arXiv:2106.08315v2 [math.OC] 9 Oct 2021

Decentralized Local Stochastic Extra-Gradient for Variational Inequalities*

Aleksandr Beznosikov[1,2,3], **Pavel Dvurechensky**[4,5], **Anastasia Koloskova**[6], **Valentin Samokhin**[1,5,7], **Sebastian U. Stich**[6], **Alexander Gasnikov**[1,2,5]

[1]Moscow Institute of Physics and Technology, Russia
[2]Higher School of Economics, Russia
[3]Yandex, Russia
[4]Weierstrass Institute for Applied Analysis and Stochastics, Germany
[5]Institute for Information Transmission Problems RAS, Russia
[6]EPFL, Switzerland
[7] Skolkovo Institute of Science and Technology, Russia

Abstract

We consider distributed stochastic variational inequalities (VIs) on unbounded domain with the problem data being heterogeneous (non-IID) and distributed across many devices. We make very general assumption on the computational network that, in particular, covers the settings of fully decentralized calculations with time-varying networks and centralized topologies commonly used in Federated Learning. Moreover, multiple local updates on the workers can be made for reducing the communication frequency between workers. We extend stochastic extragradient method to this very general setting and theoretically analyze its convergence rate in the strongly monotone, monotone, and non-monotone setting when an Minty solution exists. The provided rates have explicit dependence on network characteristics and how it varies with time, data heterogeneity, variance, number of devices, and other standard parameters. As a special case, our method and analysis apply to distributed stochastic saddle-point problems (SPP), e.g., to training Deep Generative Adversarial Networks (GANs) for which the decentralized training has been reported to be extremely challenging. In experiments for decentralized training of GANs we demonstrate the effectiveness of our proposed approach.

1 Introduction

In large scale machine learning (ML) scenarios the training data is often is split over many client devices (e.g. geo-distributed datacenters or mobile devices) (Kairouz et al. 2019). Decentralized training methods can train a ML model to the same accuracy as if all data would be aggregated on one single server (Lian et al. 2017; Assran et al. 2019). Training in a non-centralized fashion can offer many advantages over traditional centralized approaches in core aspects such as data ownership, privacy, fault tolerance and scalability. A particular instance of the decentralized learning setting is Federated Learning (FL), where the training is orchestrated by a single entity that communicates with all participating clients (McMahan et al. 2016; Kairouz et al. 2019). In contrast, in fully decentralized learning (FD) scenarios the devices only communicate with their neighbours in the network topology (Lian et al. 2017). Such algorithms are important in scenarios where centralized communication is expensive or impossible.

There have been tremendous advances recently in the development, design and understanding of decentralized training schemes (Nedić and Ozdaglar 2009; Wei and Ozdaglar 2012; Shi et al. 2015; Lian et al. 2017; Scaman et al. 2017; Uribe, Lee, and Gasnikov 2018; Tang et al. 2018; Wang and Joshi 2018). In particularly, aspects such as data-heterogeneity (Tang et al. 2018; Pu and Nedić 2020; Lin et al. 2021), communication efficiency (through local updates (Lan, Lee, and Zhou 2018; Koloskova et al. 2020) or compression (Tang et al. 2019; Koloskova, Stich, and Jaggi 2019a)), or personalization (Vanhaesebrouck, Bellet, and Tommasi 2017; Bellet et al. 2018) have been studied recently. However, all these methods have been developed for single objective loss functions (minimization objective) and are not applicable to more general problem classes. For example, the training of Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) requires the joint optimization of the generator and discriminator objective, i.e. solving a non-convex non-concave saddle-point problem (SPP). This problem structure makes GANs notoriously difficult to train in the single machine setting (Gidel et al. 2019; Chavdarova et al. 2019, 2021) and in particular over decentralized data (Liu et al. 2020; Mukherjee and Chakraborty 2020; Rogozin et al. 2021).

In this paper, we present a novel algorithm for solving decentralized SPP, and more generally, decentralized stochastic Minty variational inequalities (MVI) (Minty 1962; Juditsky, Nemirovski, and Tauvel 2011). In a decentralized stochastic MVI, the data is distributed over $M \geq 1$ devices and each device $m \in [M]$ has access to its local stochastic oracle $F_m(z, \xi_m)$ for the operator $F_m(z) := \mathbb{E}_{\xi_m \sim \mathcal{D}_m} F_m(z, \xi_m)$. The data ξ_m follows unknown distributions $\mathcal{D}_m$, different at every node $m \in [M]$. The devices are connected via network given as a communication graph

*The research of A. Beznosikov, P. Dvurechensky and A. Gasnikov was supported by Russian Science Foundation (project No. 21-71-30005).

and two nodes can exchange information only if they are connected by an edge in this graph. The goal is, respecting the communication constraints, to find cooperatively a point $z^* \in \mathbb{R}^n$ such that, for all $\forall z \in \mathbb{R}^n$,

$$\frac{1}{M}\sum_{m=1}^{M}\langle \mathbb{E}_{\xi_m\sim\mathcal{D}_m} F_m(z,\xi_m), z^*-z\rangle \leq 0. \quad (1)$$

A special instance of a decentralized MVI is the decentralized SPP with local objectives $f_m(x,y) := \mathbb{E}_{\xi_m\sim\mathcal{D}_m}[f_m(x,y,\xi_m)]$:

$$\min_{x\in\mathbb{R}^{n_x}}\max_{y\in\mathbb{R}^{n_y}}\left[f(x,y) := \frac{1}{M}\sum_{m=1}^{M} f_m(x,y)\right]. \quad (2)$$

This can be seen by considering $z = \begin{bmatrix} x \\ y \end{bmatrix}$ and the gradient field $F_m(z) = \begin{bmatrix} \nabla_x f_m(x,y) \\ -\nabla_y f_m(x,y) \end{bmatrix}$. In the special case when $f(x,y)$ is convex-concave, the corresponding operator $F(z) = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\xi_m} F_m(z,\xi_m)$ is monotone. However, in the context of GANs—where x and y are the parameters of the generator and discriminator, respectively—in general the local $f_m(x,y)$ are possibly non-convex, non-concave in x, y and we cannot assume monotonicity of F in general (see also (Diakonikolas, Daskalakis, and Jordan 2021)).

In this paper, we develop a novel algorithm for problems (1) or (2). Gradient descent-ascent scheme on objective (2) may diverge even in the simple convex-concave setting and $M = 1$ worker (Chavdarova et al. 2019). Thus, unlike (Liu et al. 2020), we use extragradient updates (Korpelevich 1976; Juditsky, Nemirovski, and Tauvel 2011; Gidel et al. 2019) as a building block and combine it with gossip-type communication protocol (Xiao and Boyd 2004; Boyd et al. 2006) on arbitrary, possibly time-varying, network topologies. One of the main challenges due to communication constraints is a "network error" induced by impossibility of all the nodes to reach the exact consensus, i.e. to have the same information on the current iterate of an algorithm. Thus, each device stores a local variable and only partial consensus among nodes can be achieved by gossip steps (Kong et al. 2021). Unlike other decentralized algorithms (Scaman et al. 2017; Liu et al. 2020) our method avoids multiple gossip steps per iteration that leads to better practical performance and possibility to work on time-varying networks. Moreover, our method allows multiple local updates between communication rounds making it suitable for communication- and privacy-restricted FL or fully decentralized settings (Zinkevich et al. 2010).

Our contributions. 1) We develop a novel algorithm, based on extragradient updates, for distributed stochastic MVIs with heterogeneous data and, in particular, distributed stochastic SPP. Our scheme supports very general communication protocol that covers centralized settings as in Federated Learning, fully decentralized settings, local steps in the both centralized/decentralized settings, and time-varying topology. In particular, we are not aware of earlier works proposing or analyzing extragradient method with local steps for the fully decentralized setting or decentralized algorithms for stochastic MVIs over time-varying networks.

2) Under the very general communication protocol and in the three settings of MVIs with an operator that is strongly-monotone, monotone, or non-monotone with the Minty condition, we prove convergence for our algorithm and give an explicit dependence of the rates on the problem parameters: characteristics of the network and how it changes with time, data heterogeneity, variance of the data, number of devices, and other standard parameters. These theoretical results clearly translate to the corresponding three settings of SPP (strongly-convex-strongly-concave, convex-concave, non convex-concave under Minty condition). All our theoretical results are valid in the important heterogeneous data regime and allow judging in a quantifiable way how different properties, e.g., data heterogeneity, scale of the noise in the data, and network characteristics influence the convergence rate of the algorithm. Even for decentralized setting, our results are novel for time-varying graphs and three different settings of monotonicity. See also Table 1 that gives more details on our contribution compared to existing literature. The main challenge of our analysis is to deal with very general assumption on the communication protocol and cope with the errors caused by stochastic nature and heretogeneity of the data and limited information exchange between the nodes of the communication network. As a byproduct of independent interest, we analyze stochastic extragradient method with biased oracle on unbounded domains that was not done so far in the literature.

3) We verify our theoretical results in numerical experiments and demonstrate the practical effectiveness of the proposed scheme we train DCGAN (Radford, Metz, and Chintala 2015) architecture on the CIFAR-10 (Krizhevsky, Nair, and Hinton 2009) dataset.

1.1 Related Work

The research on MVIs dates back at least to 1962 (Minty 1962) with the classic book (Kinderlehrer and Stampacchia 2000) and the recent works (Liu et al. 2019b; Lin, Jin, and Jordan 2020; Bullins and Lai 2020; Diakonikolas, Daskalakis, and Jordan 2021). VIs arise in a broad variety of settings: image denoising (Esser, Zhang, and Chan 2010; Chambolle and Pock 2011), game theory and optimal control (Facchinei and Pang 2007), robust optimization (Ben-Tal, Ghaoui, and Nemirovski 2009) or non-smooth oprimization via smooth reformulations (Nesterov 2005; Nemirovski 2004). In ML, MVIs and SPP arise in training GANs (Daskalakis et al. 2018; Chavdarova et al. 2019, 2021), reinforcement learning (Omidshafiei et al. 2017; Jin and Sidford 2020), and adversarial training (Madry et al. 2018).

Extragradient. The extragradient method (EGM) was first proposed in (Korpelevich 1976) and extended as mirror-prox method for deterministic problems in (Nemirovski 2004) and for stochastic ones with bounded variance in (Juditsky, Nemirovski, and Tauvel 2011). Yet, if the stochastic noise is not uniformly bounded, the EGM can diverge, see (Chavdarova et al. 2019; Mishchenko et al. 2019).

