bar{z}^{k+1/2})+F_m(z_m^{k+1/2})\right\|^2$$

$$\overset{(13)}{\leq}\frac{6}{M}\sum_{m=1}^{M}\mathsf{E}\left\|-F_m(\bar{z}^{k+1/2})+F_m(z_m^{k+1/2})\right\|^2+3D^2$$

$$\overset{(4)}{\leq}\frac{6L_{\max}^2}{M}\sum_{m=1}^{M}\mathsf{E}\|\bar{z}^{k+1/2}-z_m^{k+1/2}\|^2+3D^2$$

$$=6L_{\max}^2\mathsf{E}\left[\mathrm{Err}(k+1/2)\right]+3D^2$$

and

$$\frac{1}{M}\sum_{m=1}^{M}\mathsf{E}\left\|\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{k+1/2})-\bar{g}^{k+1/2}-F_m(z_m^{k+1/2})+F_m(z_m^{k+1/2},\xi_m^{k+1/2})\right\|^2$$

$$\overset{(15)}{\leq}2\left[\mathsf{E}\left\|\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{k+1/2})-\bar{g}^{k+1/2}\right\|^2\right.$$

$$\left.+\frac{2}{M}\sum_{m=1}^{M}\mathsf{E}\left\|F_m(z_m^{k+1/2})+F_m(z_m^{k+1/2},\xi_m^{k+1/2})\right\|^2\right]$$

$$\overset{(7)}{\leq}4\sigma^2.$$

Finally, we get

$$\mathsf{E}\left[\mathrm{Err}(t+1/2)\right] \leq 48\gamma^2 L_{\max}^2 H \sum_{k=t_p+1}^{t-1} \mathsf{E}\left[\mathrm{Err}(k+1/2)\right] + 48\gamma^2 L_{\max}^2 H \mathsf{E}\left[\mathrm{Err}(t)\right] + 32(D^2 H + \sigma^2) \sum_{k=t_p+1}^{t-1} \gamma^2 + 32\gamma^2\left(\sigma^2 + D^2\right). \qquad (54)$$

The estimate for $\mathsf{E}\left[\mathrm{Err}(t+1/3)\right]$ is done in a similar way:

$$\mathsf{E}\left[\mathrm{Err}(t)\right] \leq 48\gamma^2 L_{\max}^2 H \sum_{k=t_p+1}^{t-1} \mathsf{E}\left[\mathrm{Err}(k+1/2)\right] + 32(D^2 H + \sigma^2) \sum_{k=t_p+1}^{t-1} \gamma^2. \qquad (55)$$

Substituting $\mathsf{E}\left[\mathrm{Err}(t)\right]$ to $\mathsf{E}\left[\mathrm{Err}(t+1/2)\right]$, we get

$$\begin{aligned}\mathsf{E}\left[\mathrm{Err}(t+1/2)\right] \leq\ & 48\gamma^2 L_{\max}^2 H \sum_{k=t_p+1}^{t-1} \mathsf{E}\left[\mathrm{Err}(k+1/2)\right] \\ & + 48\gamma^2 L_{\max}^2 H \left(48\gamma^2 L_{\max}^2 H \sum_{k=t_p+1}^{t-1} \mathsf{E}\left[\mathrm{Err}(k+1/2)\right] + 32(D^2 H + \sigma^2) \sum_{k=t_p+1}^{t-1} \gamma^2\right) \\ & + 32(D^2 H + \sigma^2) \sum_{k=t_p+1}^{t-1} \gamma^2 + 32\gamma^2\left(\sigma^2 + D^2\right).\end{aligned}$$

With $\gamma \leq \frac{1}{21 H L_{\max}}$

$$\mathsf{E}\left[\mathrm{Err}(t+1/2)\right] \leq \frac{1}{8H} \sum_{k=t_p+1}^{t-1} \mathsf{E}\left[\mathrm{Err}(k+1/2)\right] + 72(D^2 H + \sigma^2)\gamma^2 (t - t_p - 1).$$