Decentralized algorithms for MVIs and SPP are the most closely related to our work and we summarize them in Table 1 and compare with our contribution, e.g. existing methods do not support arbitrary time-varying network typologies. The methods that use multiple rounds of gossip averaging (sparse communication) per iteration (Liu et al. 2020; Beznosikov, Samokhin, and Gasnikov 2021) can give

Reference	base method	arbitrary network	time-varying	local updates	no multiple gossip steps	SM	M	NM
Liu et al. (Liu et al. 2020)	Stoch. ES	✔	✘	✘	✘	✘	✘	✔†
Beznosikov et al. (Beznosikov, Samokhin, and Gasnikov 2021) Algorithm 2	Stoch. ES	✔	✘	✘	✘	✔	✔	✘
Liu et al. (Liu et al. 2019b)	Deter. prox	✔	✘	✘	✔	✘	✘	✔
Mukherjee and Chakraborty (Mukherjee and Chakraborty 2020)	Deter. ES	✔	✘	✘	✔	✔	✘	✘
Rogozin et al. (Rogozin et al. 2021)	Deter. ES	✔	✘	✘	✔	✘	✔	✘
Beznosikov et al. (Beznosikov, Samokhin, and Gasnikov 2021) Algorithm 3	Stoch. ES	✘	✘	✔	-§	✔	✘	✘
Deng and Mahdavi (Deng and Mahdavi 2021)	Stoch. DA	✘	✘	✔	-	✔	✘	✔‡
Hou et al. (Hou et al. 2021)	Stoch. DA	✘	✘	✔	-	✔	✘	✘
Ours	Stoch. ES	✔	✔	✔	✔	✔	✔	✔

† – homogeneous case, ‡ – non-convex-concave SPP, § – this column does not apply to centralized algorithms.

Table 1: Comparison of approaches for distributed strongly monotone (SM), monotone (M) and non-monotone (NM) VI or, respectively, strongly-convex-strongly-concave, convex-concave, non-convex-non-concave SPPs.
Definition of columns: **base method**—the non-distributed algorithm that is taken as a basis for the distributed method, typically it is either an extragradient (ES) method or descent-ascent (DA); **arbitrary network**—supporting fully decentralized vs. only centralized topology; **time-varying**—decentralized method with time-varying network topology; **local updates**—method supporting local steps between communications; **no multiple gossip steps**—at one global iteration method does not use many iterations of gossip averaging to good accuracy; **SM,M,NM**—monotonicity assumption, see Assumption 3.

optimal theoretical rates, but are often unstable in practice. Thus, it is preferred to have only one sparse communication per iteration (Liu et al. 2019b; Mukherjee and Chakraborty 2020; Rogozin et al. 2021). The second column of the table refers to standard algorithms that are extended to distributed setting in the corresponding work. In particular, (Liu et al. 2019b) require expensive proximal updates. The closest work to ours is (Beznosikov, Samokhin, and Gasnikov 2021), where decentralized EGM without local steps is analyzed in (strongly-)monotone setting. Unlike our more general algorithm with local steps and analysis, they require multiple gossip updates in each iteration which is not desired in practice. For the FL (centralized) setting, the same work studies EGM with local steps in the strongly-monotone setting, and (Deng and Mahdavi 2021; Hou et al. 2021) study descent-ascent method with local steps, and all three do not consider arbitrary time-varying graphs as we do.

2 Algorithm

In this section we present and discuss the proposed algorithm that is based on two main ideas: (i) an extragradient step, as in the classical methods for VIs (Korpelevich 1976; Nemirovski 2004), and (ii) gossip averaging (Boyd et al. 2006; Nedić and Ozdaglar 2009) widely used in decentralized optimization methods.

The algorithm can be divided into two phases. The local phase (lines 4–6) consists of a step of the stochastic extragradient method at each node using only local information. As in the non-distributed case, the nodes make first an extrapolation step—to "look into the future"—and then an update based on the operator value at the "future" point. This is followed by the communication phase (line 7), during which the devices average local iterates with their neighbors $\mathcal{N}_m^k$ in the current network graph corresponding to the iteration k. The averaging process involves the weights $w_{m,i}^k$ – elements of the matrix W^k,which is called the mixing matrix:

Definition 1 (Mixing matrix) *We call a matrix* $W \in [0;1]^{M\times M}$ *a mixing matrix if it satisfies the following conditions: 1)* W *is symmetric, 2)* W *is doubly stochastic (*$W\mathbf{1} = \mathbf{1}$*,* $\mathbf{1}^T W = \mathbf{1}^T$*, where* $\mathbf{1}$ *denotes the all-one vector), 3)* W *is aligned with the network:* $w_{ij} \neq 0$ *if and only if* $i = j$ *or edge* (i, j) *is in the network graph.*

Algorithm 1: Extra Step Time-Varying Gossip Method

parameters: stepsize $\gamma > 0$, $\{\mathcal{W}^k\}_{k\geq 0}$ – rules or distributions for selecting mixing matrix in iteration k.
initialize: $z^0 \in \mathcal{Z}, \forall m : z_m^0 = z^0$

1: **for** $k = 0, 1, 2, \dots$ **do**
2: Sample matrix W^k from $\mathcal{W}^k$
3: **for** each machine m **do**
4: Generate independently $\xi_m^k \sim \mathcal{D}_k$, $\xi_m^{k+1/3} \sim \mathcal{D}_k$
5: $z_m^{k+1/3} = z_m^k - \gamma F_m(z_m^k, \xi_m^k)$
6: $z_m^{k+2/3} = z_m^k - \gamma F_m(z_m^{k+1/3}, \xi_m^{k+1/3})$
7: $z_m^{k+1} = \sum_{i\in\mathcal{N}_m^k} w_{m,i}^k z_i^{k+2/3}$
8: **end for**
9: **end for**

Typical choices of mixing matrices are for example (i) the choice $W^k = I_M - \frac{L^k}{\lambda_{\max}(L^k)}$, where L^k denotes the Laplacian matrix of the network graph at time k and I_M the identity matrix, or (ii) local rules based on the degrees of the neighboring nodes (Xiao and Boyd 2004). Note that our setting allows a great flexibility as in between the iterations the topology of the communication graph can change, and the matrix W^k, that encodes the current structure of the network, changes accordingly. This is encoded in line 2, where the matrix W^k is generated by some rule $\mathcal{W}^k$ which can have different nature. Examples include deterministic choice of a sequence of matrices W^k, sampling from a time-varying probability distribution on matrices. Even local steps with no communication can be encoded with a diagonal matrix W^k. To ensure that it is possible to approach the consensus between the agents, we need the following assumption on the mixing properties of the matrix sequence W^k.

Assumption 1 (Expected Consensus Rate) *We assume that there exist two constants* $p \in (0, 1]$ *and integer* $\tau \geq 1$ *such that for all matrices* $Z \in \mathbb{R}^{d\times M}$ *and all integers* $l \in \{0, \dots, T/\tau\}$*,*

$$\mathbb{E}_W[\|ZW_{l,\tau} - \bar{Z}\|_F^2] \leq (1-p)\|Z - \bar{Z}\|_F^2, \qquad (3)$$

where $W_{l,\tau} = W^{l\tau} \cdot \ldots \cdot W^{(l+1)\tau-1}$*, we use matrix notation*

$Z = [z_1, \ldots, z_M]$, $\bar{Z} = [\bar{z}, \ldots, \bar{z}]$ *with* $\bar{z} = \frac{1}{M}\sum_{m=1}^{M} z_m$, *and the expectation* $\mathbb{E}_W$ *is taken over distributions of* W^t *and indices* $t \in \{l\tau, ..., (l+1)\tau - 1\}$.

This assumption ensures that after τ steps of the gossip algorithm with such time-varying matrices we improve the averaging between nodes by the factor of $\frac{1}{1-p}$. It is important that in this case some matrices W^k can be, for example, the identity matrix (which corresponds to performing local steps only in this round).

Such an assumption about time-varying networks (more precisely, about their mixing matrices) first appeared in (Koloskova et al. 2020) in a different to ours setting of optimization problems. As the authors note, Assumption 1 is tighter than many other already classical assumptions about time-varying graphs and covers many special cases of decentralized and centralized algorithms. For example, if we fix $W^k = W$ for some connected graph we get a decentralized algorithm on a constant topology. If at the same time we set the matrix $W = \frac{1}{M}\mathbf{1}\mathbf{1}^T$, then it is easy to make sure that then we get an analogue of centralized learning with averaging over all nodes in one communication step. If we take the matrix $W^k = W$ for some connected graph at every τth iteration, and in other cases use $W^k = I_M$, we have a decentralized (or centralized) algorithm with local iterations and communication once in τ iterations. Generic Assumption 1 covers also many other settings of time-varying decentralized topologies, e.g. random topologies, cliques, B-connected graphs (Jadbabaie, Lin, and Morse 2003; Nedic et al. 2009). Below we provably show that under an appropriate choice of the stepsize our extragradient method provably converges under such a general assumption that covers centralized, decentralized settings, local steps in the both centralized/decentralized settings, and changing topology. Even for decentralized setting, this is novel for time-varying graphs and three different settings of monotonicity.

3 Setting and assumptions

In this section we introduce necessary assumptions that are used to analyze the proposed algorithm.

Assumption 2 (Lipschitzness) *For all m, the operator $F_m(z)$ is Lipschitz with constant L, i.e. for all z_1, z_2*

$$\|F_m(z_1) - F_m(z_2)\| \le L\|z_1 - z_2\|. \tag{4}$$

This is a standard assumption that is used in the analysis of all the methods displayed in Table 1.

Assumption 3 *We consider three scenarios for the operator F, namely when F is strongly monotone, monotone and non-monotone, but with an additional assumption:*

(SM) Strong monotonicity. *There exists $\mu > 0$ such that*

$$\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \ge \mu\|z_1 - z_2\|^2, \forall z_1, z_2. \tag{SM}$$

(M) Monotonicity. *For all z_1, z_2, it holds that:*

$$\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \ge 0. \tag{M}$$

(NM) Non-monotonicity. *There exists z^* such that for all z:*

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

Assumptions (SM), (M), 2 are standard classical assumptions in the literature on VIs. Assumption (NM) is quite weak and sometimes referred to as the *Minty condition*. It is quite standard for the analysis of algorithms for non-monotone VIs in different settings (Dang and Lan 2015; Diakonikolas, Daskalakis, and Jordan 2021), including distributed (Liu et al. 2020, 2019b).

The next assumption is standard for stochastic setting.

Assumption 4 (Bounded noise) *$F_m(z, \xi)$ is unbiased and has bounded variance, i.e. for all $z \in \mathcal{Z}$*

$$\mathbb{E}[F_m(z,\xi)] = F_m(z),\ \mathbb{E}[\|F_m(z,\xi) - F_m(z)\|^2] \le \sigma^2. \tag{5}$$

The last assumption reflects the variability of the local operators compared to their mean and can often be found in the literature on local and decentralized methods, where it is called D-heterogeneity.