Let us run the recursion:

$$\begin{aligned}\mathsf{E}\left[\mathrm{Err}(t+1/2)\right] &\leq \frac{1}{8H}\left(1 + \frac{1}{8H}\right) \sum_{k=t_p+1}^{t-2} \mathsf{E}\left[\mathrm{Err}(k+1/2)\right] \\ &\quad + \frac{1}{8H} \cdot 72(D^2 H + \sigma^2)\gamma^2 (t - t_p - 2)\gamma^2 (t - t_p - 1) \\ &\leq 72(D^2 H + \sigma^2)\gamma^2 \sum_{k=t_p+1}^{t-1} \left(1 + \frac{1}{8H}\right)^{t-1-j}.\end{aligned}$$

Then one can note that $\left(1 + \frac{1}{8H}\right)^{t-1-j} \leq \left(1 + \frac{1}{2H}\right)^{2H} \leq \exp(1) \leq 3$ and then

$$\mathsf{E}\left[\mathrm{Err}(t+1/2)\right] \leq 216(D^2 H + \sigma^2) \sum_{k=t_p+1}^{t-1} \gamma^2 \leq 216(D^2 H + \sigma^2) H \gamma^2.$$

Note that in the general case $\mathsf{E}\left[\mathrm{Err}(t+1/3)\right]$ may be less than $\mathsf{E}\left[\mathrm{Err}(t)\right]$, but since recurrent (54) is stronger than (55), we assume for simplicity that $\mathsf{E}\left[\mathrm{Err}(k+1/3)\right] \geq \mathsf{E}\left[\mathrm{Err}(k)\right]$. Than (52) can be rewritten as

$$\begin{aligned} \mathsf{E}\left[k\bar z^{t+1} - z^* k^2\right] &\leq \left(1 - \frac{\mu\gamma}{2}\right)\mathsf{E}\left[k\bar z^{t} - z^* k^2\right] + \frac{10\gamma^2\sigma^2}{M} \\ &\quad + \left(\frac{7\gamma L_{\max}^2}{\mu} + 5\gamma^2 L_{\max}^2\right)\mathsf{E}\left[\mathrm{Err}(t+1/2)\right] \\ &\leq \left(1 - \frac{\mu\gamma}{2}\right)\mathsf{E}\left[k\bar z^{t} - z^* k^2\right] + \frac{10\gamma^2\sigma^2}{M} \\ &\quad + \left(\frac{7\gamma L_{\max}^2}{\mu} + 5\gamma^2 L_{\max}^2\right)\left(216(D^2H+\sigma^2)H\gamma^2\right). \end{aligned}$$

Running the recursion, we obtain:

$$\mathsf{E}\left[k\bar z^{T} - z^* k^2\right] = O\left(\left(1 - \frac{\mu\gamma}{2}\right)^T k z^0 - z^* k^2 + \frac{\gamma\sigma^2}{\mu M} + \frac{\gamma^2(D^2H+\sigma^2)HL_{\max}^2}{\mu^2}\right),$$

or

$$\mathsf{E}\left[k\bar z^{T} - z^* k^2\right] = O\left(\exp\left(-\frac{\mu\gamma T}{2}\right) k z^0 - z^* k^2 + \frac{\gamma\sigma^2}{\mu M} + \frac{\gamma^2(D^2H+\sigma^2)HL_{\max}^2}{\mu^2}\right)$$

Finally, we need tuning of $\gamma \leq \frac{1}{21HL_{\max}}$:

· If $\frac{1}{21HL_{\max}} \geq \frac{2\ln\left(\max f 2, \mu k z^0 - z^* k^2 TM/\sigma^2 g\right)}{\mu T}$ then $\gamma = \frac{2\ln\left(\max f 2, \mu k z^0 - z^* k^2 TM/\sigma^2 g\right)}{\mu T}$ gives

$$\begin{aligned} &\tilde O\left(\exp\left(-\ln\left(\max f 2, \mu k z^0 - z^* k^2 TM/\sigma^2 g\right)\right) k z^0 - z^* k^2 + \frac{\sigma^2}{\mu^2 MT} + \frac{(D^2H+\sigma^2)HL_{\max}^2}{\mu^4T^2}\right) \\ &\qquad = \tilde O\left(\frac{\sigma^2}{\mu^2 MT} + \frac{(D^2H+\sigma^2)HL_{\max}^2}{\mu^4T^2}\right) \end{aligned}$$