Assumption 5 (D-heterogeneity.) *The values of the local operator have bounded variablility, i.e. for all z*

$$\|F_m(z) - F(z)\|^2 \le D^2. \tag{6}$$

4 Main results

In this section, we present the convergence results for the proposed method under different settings of Assumption 3. To present the main result, we introduce notation $\bar{z}^k := \frac{1}{M}\sum_{m=1}^{M} z_m^k$, $\bar{z}^{k+1/3} := \frac{1}{M}\sum_{m=1}^{M} z_m^{k+1/3}$ for the averaged among the devices iterates and $\hat{z}^k = \frac{1}{k+1}\sum_{i=0}^{k} \bar{z}^{i+1/3}$ for the averaged among the devices and iterates sequence, a.k.a. ergodic sequence. Finally, we denote $\Delta = \frac{\tau}{p}\left(\frac{D^2\tau}{p} + \sigma^2\right)$ which plays a role of the consensus error—the error that is due to impossibility to reach the exact consensus between the agents. Note that the data heterogeneity appears in the convergence rates only through the quantity Δ.

Theorem 1 (Main theorem) *Let Assumptions 1, 2, 4, 5 hold and the sequences $\bar{z}^k$, $\hat{z}^k$ be generated by Algorithm 1 that is run for $K > 0$ iterations. Then, with an appropriate choice of a constant step γ depending on the problem parameters listed in the assumptions and the iteration budget K (see the details in the supplementary material), the following convergence estimates are valid.*

- ***Strongly-monotone case:*** *under Assumption 3(SM) with $\gamma \le \frac{p}{120L\tau}$ it holds that $\mathbb{E}\left[\|\bar{z}^{K+1} - z^*\|^2\right]$ is*

$$\tilde{\mathcal{O}}\left(\|z^0 - z^*\|^2 \cdot \exp\left(-\frac{\mu K p}{240L\tau}\right) + \frac{\sigma^2}{\mu^2 MK} + \frac{L^2\Delta}{\mu^4K^2}\right); \tag{7}$$

- ***Monotone case:*** *under Assumption 3(M) with $\gamma \le \frac{1}{3L}$, for any convex compact $\mathcal{C}$ s.t. $z^0, z^* \in \mathcal{C}$ and $\max_{z,z'\in\mathcal{C}}\|z - z'\| \le \Omega_{\mathcal{C}}$ it holds that $\sup_{z\in\mathcal{C}}\mathbb{E}\left[\langle F(z), \hat{z}^K - z\rangle\right]$ is*

$$\mathcal{O}\Big(\frac{L\Omega_{\mathcal{C}}^2}{K} + \frac{\sigma\Omega_{\mathcal{C}}}{\sqrt{MK}} + \frac{\sqrt{L\Omega_{\mathcal{C}}^3\sqrt{\Delta}}}{\sqrt{K}} + \sqrt{\frac{(\Delta + L^2\Omega_{\mathcal{C}}^2)\Omega_{\mathcal{C}}\sqrt{\Delta}}{KL}}\Big) \tag{8}$$

Under the additional assumption that, for all k, $\|\bar{z}^k\| \le \Omega$, we have that $\sup_{z\in\mathcal{C}}\mathbb{E}\left[\langle F(z), \hat{z}^K - z\rangle\right]$ is

$$\mathcal{O}\Big(\frac{L\Omega_{\mathcal{C}}^2}{K} + \frac{\sigma\Omega_{\mathcal{C}}}{\sqrt{MK}} + \frac{\sqrt{L\Omega_{\mathcal{C}}^3\sqrt{\Delta}}}{K^{3/4}} + \sqrt{\frac{((\Omega+\Omega_{\mathcal{C}})L\sqrt{\Delta}+\Delta)\Omega_{\mathcal{C}}^2}{K}}\Big); \tag{9}$$

• ***Non-monotone case:*** *under Assumption 3(NM) with* $\gamma \leq \frac{1}{4L}$ *it holds that* $\mathbb{E}\Big[\frac{1}{K+1}\sum_{k=0}^{K}\|F(\bar{z}^k)\|^2\Big]$ *is*

$$\mathcal{O}\Big(\frac{L^2\|z^0-z^*\|^2}{K} + \frac{\sigma^2}{M} + L\|z^0-z^*\|\sqrt{\Delta} + \frac{\sqrt{L\|z^0-z^*\|\Delta^{3/4}}}{\sqrt{K}}\Big). \tag{10}$$

Under an additional assumption that $\|z^*\| \leq \Omega$ *and, for all* k, $\|\bar{z}^k\| \leq \Omega$, *we have that* $\mathbb{E}\Big[\frac{1}{K+1}\sum_{k=0}^{K}\|F(\bar{z}^k)\|^2\Big]$ *is*

$$\mathcal{O}\Big(\frac{L^2\Omega^2}{K} + \frac{\sigma^2}{M} + \frac{(L\Omega\Delta)^{2/3}}{K^{1/3}} + L\Omega\sqrt{\Delta}\Big). \tag{11}$$

The proof of the theorem is given in the supplementary material. We underline that the standard analysis (Juditsky, Nemirovski, and Tauvel 2011) does not apply for the following reasons. Firstly, unlike (Juditsky, Nemirovski, and Tauvel 2011) in our problem (1) the feasible set is not bounded, which is especially important for the analysis in the monotone and non-monotone settings. Secondly, our algorithm has an additional communication step (Step 7) between the computational nodes, which leads to impossibility for all the nodes to have the same information about the global operator $F(z)$ and about the current iterate z. This, in order, leads to biased oracle that, unlike existing works, has to be analyzed in the setting of unbounded feasible set, which is quite challenging. To analyze our variant of the extragradient method we successfully handle this challenge. Our key steps are to bound the bias (see, e.g. the last two terms in the r.h.s. of Lemma 7 that are caused by the network errors), prove the boundedness in expectation of the sequence of the iterates for monotone (see Section C.3 of the supplementary material) and non-monotone (see Section C.4 of the supplementary material) cases, which may be of independent interest and which we haven't seen in the literature, even in the non-distributed setting with biased stochastic oracles. Proving the boundedness is challenging due to noise caused by stochasticity and heterogeneity of the data and network effects due to the imperfect exchange of information. Surprisingly, in the end we still manage to analyze our algorithm under the very general Assumption 1 and we are not aware of any results with similar generality of the settings: different networks topologies (including time-varying), distributed architectures, different monotonicity assumptions.

The provided convergence rates have explicit dependence on the problem parameters: network that is characterized by mixing time τ and mixing factor p, data heterogeneity D (that appears in the convergence rates only through the quantity Δ), variance σ^2 of the noise in the data, Lipschitz constant L and strong monotonicity parameter μ, number of devices M. Thus, our rates allow judging how different properties, e.g., data heterogeneity, amount of noise, network characteristics influence the convergence of the algorithm. This opens an opportunity for meta-optimization process if we can design the network by changing M, τ, p to achieve faster convergence.

We now discuss the convergence results obtained in the theorem, and also compare them with the already existing algorithms (see Table 1) and their guarantees. Firstly, all the estimates have similar there-component structure. The first term corresponds to the deterministic setting and is similar to existing methods for smooth VIs in non-distributed setting. Only in the strongly convex case there is an additional factor τ/p that increases the condition number L/μ of the problem. The second (stochastic) term is also standard for non-distributed setting and corresponds to the stochastic nature of the problem. Note that, for a very general distributed setting we have managed to obtain the corresponding terms similar to non-distributed setting. Moreover, we can see the benefit of exploiting distributed computations: the leading stochastic term depends on σ^2/M that decreases as the number M of nodes increases. The other terms correspond to the consensus error Δ and are caused by imperfect communications between the agents, i.e. that it is impossible for the agents to have exactly the same information about the current iterate. Importantly, in all the cases this error does not make the overall convergence worse since the dependence on K is no worse for these terms than the dependence on K in the stochastic term. In the experimental section we illustrate that the network error is not an artefact of the analysis, but indeed is present in practice.

Before we move to the specific comments for each setting, we remark that the K-dependent terms can be made arbitrarily small by increasing the total budget of iterations K and choosing the corresponding step γ. Further, the any-time convergence can be achieved by a restart technique when once in a while we increase the budget K and restart the algorithm with the new fixed stepsize. It is also possible, based on the estimates in the above theorem for strongly monotone and monotone cases, to achieve any desired accuracy by choosing an appropriate K and the corresponding step γ.

• **Strongly-monotone case:** In the centralized setting with local updates our rate is slightly better than in (Beznosikov, Samokhin, and Gasnikov 2021). Unlike our algorithm, centralized algorithms with local steps for SPP in (Deng and Mahdavi 2021; Hou et al. 2021) are based on gradient descent-ascent that may diverge in the stochastic setting even for bilinear problems. Moreover, their analysis implies a very small stepsize $\gamma \sim \frac{\mu p}{L^2\tau}$ (cf. ours $\gamma \sim \frac{p}{L\tau}$) that greatly slows down the convergence of the algorithm.

For the decentralized setting (Beznosikov, Samokhin, and Gasnikov 2021) propose an optimal algorithm with the rate matching the lower bound which they give. Our rate is worse probably because of the generality of the Assumption 1. On the other hand, we do not rely on many gossip iterations at once that should be avoided in practice. Also, our algorithm is more general, allowing us to work with time-varying topology and local steps even in the decentralized setting.

• **Monotone case:** The quantity $\sup_{z\in\mathcal{C}} \mathbb{E}\left[\left\langle F(z), \widehat{z}^K - z\right\rangle\right]$ in the estimates reflects the stochastic nature of the problem and is a counterpart of the standard restricted gap (or merit) function (Nesterov 2007): $\mathrm{Gap}_{\mathcal{C}}(u) := \sup_{z\in\mathcal{C}}\left[\langle F(z), u - z\rangle\right]$. When F is a monotone operator, if $\mathrm{Gap}_{\mathcal{C}}(\hat{u}) = 0$ and $\mathcal{C}$ contains a neighborhood of $\hat{u}$, then (Nesterov 2007; Antonakopoulos, Belmega, and Mertikopoulos 2019) $\hat{u}$ is a solution to (1) and even more: it is a strong solution to the corresponding variational inequality, i.e., for all z, $\langle F(\hat{u}), \hat{u}-z \rangle \leq 0$. Thus, $\text{Gap}_{\mathcal{C}}(u)$ is an appropriate measure of suboptimality in this setting and (8) guarantees that after a sufficient number of iterations, we obtain an approximate solution in expectation. Importantly, for (8), neither z nor $\bar{z}^k$ are assumed to be bounded. As in the previous works on non-distributed algorithms for MVIs (Nesterov 2007; Antonakopoulos, Belmega, and Mertikopoulos 2019), we use $\text{Gap}_{\mathcal{C}}(u)$ for an arbitrary compact set $\mathcal{C}$ that contains z^0 and z^* (this can be a large set). Further, (9) is a refined version of the general result (8) under additional assumptions of boundedness. If boundedness does not hold, we still have (8). Moreover, (8) and (9) hold for the same method and to run the algorithm, there is no need to know in advance whether the generated sequence is bounded or not.