· If $\frac{1}{21HL_{\max}} \leq \frac{2\ln\left(\max f 2, \mu k z^0 - z^* k^2 TM/\sigma^2 g\right)}{\mu T}$ then $\gamma = \frac{1}{21HL_{\max}}$ gives

$$\begin{aligned} &\tilde O\left(\exp\left(-\frac{\mu T}{42HL_{\max}}\right) k z^0 - z^* k^2 + \frac{\gamma\sigma^2}{\mu M} + \frac{\gamma^2(D^2H+\sigma^2)HL_{\max}^2}{\mu^2}\right) \\ &\leq \tilde O\left(\exp\left(-\frac{\mu T}{42HL_{\max}}\right) k z^0 - z^* k^2 + \frac{\sigma^2}{\mu^2 MT} + \frac{(D^2H+\sigma^2)HL_{\max}^2}{\mu^4T^2}\right). \end{aligned}$$

11.2. Non-convex-non-concave problems

Theorem 19 (Theorem 7) *Let $f z_m^t g_{t\geq 0}$ denote the iterates of Algorithm 3 for solving problem* (1). *Let Assumptions 1(l), 2(nc), 3 and 5 be satisfied. Also let* $H = \max_p j k_{p+1} - k_p j$

is a maximum distance between moments of communication ($k_p \in I$) and $\|\bar{z}^t\| \leq \Omega$ (for all t). Then, if $\gamma \leq \frac{1}{4L_{\max}}$, we have the following estimate:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|F(\bar{z}^t)\|^2\right] = \mathcal{O}\Bigg(\frac{L_{\max}^2\|\bar{z}^0 - z^*\|^2}{T} + \frac{(L_{\max}\Omega(D^2H+\sigma^2)H)^{2/3}}{T^{1/3}} + \frac{\sigma^2}{M} + L_{\max}\Omega\sqrt{(D^2H+\sigma^2)H}\Bigg).$$

Proof: Most of the necessary estimates have already been made in the previous subsection. In particular, Lemmas 13 and 14 are valid for us. But Lemma 12 needs modification:

Lemma 15. *It holds:*

$$-2\gamma\mathbb{E}\left[\langle\bar{g}^{t+1/2}, \bar{z}^{t+1/2} - z^*\rangle\right] \leq 2\gamma L_{\max}\sqrt{\mathbb{E}\left[\|\bar{z}^{t+1/2} - z^*\|^2\right]}\sqrt{\mathbb{E}\left[\mathrm{Err}(t+1/2)\right]} + \gamma L_{\max}\mathbb{E}\left[\|\bar{z}^{t+1/2} - \bar{z}^t\|^2\right] + \gamma L\mathbb{E}\left[\mathrm{Err}(t+1/2)\right]. \quad (56)$$

Proof: First of all, we use the independence of all random vectors $\xi^i = (\xi_1^i, \ldots, \xi_m^i)$ and select only the conditional expectation $\mathbb{E}_{\xi^{t+1/2}}$ on vector $\xi^{t+1/2}$ and get the following chain of inequalities:

$$\begin{aligned}
-2\gamma\mathbb{E}\left[\langle\bar{g}^{t+1/2}, \bar{z}^{t+1/2} - z^*\rangle\right] &= -2\gamma\mathbb{E}\left[\left\langle\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\xi^{k+1/2}}[F_m(z_m^{t+1/2}, \xi_m^{t+1/2})], \bar{z}^{t+1/2} - z^*\right\rangle\right] \\
&\overset{(7)}{=} -2\gamma\mathbb{E}\left[\left\langle\frac{1}{M}\sum_{m=1}^{M}F_m(z_m^{t+1/2}), \bar{z}^{t+1/2} - z^*\right\rangle\right] \\
&= -2\gamma\mathbb{E}\left[\left\langle\frac{1}{M}\sum_{m=1}^{M}F_m(\bar{z}^{t+1/2}), \bar{z}^{t+1/2} - z^*\right\rangle\right] \\
&\quad + 2\gamma\mathbb{E}\left[\left\langle\frac{1}{M}\sum_{m=1}^{M}[F_m(\bar{z}^{t+1/2}) - F_m(z_m^{t+1/2})], \bar{z}^{t+1/2} - z^*\right\rangle\right] \\
&= -2\gamma\mathbb{E}\left[\langle F(\bar{z}^{t+1/2}), \bar{z}^{t+1/2} - z^*\rangle\right] \\
&\quad + 2\gamma\mathbb{E}\left[\left\langle\frac{1}{M}\sum_{m=1}^{M}[F_m(\bar{z}^{t+1/2}) - F_m(z_m^{t+1/2})], \bar{z}^{t+1/2} - z^*\right\rangle\right] \\
&\overset{(6)}{\leq} 2\gamma\mathbb{E}\left[\left\langle\frac{1}{M}\sum_{m=1}^{M}[F_m(\bar{z}^{t+1/2}) - F_m(z_m^{k+1/2})], \bar{z}^{t+1/2} - z^*\right\rangle\right] \\
&\leq 2\gamma\mathbb{E}\left[\|\bar{z}^{t+1/2} - z^*\| \cdot \left\|\frac{1}{M}\sum_{m=1}^{M}F_m(\bar{z}^{t+1/2}) - F_m(z_m^{t+1/2})\right\|\right]
\end{aligned}$$

$$\leq 2\gamma \mathbb{E}\left[\|\bar{z}^{t+1/2} - z^*\| \cdot \frac{1}{M}\sum_{m=1}^{M}\left\|F_m(\bar{z}^{t+1/2}) - F_m(z_m^{t+1/2})\right\|\right]$$

$$\overset{(4)}{\leq} 2\gamma L_{\max}\mathbb{E}\left[\|\bar{z}^{t+1/2} - z^*\| \cdot \frac{1}{M}\sum_{m=1}^{M}\left\|z_m^{t+1/2} - \bar{z}^{t+1/2}\right\|\right]$$

$$\begin{aligned}
-2\gamma \mathbb{E}&\left[\langle \bar{g}^{t+1/2}, \bar{z}^{t+1/2} - z^*\rangle\right] \\
&\leq 2\gamma L_{\max}\mathbb{E}\left[\|\bar{z}^{t} - z^*\| \cdot \frac{1}{M}\sum_{m=1}^{M}\left\|z_m^{t+1/2} - \bar{z}^{t+1/2}\right\|\right] \\
&\quad + 2\gamma L_{\max}\mathbb{E}\left[\|\bar{z}^{t+1/2} - \bar{z}^t\| \cdot \frac{1}{M}\sum_{m=1}^{M}\left\|z_m^{t+1/2} - \bar{z}^{t+1/2}\right\|\right] \\
&\leq 2\gamma L_{\max}\sqrt{\mathbb{E}\left[\|\bar{z}^t - z^*\|^2\right]} \cdot \sqrt{\mathbb{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M}\left\|z_m^{t+1/2} - \bar{z}^{t+1/2}\right\|\right)^2\right]} \\
&\quad + \gamma L_{\max}\mathbb{E}\left[\|\bar{z}^{t+1/2} - \bar{z}^t\|^2\right] + \gamma L_{\max}\mathbb{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M}\|\bar{z}^{t+1/2} - z_m^{t+1/2}\|\right)^2\right].
\end{aligned}$$

By (15) it is easy to see that

$$\mathbb{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M}\|\bar{z}^{t+1/2} - z_m^{t+1/2}\|\right)^2\right] \leq \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M}\|\bar{z}^{t+1/2} - z_m^{t+1/2}\|^2\right].$$

This completes the proof.