Only (Beznosikov, Samokhin, and Gasnikov 2021; Rogozin et al. 2021) consider MVIs with monotone operator in distributed setting. Our algorithm is more general than theirs: our algorithm supports time-varying networks and local steps in between communications. The algorithm in (Beznosikov, Samokhin, and Gasnikov 2021) uses multiple gossip steps between the updates of the iterates. On the one hand, this allows decrease the consensus error Δ. On the other hand, this leads to an additional factor in the number of communications compared to our estimates: the first term in their bound is $\sqrt{\chi}$ times larger than ours, where $\chi > 1$ is some condition number of the mixing matrix. Moreover, multiple gossip steps may be impractical if the communication is performed through unstable channels or is expensive by some reason. The paper (Rogozin et al. 2021) considers only deterministic setting.

• **Non-monotone case:** Here the same as in the previous case remark on the boundedness of $\bar{z}^k$, z^* assumed to obtain (11) can be repeated. Further, in this setting the convergence is guaranteed up to some accuracy that is governed by the stochastic nature of the problem (the σ^2 term) and by the distributed nature of the problem (Δ terms). With this respect the results are similar to non-distributed stochastic extragradient method (Barazandeh, Tarzanagh, and Michailidis 2021) and distributed method in the homogeneous case (Liu et al. 2020). To the best of our knowledge, convergence up to arbitrarily small accuracy can be guaranteed only for deterministic distributed methods (Liu et al. 2019b), i.e. in a much simpler setting than ours. Moreover, the methods of (Liu et al. 2019b) are not the most robust since they require evaluating the proximal operator of a function and it is assumed that this can be done in closed-form, which is computationally expensive and may not hold in practice.

Note that based on our result it is possible to achieve convergence up to arbitrarily small accuracy if one considers the homogeneous case ($D = 0$). Indeed, choosing the right batch size, for example, $\sim K^\alpha$ with $\alpha > 0$, one can replace σ^2 by $\frac{\sigma^2}{K^\alpha}$ in (10) and (11) and get convergence.

5 Experiments

We present two set of experiments to validate the performance of Algorithm 1. In Section 5.1 we verify the proven convergence guarantee on a strong-monotone and on a monotone bilenar problem, and in Section 5.2 we explore the non-monotone case with application of GAN training. Extended details about the exerimental setup can be found in the appendix.

5.1 Verifying Theoretical Convergence Rate

First, we focus on verifying whether the Algorithm 1 behaviour is predicted by the theoretical convergence rate (Theorem 1).

Setup. We consider a distributed bi-linear SPP (2) with objective functions $f_m(x, y) = \frac{a}{2}\|x\|^2 + \frac{b}{2}x^\top y - \frac{a}{2}\|y\|^2 + c_m^\top x$, where $x, y, c_m \in \mathbb{R}^n$, $a, b \in \mathbb{R}$ and $m \in \{1, \ldots, M\}$. This set of functions satisfy Assumptions 2, 3, 5 with constants $\mu = a, L = a^2 + \frac{b^2}{4}, D = \max_m \|c_m - \bar{c}\|$. In this section we use a ring topology on $M = 9$ nodes, with uniform averaging weights, and we set the dimension $n = 2$, $a = b = 1$ and we set $D = 1$ and keep $\tau = 1$. The value of the parameter p in this setting is approximately 0.288 (Koloskova, Stich, and Jaggi 2019b, Table 1). To satisfy Assumption 4, we generate stochastic gradients by adding to the real gradients unbiased Gaussian noise with variance σ^2.

Convergence Behaviour. In Figure 1 we show the convergence of Algorithm 1 with a fixed stepsize on a strongly-monotone ($a = 1$) and monotone ($a = 0$) instance. In the strongy-convex case we see linear convergence up to the level of the heterogenity parameter and the noise. The convergence on the non-strongly monotone problem is stronger affected by the noise, but interestingly we also see linear convergence (with oscillations) when there is no noise. Note that convergence to some limit accuracy is expected since when a constant stepsize is used in stochastic optimization/stochastic variational inequalities with strong convexity/monotonicity, the algorithm is usually guaranteed to converge to a vicinity of the solution, see, e.g., Theorem 2 in (Mishchenko et al. 2019). This is also in accordance to Theorem 1 which for a fixed stepsize gurantees convergence to some non-zero limit accuracy and says that the error to drop to zero one needs to choose a decreasing stepsize. We additionally validate in the Appendix A.2 that with decreasing stepsize, algorithm can converge to zero error.

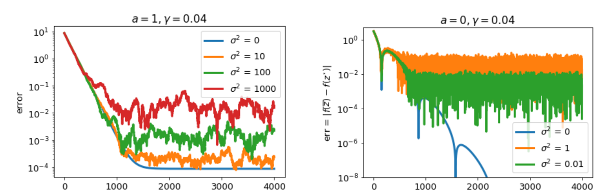


Figure 1: Impact of the stochastic noise in strongly monotone (left) and monotone (right) cases.

Depencence on the Heterogenity parameter D. In a second set of experiments we aim to verify the dependence on the data heterogeneity parameter D. Therefore, we consider the setting when $\sigma^2 = 0$. From our theory, equation (7), we predict that the most significant term in the convergence rate (when $\sigma^2 = 0$) scales as $\mathcal{O}\left(\frac{D^2}{p^2 K^2}\right)$ (since the primary

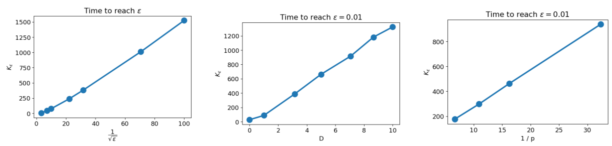


Figure 2: Verifying the $\mathcal{O}\left(\frac{D^2}{p^2K^2}\right)$ convergence for the strongly monotone noiseless ($\hat{\sigma}^2 = 0$) case.

goal of this experiment is to study the dependence on p, D, K, we omit all the other fixed parameters for simplicity). We repeat experiments for different $a > 0$ (strongly-monotone case) with number of iterations needed until the error $\frac{1}{M}\sum_{m=1}^{M}\|z_m^k - z^*\|^2 < \epsilon$, for different ϵ. In all these experiments the step size is tuned individually.

First, we verify the power of K. For this experiment, we keep D, L, μ, p constant and vary the accuracy ϵ. We can seen from the leftmost subplot in Figure 2 that the number of iterations scale as $K \propto \frac{1}{\sqrt{\epsilon}}$, confirming the predicted $T = \mathcal{O}\left(\frac{1}{K^2}\right)$ dependency. Next, we measure the number of iteration it takes to reach error $\epsilon = 0.01$ while varying D. The middle plot shows that the number of iterations scales proportional to D (showing $D \propto K$). Lastly, we depict the time to reach $\epsilon = 0.01$ while changing the graph parameter p and again observe $\frac{1}{p} \propto K$. All together, these experiments verify the $\mathcal{O}\left(\frac{D^2}{p^2K^2}\right)$ term in the convergence rate.

5.2 GANs

Our algorithm allows you to combine different communication topologies of devices, as well as local steps in distributed learning. This is what we want to compare in the next experiment with GAN (in Appendix A.1, we discuss how close the theory is to GANs.).

Data and model. We consider the CIFAR-10 (Krizhevsky, Nair, and Hinton 2009) dataset. It contains 60000 images (but we increased the size of data by 4 times due to transformations and adding noise), equally distributed over 10 classes. We simulate a distributed setup of 16 nodes on two GPUs, we are using Ray (Moritz et al. 2018b). To emulate the heterogeneous, we partition the dataset into 16 subsets. For each subset, we select a major class that forms 20% of the data, while the rest of the data split is filled uniformly by the other classes. As a basic architecture we choose DCGAN (Radford, Metz, and Chintala 2015), conditioned by class labels, similarly to (Mirza and Osindero 2014) (the network architecture can be found in Appendix A.1). We chose Adam (Kingma and Ba 2014) as the optimizer. We make one local Adam step, and then one gossip averaging step with time-varying matrix W—similar to how it works in Algorithm 1.

Setting. We compare the following three topologies:

- **Full.** Full graph at the end of each epoch, otherwise local steps. This means that we make 120 communication rounds (by communication round we mean the exchange of information between a pair of devices) in an epoch.
- **Local.** Full graph at the end of each 5th epoch, otherwise local steps. This means that we make 24 communication rounds in an epoch (in average: 4 epochs without communications and 1 epoch with 120 rounds).
- **Clusters.** At the end of each epoch, clique clusters of size 4 are randomly formed (in total 4 cliques). This means that we make 24 rounds of communication in an epoch.

It turns out that the communication budget of the first approach is higher is 5 times higher.

We use the same learning rate for the generator and discriminator equal to 0.002. The rest of the parameters and features of the architecture are contained in the Appendix.

Results. The results of the experiment can be found on Figure 3 and Figure 4. Note that all methods from the point of view of local epochs worked approximately the same and produced similar pictures. But from the point of view of communications, Local and Cluster topologies are much better. In turn, it can also be noted that the Cluster topology is slightly ahead of Local.

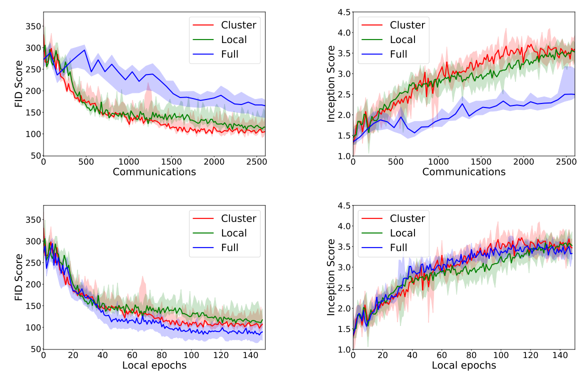
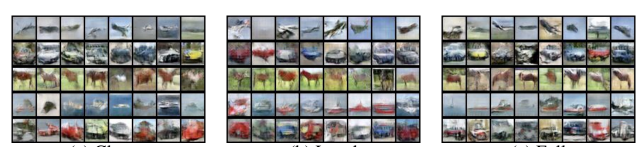


Figure 3: Comparison of three topological approaches in DCGAN distributed decentralized learning on CIFAR-10. FID Score and Inception Score in terms of the number of communications (top row), and Scores in terms of local epochs (bottom row). The experiment was repeated 5 times on different data random splitting, the maximum and minimum deviations are depicted in the plots by shade.



(a) Cluster (b) Local (c) Full

Figure 4: Pictures generated by DCGAN trained distributed on different communication topologies: (a) Cluster, (b) Local, (c) Full.

6 Conclusion

We developed a novel algorithm to efficiently solve decentalized MVIs and SPPs. Our method is the first extragradient method with local steps for time-varying network topologies. We give convergence analyses for the SM, M and NM cases. In numerical experiments we verified that that the dependency of our result on the data parameter D is tight in the SM case, and cannot be further improved in general. By training DCGAN on a decentralized topology we demonstrate that our method is effective on practical DL tasks.

References

Antonakopoulos, K.; Belmega, V.; and Mertikopoulos, P. 2019. An adaptive Mirror-Prox method for variational inequalities with singular operators. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 8455–8465. Curran Associates, Inc.

Assran, M.; Loizou, N.; Ballas, N.; and Rabbat, M. 2019. Stochastic Gradient Push for Distributed Deep Learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*. PMLR.

Barazandeh, B.; Tarzanagh, D. A.; and Michailidis, G. 2021. Solving a Class of Non-Convex Min-Max Games Using Adaptive Momentum Methods. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3625–3629.

Bellet, A.; Guerraoui, R.; Taziki, M.; and Tommasi, M. 2018. Personalized and Private Peer-to-Peer Machine Learning. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84, 473–481. PMLR.

Ben-Tal, A.; Ghaoui, L. E.; and Nemirovski, A. 2009. *Robust Optimization*. Princeton University Press.

Beznosikov, A.; Samokhin, V.; and Gasnikov, A. 2021. Distributed Saddle-Point Problems: Lower Bounds, Optimal Algorithms and Federated GANs. *arXiv preprint arXiv:2010.13112*.

Boyd, S.; Ghosh, A.; Prabhakar, B.; and Shah, D. 2006. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6): 2508–2530.

Bullins, B.; and Lai, K. A. 2020. Higher-order methods for convex-concave min-max optimization and monotone variational inequalities. *arXiv preprint arXiv:2007.04528*.

Chambolle, A.; and Pock, T. 2011. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1): 120–145.

Chavdarova, T.; Gidel, G.; Fleuret, F.; and Lacoste-Julien, S. 2019. Reducing Noise in GAN Training with Variance Reduced Extragradient. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*.

Chavdarova, T.; Pagliardini, M.; Stich, S. U.; Fleuret, F.; and Jaggi, M. 2021. Taming GANs with Lookahead-Minmax. In *International Conference on Learning Representations (ICLR)*.

Dang, C. D.; and Lan, G. 2015. On the convergence properties of non-Euclidean extragradient methods for variational inequalities with generalized monotone operators. *Computational Optimization and Applications*, 60(2): 277–310.

Daskalakis, C.; Ilyas, A.; Syrgkanis, V.; and Zeng, H. 2018. Training GANs with Optimism. In *International Conference on Learning Representations (ICLR)*.

Deng, Y.; and Mahdavi, M. 2021. Local Stochastic Gradient Descent Ascent: Convergence Analysis and Communication Efficiency. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 1387–1395. PMLR.

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

Esser, E.; Zhang, X.; and Chan, T. F. 2010. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4): 1015–1046.

Facchinei, F.; and Pang, J. 2007. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research and Financial Engineering. Springer New York. ISBN 9780387218151.

Gidel, G.; Berard, H.; Vincent, P.; and Lacoste-Julien, S. 2019. A Variational Inequality Perspective on Generative Adversarial Nets. In *International Conference on Learning Representations (ICLR)*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NIPS)*.

Gorbunov, E.; Dvurechensky, P.; and Gasnikov, A. 2018. An Accelerated Method for Derivative-Free Smooth Stochastic Convex Optimization. *arXiv preprint arXiv:1802.09022*.

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

Jadbabaie, A.; Lin, J.; and Morse, A. 2003. Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, 48(6): 988–1001.

Jin, Y.; and Sidford, A. 2020. Efficiently Solving MDPs with Stochastic Mirror Descent. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, 4890–4900. PMLR.

Juditsky, A.; Nemirovski, A.; and Tauvel, C. 2011. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1): 17–58.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; D'Oliveira, R. G. L.; Rouayheb, S. E.; Evans, D.; Gardner, J.; Garrett, Z.; Gascón, A.; Ghazi, B.; Gibbons, P. B.; Gruteser, M.; Harchaoui, Z.; He, C.; He, L.; Huo, Z.; Hutchinson, B.; Hsu, J.; Jaggi, M.; Javidi, T.; Joshi, G.; Khodak, M.; Konečný, J.; Korolova, A.; Koushanfar, F.; Koyejo, S.; Lepoint, T.; Liu, Y.; Mittal, P.; Mohri, M.; Nock, R.; Özgür, A.; Pagh, R.; Raykova, M.; Qi, H.; Ramage, D.; Raskar, R.; Song, D.; Song, W.; Stich, S. U.; Sun, Z.; Suresh, A. T.; Tramèr, F.; Vepakomma, P.; Wang, J.; Xiong, L.; Xu, Z.; Yang, Q.; Yu, F. X.; Yu, H.; and Zhao, S. 2019. Advances and Open Problems in Federated Learning. *arXiv preprint arXiv:1912.04977*.

Kinderlehrer, D.; and Stampacchia, G. 2000. *An Introduction to Variational Inequalities and Their Applications*. Society for Industrial and Applied Mathematics.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kleinberg, B.; Li, Y.; and Yuan, Y. 2018. An alternative view: When does SGD escape local minima? In *International Conference on Machine Learning*, 2698–2707. PMLR.

Koloskova, A.; Loizou, N.; Boreiri, S.; Jaggi, M.; and Stich, S. U. 2020. A Unified Theory of Decentralized SGD with Changing Topology and Local Updates. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*. PMLR.

Koloskova, A.; Stich, S.; and Jaggi, M. 2019a. Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Communication. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, 3478–3487. PMLR.

Koloskova, A.; Stich, S.; and Jaggi, M. 2019b. Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Communication. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 3478–3487. PMLR.

Kong, L.; Lin, T.; Koloskova, A.; Jaggi, M.; and Stich, S. U. 2021. Consensus Control for Decentralized Deep Learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. PMLR.

Korpelevich, G. 1976. The extragradient method for finding saddle points and other problems. *Eknomika i Matematicheskie Metody*, 12: 747–756.

Krizhevsky, A.; Nair, V.; and Hinton, G. 2009. CIFAR-10 (Canadian Institute for Advanced Research). http://www.cs.toronto.edu/~kriz/cifar.html.

Lan, G.; Lee, S.; and Zhou, Y. 2018. Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming*.

Li, Y.; and Yuan, Y. 2017. Convergence analysis of two-layer neural networks with relu activation. *arXiv preprint arXiv:1705.09886*.

Lian, X.; Zhang, C.; Zhang, H.; Hsieh, C.-J.; Zhang, W.; and Liu, J. 2017. Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 5330–5340. Curran Associates, Inc.

Lin, T.; Jin, C.; and Jordan, M. I. 2020. Near-Optimal Algorithms for Minimax Optimization. In *Proceedings of Thirty Third Conference on Learning Theory (COLT)*, volume 125, 2738–2779. PMLR.

Lin, T.; Karimireddy, S. P.; Stich, S. U.; and Jaggi, M. 2021. Quasi-Global Momentum: Accelerating Decentralized Deep Learning on Heterogeneous Data. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Liu, M.; Mroueh, Y.; Ross, J.; Zhang, W.; Cui, X.; Das, P.; and Yang, T. 2019a. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. *arXiv preprint arXiv:1912.11940*.

Liu, M.; Zhang, W.; Mroueh, Y.; Cui, X.; Ross, J.; Yang, T.; and Das, P. 2020. A Decentralized Parallel Algorithm for Training Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Liu, W.; Mokhtari, A.; Ozdaglar, A.; Pattathil, S.; Shen, Z.; and Zheng, N. 2019b. A decentralized proximal point-type method for saddle point problems. *arXiv preprint arXiv:1910.14380*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations (ICLR)*.

McMahan, H. B.; Moore, E.; Ramage, D.; and y Arcas, B. A. 2016. Federated Learning of Deep Networks using Model Averaging. *arXiv preprint arXiv:1602.05629*.

Mertikopoulos, P.; Lecouat, B.; Zenati, H.; Foo, C.-S.; Chandrasekhar, V.; and Piliouras, G. 2018. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*.

Minty, G. J. 1962. Monotone (nonlinear) operators in Hilbert space. *Duke Mathematical Journal*, 29(3): 341 – 346.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Mishchenko, K.; Kovalev, D.; Shulgin, E.; Richtárik, P.; and Malitsky, Y. 2019. Revisiting Stochastic Extragradient. *arXiv preprint arXiv:1905.11373*.

Moritz, P.; Nishihara, R.; Wang, S.; Tumanov, A.; Liaw, R.; Liang, E.; Elibol, M.; Yang, Z.; Paul, W.; Jordan, M. I.; and Stoica, I. 2018a. Ray: A Distributed Framework for Emerging AI Applications. arXiv:1712.05889.

Moritz, P.; Nishihara, R.; Wang, S.; Tumanov, A.; Liaw, R.; Liang, E.; Elibol, M.; Yang, Z.; Paul, W.; Jordan, M. I.; et al. 2018b. Ray: A distributed framework for emerging {AI} applications. In *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*, 561–577.

Mukherjee, S.; and Chakraborty, M. 2020. A decentralized algorithm for large scale min-max problems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, 2967–2972.

Nedic, A.; Olshevsky, A.; Ozdaglar, A.; and Tsitsiklis, J. N. 2009. On Distributed Averaging Algorithms and Quantization Effects. *IEEE Transactions on Automatic Control*, 54(11): 2506–2517.

Nedić, A.; and Ozdaglar, A. 2009. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1): 48–61.

Nemirovski, A. 2004. 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*, 15(1): 229–251.

Nesterov, Y. 2005. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1): 127–152.

Nesterov, Y. 2007. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2): 319–344.

Omidshafiei, S.; Pazis, J.; Amato, C.; How, J. P.; and Vian, J. 2017. Deep Decentralized Multi-task Multi-Agent Reinforcement Learning under Partial Observability. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, 2681–2690. PMLR.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32, 8026–8037. Curran Associates, Inc.