Then we have the same as (50):

$$\begin{aligned}
\mathbb{E}\left[\|\bar{z}^{t+1} - z^*\|^2\right] &\leq \mathbb{E}\left[\|\bar{z}^{t} - z^*\|^2\right] - \mathbb{E}\left[\|\bar{z}^{t+1/2} - \bar{z}^t\|^2\right] \\
&\quad + 2\gamma L_{\max}\sqrt{\mathbb{E}\left[\|\bar{z}^t - z^*\|^2\right]}\sqrt{\mathbb{E}\left[\mathrm{Err}(t+1/2)\right]} \\
&\quad + \gamma L_{\max}\mathbb{E}\left[\|\bar{z}^{t+1/2} - \bar{z}^t\|^2\right] + \gamma L_{\max}\mathbb{E}\left[\mathrm{Err}(t+1/2)\right] \\
&\quad + \gamma^2\left(5L_{\max}^2\mathbb{E}\left[\|\bar{z}^{t+1/2} - \bar{z}^t\|^2\right] + \frac{10\sigma^2}{M} + 5L_{\max}^2\mathbb{E}\left[\mathrm{Err}(t+1/2)\right] + 5L_{\max}^2\mathbb{E}\left[\mathrm{Err}(t)\right]\right).
\end{aligned}$$

Choosing $\gamma \leq \frac{1}{4L_{\max}}$ gives

$$\begin{aligned}
\frac{1}{2}\mathbb{E}\left[\|\bar{z}^{t+1/2} - \bar{z}^t\|^2\right] &\leq \mathbb{E}\left[\|\bar{z}^{t} - z^*\|^2\right] - \mathbb{E}\left[\|\bar{z}^{t+1} - z^*\|^2\right] \\
&\quad + 2\gamma L_{\max}\sqrt{\mathbb{E}\left[\|\bar{z}^t - z^*\|^2\right]}\sqrt{\mathbb{E}\left[\mathrm{Err}(t+1/2)\right]} \\
&\quad + (5\gamma^2 L_{\max}^2 + \gamma L_{\max})\mathbb{E}\left[\mathrm{Err}(t+1/2)\right] + 5\gamma^2 L_{\max}^2\mathbb{E}\left[\mathrm{Err}(t)\right] + \frac{10\gamma^2\sigma^2}{M}.
\end{aligned}$$

Next we work with

$$
\begin{aligned}
&\mathsf{E}\left[k\bar{z}^{t+1/2} - \bar{z}^t k^2\right] \\
&= \gamma^2 \mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} F_m(z_m^t, \xi_m^t) - F_m(z_m^t) + F_m(z_m^t) - F_m(\bar{z}^t) + F_m(\bar{z}^t)\right\|^2\right] \\
&\geq \frac{\gamma^2}{2}\mathsf{E}\left\|F(\bar{z}^t)\right\|^2 - \gamma^2 \mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} F_m(z_m^t, \xi_m^t) - F_m(z_m^t) + F_m(z_m^t) - F_m(\bar{z}^t)\right\|^2\right] \\
&\geq \frac{\gamma^2}{2}\mathsf{E}\left\|F(\bar{z}^t)\right\|^2 - 2\gamma^2 \mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} F_m(z_m^t, \xi_m^t) - F_m(z_m^t)\right\|^2\right] \\
&\quad - 2\gamma^2 \mathsf{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} F_m(z_m^t) - F_m(\bar{z}^t)\right\|^2\right] \\
&\overset{(4)}{\geq} \frac{\gamma^2}{2}\mathsf{E}\left\|F(\bar{z}^t)\right\|^2 - \frac{2\gamma^2\sigma^2}{M} - \frac{2\gamma^2 L_{\max}^2}{M}\sum_{m=1}^{M}\mathsf{E}\left[\left\|z_m^t - \bar{z}^t\right\|^2\right] \\
&= \frac{\gamma^2}{2}\mathsf{E}\left\|F(\bar{z}^t)\right\|^2 - \frac{2\gamma^2\sigma^2}{M} - 2\gamma^2 L_{\max}^2 \mathsf{E}\left[\mathrm{Err}(t)\right].
\end{aligned}
$$

Connecting with previous gives

$$
\begin{aligned}
\frac{\gamma^2}{4}\mathsf{E}\left[kF(\bar{z}^t)k^2\right] \leq\ & \mathsf{E}\left[k\bar{z}^t - z^\star k^2\right] - \mathsf{E}\left[k\bar{z}^{t+1} - z^\star k^2\right] \\
& + 2\gamma L_{\max}\sqrt{\mathsf{E}\left[k\bar{z}^t - z^\star k^2\right]}\sqrt{\mathsf{E}\left[\mathrm{Err}(t+1/2)\right]} \\
& + (\gamma L_{\max} + 5\gamma^2 L_{\max}^2)\mathsf{E}\left[\mathrm{Err}(t+1/2)\right] + 6\gamma^2 L_{\max}^2 \mathsf{E}\left[\mathrm{Err}(t)\right] + \frac{11\gamma^2\sigma^2}{M}.
\end{aligned}
$$

With result of Lemma 14 we get

$$
\begin{aligned}
\frac{\gamma^2}{4}\mathsf{E}\left[kF(\bar{z}^t)k^2\right] \leq\ & \mathsf{E}\left[k\bar{z}^t - z^\star k^2\right] - \mathsf{E}\left[k\bar{z}^{t+1} - z^\star k^2\right] \\
& + 2\gamma L_{\max}\sqrt{\mathsf{E}\left[k\bar{z}^t - z^\star k^2\right]}\sqrt{216(D^2H + \sigma^2)H\gamma^2} \\
& + \frac{11\gamma^2\sigma^2}{M} + 216(\gamma L_{\max} + 11\gamma^2 L_{\max}^2)(D^2H + \sigma^2)H\gamma^2.
\end{aligned}
$$

Summing over all t from 0 to $T - 1$ and averaging gives:

$$
\begin{aligned}
\mathsf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} kF(\bar{z}^t)k^2\right] \leq\ & \frac{4kz^0 - z^\star k^2}{\gamma^2 T} + \frac{44\sigma^2}{M} \\
& + 1000(\gamma L_{\max} + 11\gamma^2 L_{\max}^2)(D^2H + \sigma^2)H \\
& + \frac{120 L_{\max}\sqrt{(D^2H + \sigma^2)H}}{T}\sum_{t=0}^{T-1}\sqrt{\mathsf{E}\left[k\bar{z}^t - z^\star k^2\right]}. \qquad (57)
\end{aligned}
$$

Under the additional assumption that $\|z^*\| \leq \Omega$ and $\|\bar{z}^t\| \leq \Omega$, from (57), we obtain

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|F(\bar{z}^t)\|^2\right] = \mathcal{O}\Bigg(\frac{\|z^0 - z^*\|^2}{\gamma^2 T} + (\gamma L_{\max} + \gamma^2 L_{\max}^2)(D^2H + \sigma^2)H \\ + \frac{\sigma^2}{M} + L_{\max}\Omega\sqrt{(D^2H + \sigma^2)H}\Bigg).$$

With $\gamma = \min\left\{\frac{1}{4L_{\max}}; \left(\frac{\|\bar{z}^0 - z^*\|^2}{TL_{\max}(D^2H+\sigma^2)H}\right)^{1/3}\right\}$ we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|F(\bar{z}^t)\|^2\right] = \mathcal{O}\Bigg(\frac{L_{\max}^2\|z^0 - z^*\|^2}{T} + \frac{(L_{\max}\Omega(D^2H + \sigma^2)H)^{2/3}}{T^{1/3}} + \frac{\sigma^2}{M} \\ + L_{\max}\Omega\sqrt{(D^2H + \sigma^2)H}\Bigg).$$