Pu, S.; and Nedić, A. 2020. Distributed stochastic gradient tracking methods. *Math. Program.*

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Rogozin, A.; Beznosikov, A.; Dvinskikh, D.; Kovalev, D.; Dvurechensky, P.; and Gasnikov, A. 2021. Decentralized Distributed Optimization for Saddle Point Problems. *arXiv preprint arXiv:2102.07758*.

Scaman, K.; Bach, F.; Bubeck, S.; Lee, Y. T.; and Massoulié, L. 2017. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Shi, W.; Ling, Q.; Wu, G.; and Yin, W. 2015. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2): 944–966.

Tang, H.; Lian, X.; Qiu, S.; Yuan, L.; Zhang, C.; Zhang, T.; and Liu, J. 2019. DeepSqueeze: Decentralization Meets Error-Compensated Compression. *arXiv preprint arXiv:1907.07346*.

Tang, H.; Lian, X.; Yan, M.; Zhang, C.; and Liu, J. 2018. D^2: Decentralized Training over Decentralized Data. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, 4848–4856. PMLR.

Uribe, C. A.; Lee, S.; and Gasnikov, A. 2018. A Dual Approach for Optimal Algorithms in Distributed Optimization over Networks. *arXiv preprint arXiv:1809.00710*.

Vanhaesebrouck, P.; Bellet, A.; and Tommasi, M. 2017. Decentralized Collaborative Learning of Personalized Models over Networks. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, 509–517. PMLR.

Wang, J.; and Joshi, G. 2018. Cooperative SGD: A unified Framework for the Design and Analysis of Communication-Efficient SGD Algorithms. *arXiv preprint arXiv:1808.07576*.

Wei, E.; and Ozdaglar, A. 2012. Distributed Alternating Direction Method of Multipliers. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, 5445–5450.

Xiao, L.; and Boyd, S. 2004. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1): 65–78.

Zhou, Y.; Yang, J.; Zhang, H.; Liang, Y.; and Tarokh, V. 2019. SGD converges to global minimum in deep learning via star-convex path. *arXiv preprint arXiv:1901.00451*.

Zinkevich, M.; Weimer, M.; Li, L.; and Smola, A. J. 2010. Parallelized Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems 23*, 2595–2603. Curran Associates, Inc.

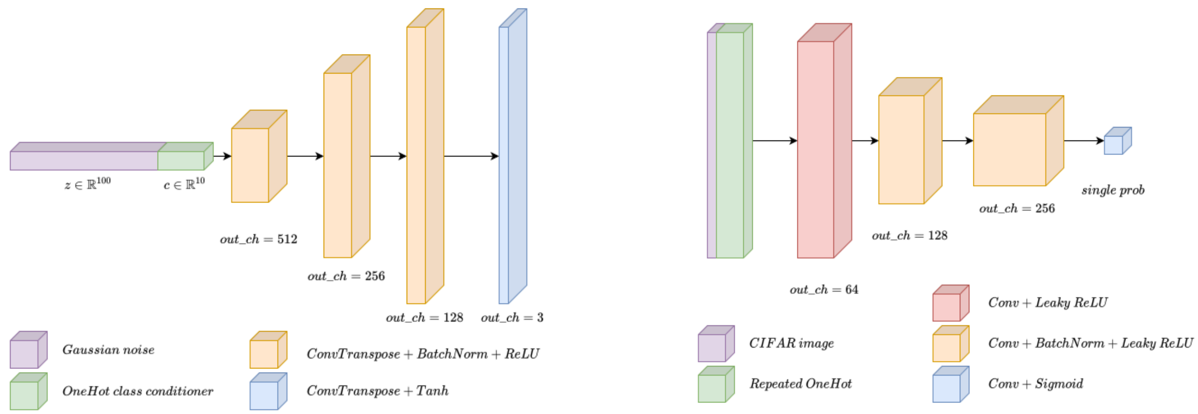
Supplementary Material

A Experiments

We implement all methods in Python 3.8 using PyTorch (Paszke et al. 2019) and Ray (Moritz et al. 2018a) and run on a machine with 24 AMD EPYC 7552 @ 2.20GHz processors, 2 GPUs NVIDIA A100-PCIE with 40536 Mb of memory each (Cuda 11.3).

A.1 Additional information about experiments with GANs

As mentioned in the main part of the paper we use DCGAN (Radford, Metz, and Chintala 2015), conditioned by class labels, similarly to (Mirza and Osindero 2014). See architecture in Figure 5. In Table 2 see hyperparameters for all experiments.



a) Generator b) Discriminator

Figure 5: DCGAN architecture.

Hyperparameters	
Batch size	=64
Weight clipping for the discriminator	=0.01
Learning rate for generator and discriminator	=0.002
Initialization:	normal
Other parameters:	default in PyTorch

Table 2: Hyperparameters for DCGAN training.

Now we will try to discuss the question of how close the GAN model is to the assumptions that we made in the theoretical part. First of all, we note that the goal of the DCGAN experiment is to study how the network topology influences the convergence of the algorithm. Even if the assumptions do not hold, we see that the algorithm performs quite well and is flexible w.r.t. choice of the topology. In the theory, we use pretty standard assumptions usually used by the community. In particular, Assumptions 2 and 4 are classic and are often used in the literature, including the literature on neural networks training. Assumption 5 holds with a small constant D when the data is uniformly split among the devices. It is very easy to realize this splitting when we compute on a cluster with a large number of pictures (data). But in Section 5.2 we look at a more complex setup and make the distribution of pictures not uniform, but heterogeneous. In fact, the constant D exists for any data partitioning, since the gradients can be considered to be bounded because we show that the iterates of the method are bounded. Assumption 3 (NM) is also used in the literature on the analysis of GANs (Liu et al. 2020; Mertikopoulos et al. 2018; Liu et al. 2019a). Moreover, this assumption holds in some nonconvex minimization problems. For example, it holds in both theory and practice when we use SGD for training neural networks (Li and Yuan 2017; Kleinberg, Li, and Yuan 2018; Zhou et al. 2019).

A.2 Additional Experiments with Decreasing Stepsize

Our theoretical results in Theorem 1 hold for (optimally chosen) fixed stepsizes. That is, one has to choose the optimal stepsize in dependency of the target accuracy ϵ (this is standard for theoretical results of this kind). However, in case when the desired target accuracy ϵ is not known, or not determined, one can resort to decreasing stepsizes. In this section, we numerically illustrate that Algorithm 1 can reach arbitrary small error when using decreasing stepsizes.

For that, we consider the same setup as in the main paper Section 5.1, Figure 1, left, i.e. strongly-monotone bi-linear objective functions distributed over the ring topology. We consider two cases: with and without stochastic noise σ, i.e. we fix $\sigma = 0$ and $\sigma = 100$. We decrease stepsize during training as $\gamma_k = \frac{\alpha}{k+\beta}$, where k is the current iteration number. We set $\alpha = 40, \beta = 800$ in the noiseless case and $\alpha = 15, \beta = 150$ with $\sigma^2 = 100$. In Figure 6 we can see that with decreasing stepsizes, algorithm indeed does not have a limiting accuracy, in contrast to constant stepsizes and the sublinearly converges to zero in both cases.

Figure 6: Convergence of Algorithm 1 with decreasing stepsizes in the noiseless (left) and stochastic (right) cases.

B Basic Facts

Upper bound for a squared sum. For arbitrary integer $n \geq 1$ and arbitrary set of vectors $a_1, \ldots, a_n$ we have

$$\left(\sum_{i=1}^{n} a_i\right)^2 \leq m \sum_{i=1}^{n} a_i^2 \tag{12}$$

Cauchy-Schwarz inequality. For arbitrary vectors a and b and any constant $c > 0$

$$2\langle a, b\rangle \leq c\|a\|^2 + c^{-1}\|b\|^2, \tag{13}$$

$$\|a + b\|^2 \leq (1 + c)\|a\|^2 + (1 + c^{-1})\|b\|^2. \tag{14}$$

Cauchy-Schwarz inequality for random variables. Let ξ and η be real valued random variables such that $\mathbb{E}[\xi^2] < \infty$ and $\mathbb{E}[\eta^2] < \infty$. Then

$$\mathbb{E}[\xi\eta] \leq \sqrt{\mathbb{E}[\xi^2]\mathbb{E}[\eta^2]}. \tag{15}$$

Frobenius norm of product. For given matrix A and B

$$\|AB\|_F \leq \|A\|_F\|B\|_2. \tag{16}$$

C Missing proofs

C.1 Notation

To begin with, we introduce auxiliary notation:

- Average z and g values across all devices:

$$\bar{z}^k := \frac{1}{M}\sum_{m=1}^{M} z_m^k, \quad \bar{g}^k := \frac{1}{M}\sum_{m=1}^{M} g_m^k = \frac{1}{M}\sum_{m=1}^{M} F_m(z_m^k, \xi_m^k),$$
$$\bar{z}^{k+1/3} = \bar{z}^k - \gamma\bar{g}^k, \quad \bar{z}^{k+2/3} = \bar{z}^k - \gamma\bar{g}^{k+1/3}, \quad \bar{z}^{k+1} = \bar{z}^{k+2/3} \tag{17}$$

  The last fact: $\bar{z}^{k+1} = \bar{z}^{k+2/3}$ follows from that one step of gossip preserves the average.

- Matrix notation of z, $\bar{z}$, g and $\bar{g}$:

$$Z^k := [z_1^k, \ldots, z_M^k], \quad \bar{Z}^k := [\bar{z}^k, \ldots, \bar{z}^k],$$
$$G^k := [g_1^k, \ldots, g_M^k], \quad \bar{G}^k := [\bar{g}^k, \ldots, \bar{g}^k],$$
$$\Phi^k := [F_1(z_1^k), \ldots, F_M(z_M^k)], \quad \bar{\Phi}^k := \left[\frac{1}{M}\sum_{m=1}^{M} F_m(z_m^k), \ldots\right],$$

following this notation one can rewrite iteration of the Algorithm 1 and "averaged" iteration (17):

$$\begin{aligned} Z^{k+1/3} &= Z^k - \gamma G^k, \quad \bar{Z}^{k+1/3} = \bar{Z}^k - \gamma \bar{G}^k, \\ Z^{k+2/3} &= Z^k - \gamma G^{k+1/3}, \quad \bar{Z}^{k+2/3} = \bar{Z}^k - \gamma \bar{G}^{k+1/3}, \\ Z^{k+1} &= Z^{k+2/3} W^k, \quad \bar{Z}^{k+1} = \bar{Z}^{k+2/3}. \end{aligned} \tag{18}$$

- Error difference between devices:

$$\text{Err}(k) = \frac{1}{M} \sum_{m=1}^{M} \|z_m^k - \bar{z}^k\|^2. \tag{19}$$

C.2 Proof of Theorem 1, strongly-monotone case.

We begin the proof with the following lemma:

Lemma 1 *Let $z, y \in \mathbb{R}^n$. We set $z^+ = z - y$, then for all $u \in \mathbb{R}^n$:*

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

Proof: Simple manipulations give

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

□

Apply this Lemma two times with $z^+ = \bar{z}^{k+2/3}$, $z = \bar{z}^k$, $u = z^*$ and $y = \gamma \bar{g}^{k+1/3}$

$$\|\bar{z}^{k+2/3} - z^*\|^2 = \|\bar{z}^k - z^*\|^2 - 2\gamma\langle \bar{g}^{k+1/3}, \bar{z}^{k+2/3} - z^* \rangle - \|\bar{z}^{k+2/3} - \bar{z}^k\|^2,$$

and with $z^+ = \bar{z}^{k+1/3}$, $z = \bar{z}^k$, $u = z^{k+2/3}$, $y = \gamma \bar{g}^k$:

$$\|\bar{z}^{k+1/3} - \bar{z}^{k+2/3}\|^2 = \|\bar{z}^k - \bar{z}^{k+2/3}\|^2 - 2\gamma\langle \bar{g}^k, \bar{z}^{k+1/3} - \bar{z}^{k+2/3} \rangle - \|\bar{z}^{k+1/3} - \bar{z}^k\|^2.$$

Summing up the two previous inequalities, we have

$$\begin{aligned} \|\bar{z}^{k+2/3} - z^*\|^2 + \|\bar{z}^{k+1/3} - \bar{z}^{k+2/3}\|^2 &= \|\bar{z}^k - z^*\|^2 - \|\bar{z}^{k+1/3} - \bar{z}^k\|^2 \\ &\quad - 2\gamma\langle \bar{g}^{k+1/3}, \bar{z}^{k+2/3} - z^* \rangle - 2\gamma\langle \bar{g}^k, \bar{z}^{k+1/3} - \bar{z}^{k+2/3} \rangle. \end{aligned}$$

A small rearrangement gives

$$\begin{aligned} &\|\bar{z}^{k+2/3} - z^*\|^2 + \|\bar{z}^{k+1/3} - \bar{z}^{k+2/3}\|^2 \\ &\quad = \|\bar{z}^k - z^*\|^2 - \|\bar{z}^{k+1/3} - \bar{z}^k\|^2 \\ &\qquad - 2\gamma\langle \bar{g}^{k+1/3}, \bar{z}^{k+1/3} - z^* \rangle + 2\gamma\langle \bar{g}^{k+1/3} - \bar{g}^k, \bar{z}^{k+1/3} - \bar{z}^{k+2/3} \rangle \\ &\quad \le \|\bar{z}^k - z^*\|^2 - \|\bar{z}^{k+1/3} - \bar{z}^k\|^2 \\ &\qquad - 2\gamma\langle \bar{g}^{k+1/3}, \bar{z}^{k+1/3} - z^* \rangle + \gamma^2\|\bar{g}^{k+1/3} - \bar{g}^k\|^2 + \|\bar{z}^{k+1/3} - \bar{z}^{k+2/3}\|^2, \end{aligned}$$

Next we take the full expectation and get

$$\begin{aligned} \mathbb{E}\left[\|\bar{z}^{k+2/3} - z^*\|^2\right] &= \mathbb{E}\left[\|\bar{z}^k - z^*\|^2\right] - \mathbb{E}\left[\|\bar{z}^{k+1/3} - \bar{z}^k\|^2\right] \\ &\quad - 2\gamma\mathbb{E}\left[\langle \bar{g}^{k+1/3}, \bar{z}^{k+1/3} - z^* \rangle\right] + \gamma^2\mathbb{E}\left[\|\bar{g}^{k+1/3} - \bar{g}^k\|^2\right]. \end{aligned}$$

With $\bar{z}^{k+1} = \bar{z}^{k+2/3}$ we deduce the following inequality per step of Algorithm

$$\begin{aligned} \mathbb{E}\left[\|\bar{z}^{k+1} - z^*\|^2\right] &= \mathbb{E}\left[\|\bar{z}^k - z^*\|^2\right] - \mathbb{E}\left[\|\bar{z}^{k+1/3} - \bar{z}^k\|^2\right] \\ &\quad - 2\gamma\mathbb{E}\left[\langle \bar{g}^{k+1/3}, \bar{z}^{k+1/3} - z^* \rangle\right] + \gamma^2\mathbb{E}\left[\|\bar{g}^{k+1/3} - \bar{g}^k\|^2\right]. \end{aligned} \tag{20}$$

It turns out that we need to estimate two terms: $-2\gamma\mathbb{E}\left[\langle \bar{g}^{k+1/3}, \bar{z}^{k+1/3} - z^* \rangle\right]$ and $\gamma^2\mathbb{E}\left[\|\bar{g}^{k+1/3} - \bar{g}^k\|^2\right]$. For this, we prove two more auxiliary lemmas.

Lemma 2 *Under Assumptions 2, 3, 4 it holds:*

$$-2\gamma\mathbb{E}\left[\langle \bar{g}^{k+1/3}, \bar{z}^{k+1/3} - z^*\rangle\right] \leq -\gamma\mu\mathbb{E}\left[\|\bar{z}^{k+1/3} - z^*\|^2\right] + \frac{\gamma L^2}{\mu}\mathbb{E}\left[\,\mathrm{Err}(k+1/3)\right]. \tag{21}$$

Proof: First of all, we use the independence of all random vectors $\xi^i = (\xi^i_1, \ldots, \xi^i_m)$ and select only the conditional expectation $\mathbb{E}_{\xi^{k+1/3}}$ on vector $\xi^{k+1/3}$:

$$\begin{aligned}
-2\gamma\mathbb{E}\left[\langle \bar{g}^{k+1/3}, \bar{z}^{k+1/3} - z^*\rangle\right] &= -2\gamma\mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\xi^{k+1/3}}[F_m(z_m^{k+1/3}, \xi_m^{k+1/3})], \bar{z}^{k+1/3} - z^*\right\rangle\right] \\
&\overset{(5)}{=} -2\gamma\mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M}F_m(z_m^{k+1/3}), \bar{z}^{k+1/3} - z^*\right\rangle\right] \\
&= -2\gamma\mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M}F_m(\bar{z}^{k+1/3}), \bar{z}^{k+1/3} - z^*\right\rangle\right] \\
&\quad + 2\gamma\mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M}[F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})], \bar{z}^{k+1/3} - z^*\right\rangle\right] \\
&= -2\gamma\mathbb{E}\left[\left\langle F(\bar{z}^{k+1/3}), \bar{z}^{k+1/3} - z^*\right\rangle\right] \\
&\quad + 2\gamma\mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M}[F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})], \bar{z}^{k+1/3} - z^*\right\rangle\right].
\end{aligned}$$

Further, we take into account that for the solution z^* it holds that $\langle F(z^*), \bar{z}^{k+1/3} - z^*\rangle \geq 0$, and then we have:

$$\begin{aligned}
-2\gamma\mathbb{E}\left[\langle \bar{g}^{k+1/3}, \bar{z}^{k+1/3} - z^*\rangle\right] &= -2\gamma\mathbb{E}\left[\left\langle F(\bar{z}^{k+1/3}) - F(z^*), \bar{z}^{k+1/3} - z^*\right\rangle\right] \\
&\quad + 2\gamma\mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M}[F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})], \bar{z}^{k+1/3} - z^*\right\rangle\right] \\
&\overset{\text{(SM)}}{\leq} -2\gamma\mu\mathbb{E}\left[\|\bar{z}^{k+1/3} - z^*\|^2\right] \\
&\quad + 2\gamma\mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M}[F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})], \bar{z}^{k+1/3} - z^*\right\rangle\right].
\end{aligned}$$

By (13) with $\mu > 0$ we get

$$\begin{aligned}
-2\gamma\mathbb{E}\left[\langle \bar{g}^{k+1/3}, \bar{z}^{k+1/3} - z^*\rangle\right] &\leq -2\gamma\mu\mathbb{E}\left[\|\bar{z}^{k+1/3} - z^*\|^2\right] \\
&\quad + \gamma\mu\mathbb{E}\left[\left\|\bar{z}^{k+1/3} - z^*\right\|^2\right] + \frac{\gamma}{\mu}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}[F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})]\right\|^2\right] \\
&= -\gamma\mu\mathbb{E}\left[\|\bar{z}^{k+1/3} - z^*\|^2\right] + \frac{\gamma}{\mu M^2}\mathbb{E}\left[\left\|\sum_{m=1}^{M}[F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})]\right\|^2\right] \\
&\leq -\gamma\mu\mathbb{E}\left[\|\bar{z}^{k+1/3} - z^*\|^2\right] + \frac{\gamma}{\mu M}\mathbb{E}\left[\sum_{m=1}^{M}\left\|F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})\right\|^2\right] \\
&\overset{(4)}{\leq} -\gamma\mu\mathbb{E}\left[\|\bar{z}^{k+1/3} - z^*\|^2\right] + \frac{\gamma L^2}{\mu M}\mathbb{E}\left[\sum_{m=1}^{M}\left\|\bar{z}^{k+1/3} - z_m^{k+1/3}\right\|^2\right].
\end{aligned}$$

Definition (19) ends the proof.

□

Lemma 3 *Under Assumptions 2, 4 it holds that*

$$\mathbb{E}\left[\|\bar{g}^{k+1/3}-\bar{g}^k\|^2\right] \leq 5L^2\mathbb{E}\left[\|\bar{z}^{k+1/3}-\bar{z}^k\|^2\right] + \frac{10\sigma^2}{M} + 5L^2\mathbb{E}\left[Err(k+1/3)\right] + 5L^2\mathbb{E}\left[Err(k)\right]. \quad (22)$$

Proof: Consider the following chain of inequalities:

$$\begin{aligned}
\mathbb{E}\left[\|\bar{g}^{k+1/3}-\bar{g}^k\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M F_m(z_m^{k+1/3},\xi_m^{k+1/3}) - \frac{1}{M}\sum_{m=1}^M F_m(z_m^k,\xi_m^k)\right\|^2\right] \\
&\overset{(12)}{\leq} 5\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^{k+1/3},\xi_m^{k+1/3}) - F_m(z_m^{k+1/3})]\right\|^2\right] \\
&\quad + 5\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^{k+1/3}) - F_m(\bar{z}^{k+1/3})]\right\|^2\right] + 5\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(\bar{z}^{k+1/3}) - F_m(\bar{z}^k)]\right\|^2\right] \\
&\quad + 5\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^k) - F_m(\bar{z}^k)]\right\|^2\right] + 5\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^k,\xi_m^k) - F_m(z_m^k)]\right\|^2\right] \\
&\overset{(12)}{\leq} 5\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^{k+1/3},\xi_m^{k+1/3}) - F_m(z_m^{k+1/3})]\right\|^2\right] \\
&\quad + 5\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^k,\xi_m^k) - F_m(z_m^k)]\right\|^2\right] \\
&\quad + \frac{5}{M}\sum_{m=1}^M \mathbb{E}\left[\left\|F_m(z_m^{k+1/3}) - F_m(\bar{z}^{k+1/3})\right\|^2\right] + \frac{5}{M}\sum_{m=1}^M \mathbb{E}\left[\left\|F_m(z_m^k) - F_m(\bar{z}^k)\right\|^2\right] \\
&\quad + 5\mathbb{E}\left[\left\|F(\bar{z}^{k+1/3}) - F(\bar{z}^k)\right\|^2\right] \\
&\overset{(4),(19)}{\leq} 5\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^{k+1/3},\xi_m^{k+1/3}) - F_m(z_m^{k+1/3})]\right\|^2\right] \\
&\quad + 5\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^k,\xi_m^k) - F_m(z_m^k)]\right\|^2\right] \\
&\quad + 5L^2\mathbb{E}\left[\text{Err}(k+1/3)\right] + 5L^2\mathbb{E}\left[\text{Err}(k)\right] + 5L^2\mathbb{E}\left[\|\bar{z}^{k+1/3}-\bar{z}^k\|^2\right] \\
&= 5\mathbb{E}\left[\mathbb{E}_{\xi_{k+1/3}}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^{k+1/3},\xi_m^{k+1/3}) - F_m(z_m^{k+1/3})]\right\|^2\right]\right] \\
&\quad + 5\mathbb{E}\left[\mathbb{E}_{\xi_k}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(z_m^k,\xi_m^k) - F_m(z_m^k)]\right\|^2\right]\right] \\
&\quad + 5L^2\mathbb{E}\left[\text{Err}(k+1/3)\right] + 5L^2\mathbb{E}\left[\text{Err}(k)\right] + 5L^2\mathbb{E}\left[\|\bar{z}^{k+1/3}-\bar{z}^k\|^2\right].
\end{aligned}$$

Using independence of each machine and (5), we get:

$$\mathbb{E}\left[\|\bar{g}^{k+1/3}-\bar{g}^k\|^2\right] \leq \frac{10\sigma^2}{M} + 5L^2\mathbb{E}\left[\text{Err}(k+1/3)\right] + 5L^2\mathbb{E}\left[\text{Err}(k)\right] + 5L^2\mathbb{E}\left[\|\bar{z}^{k+1/3}-\bar{z}^k\|^2\right].$$

□

Let's go back to the proof of Theorem and connect (20), (21) and (22):

$$\begin{aligned}\mathbb{E}\left[\|\bar{z}^{k+1}-z^*\|^2\right] &\leq \mathbb{E}\left[\|\bar{z}^{k}-z^*\|^2\right] - \mathbb{E}\left[\|\bar{z}^{k+1/3}-\bar{z}^k\|^2\right] \\ &- \gamma\mu\mathbb{E}\left[\|\bar{z}^{k+1/3}-z^*\|^2\right] + \frac{\gamma L^2}{\mu}\mathbb{E}\left[\mathrm{Err}(k+1/3)\right] \\ &+ \gamma^2\left(5L^2\mathbb{E}\left[\|\bar{z}^{k+1/3}-\bar{z}^k\|^2\right] + \frac{10\sigma^2}{M} + 5L^2\mathbb{E}\left[\mathrm{Err}(k+1/3)\right] + 5L^2\mathbb{E}\left[\mathrm{Err}(k)\right]\right)\end{aligned}$$

By (14) with $c=1$, $a=\bar{z}^{k+1/3}-z^*$ and $b=\bar{z}^{k+1/3}-\bar{z}^k$ we get

$$\begin{aligned}\mathbb{E}\left[\|\bar{z}^{k+1}-z^*\|^2\right] &\leq \left(1-\frac{\gamma\mu}{2}\right)\mathbb{E}\left[\|\bar{z}^{k}-z^*\|^2\right] - (1-5\gamma^2L^2-\gamma\mu)\mathbb{E}\left[\|\bar{z}^{k+1/3}-\bar{z}^k\|^2\right] \\ &+ \left(\frac{\gamma L^2}{\mu}+5\gamma^2L^2\right)\mathbb{E}\left[\mathrm{Err}(k+1/3)\right] + 5\gamma^2L^2\mathbb{E}\left[\mathrm{Err}(k)\right] + \frac{10\gamma^2\sigma^2}{M}. \end{aligned} \tag{23}$$

Choosing $\gamma \leq \frac{1}{3L}$ gives

$$\begin{aligned}\mathbb{E}\left[\|\bar{z}^{k+1}-z^*\|^2\right] &\leq \left(1-\frac{\gamma\mu}{2}\right)\mathbb{E}\left[\|\bar{z}^{k}-z^*\|^2\right] \\ &+ \left(\frac{\gamma L^2}{\mu}+5\gamma^2L^2\right)\mathbb{E}\left[\mathrm{Err}(k+1/3)\right] + 5\gamma^2L^2\mathbb{E}\left[\mathrm{Err}(k)\right] + \frac{10\gamma^2\sigma^2}{M}. \end{aligned} \tag{24}$$

Now we need to bound $\mathbb{E}\left[\mathrm{Err}(k)\right]$ and $\mathbb{E}\left[\mathrm{Err}(k+1/3)\right]$. For this we need one more lemma.

Lemma 4 *Under Assumptions 2, 4, 5, 1 it holds that*

$$\begin{aligned}\mathbb{E}\left[\mathit{Err}(k)\right] &\leq \left(1-\frac{3p}{4}\right)\mathbb{E}[\mathit{Err}(h\tau)] + \frac{144\gamma^2L^2\tau}{p}\sum_{j=h\tau}^{k-1}\mathbb{E}\left[\mathit{Err}(j+1/3)\right] \\ &+ \left(\frac{72D^2\tau}{p}+8\sigma^2\right)\sum_{j=h\tau}^{k-1}\gamma^2 \end{aligned} \tag{25}$$

$$\begin{aligned}\mathbb{E}\left[\mathit{Err}(k+1/3)\right] &\leq \left(1-\frac{3p}{4}\right)\mathbb{E}[\mathit{Err}(h\tau)] + \frac{216\gamma^2L^2\tau}{p}\sum_{j=h\tau}^{k-1}\mathbb{E}\left[\mathit{Err}(j+1/3)\right] + \frac{216\gamma^2L^2\tau}{p}\mathbb{E}\left[\mathit{Err}(k)\right] \\ &+ \left(\frac{108D^2\tau}{p}+12\sigma^2\right)\sum_{j=h\tau}^{k-1}\gamma^2 + \left(\frac{108D^2\tau}{p}+12\sigma^2\right)\gamma^2. \end{aligned} \tag{26}$$

where we define $h=\lfloor k/\tau\rfloor - 1$.

Proof: Using matrix notation introduced in (18) one can get

$$\begin{aligned} M\cdot\mathbb{E}\left[\mathrm{Err}(k)\right] &= \mathbb{E}\|X^k-\bar{X}^k\|_F^2 = \mathbb{E}\|X^k-\bar{X}^{h\tau}-\bar{X}^k+\bar{X}^{h\tau}\|_F^2 \\ &= \mathbb{E}\Bigg[\Bigg\|X^{h\tau}\prod_{i=k-1}^{h\tau}W^i - \bar{X}^{h\tau} - \gamma\sum_{j=h\tau}^{k-1}G^{j+1/3}\prod_{i=k-1}^{j}W^i \\ &\quad - \left(\bar{X}^{h\tau}\prod_{i=k-1}^{h\tau}W^i - \bar{X}^{h\tau} - \gamma\sum_{j=h\tau}^{k-1}\bar{G}^{j+1/3}\prod_{i=k-1}^{j}W^i\right)\Bigg\|_F^2\Bigg] \\ &= \mathbb{E}\Bigg[\mathbb{E}_{\xi^{k-1+1/3}}\Bigg[\Bigg\|X^{h\tau}\prod_{i=k-1}^{h\tau}W^i - \bar{X}^{h\tau} - \left(\bar{X}^{h\tau}\prod_{i=k-1}^{h\tau}W^i - \bar{X}^{h\tau}\right) \\ &\quad - \gamma\sum_{j=h\tau}^{k-1}(\Phi^{j+1/3}-\bar{\Phi}^{j+1/3})\prod_{i=k-1}^{j}W^i \\ &\quad - \gamma\sum_{j=h\tau}^{k-1}(G^{j+1/3}-\Phi^{j+1/3}-\bar{G}^{j+1/3}+\bar{\Phi}^{j+1/3})\prod_{i=k-1}^{j}W^i\Bigg\|_F^2\Bigg]\Bigg]. \end{aligned}$$

Taking into account that only $G^{k-1+1/3}$ and $\Phi^{k-1+1/3}$ depend on $\xi^{k-1+1/3}$, as well as the unbiasedness of $G^{k-1+1/3}$, we have

$$
\begin{aligned}
M \cdot \mathbb{E}\left[\mathrm{Err}(k)\right] = \mathbb{E}\Bigg[\Bigg\| & X^{h\tau} \prod_{i=k-1}^{h\tau} W^i - \bar{X}^{h\tau} - \left(\bar{X}^{h\tau} \prod_{i=k-1}^{h\tau} W^i - \bar{X}^{h\tau}\right) \\
& - \gamma \sum_{j=h\tau}^{k-1} (\Phi^{j+1/3} - \bar{\Phi}^{j+1/3}) \prod_{i=k-1}^{j} W^i \\
& - \gamma \sum_{j=h\tau}^{k-2} (G^{j+1/3} - \Phi^{j+1/3} - \bar{G}^{j+1/3} + \bar{\Phi}^{j+1/3}) \prod_{i=k-1}^{j} W^i \Bigg\|_F^2\Bigg] \\
& + \gamma^2 \mathbb{E}\left[\left\|(G^{k-1+1/3} - \Phi^{k-1+1/3} - \bar{G}^{k-1+1/3} + \bar{\Phi}^{k-1+1/3}) W^{k-1}\right\|_F^2\right].
\end{aligned}
$$

We want to continue the same way, but note that $X^{k-1+1/3}$ (and $\Phi^{k-1+1/3}$, $\bar{\Phi}^{k-1+1/3}$) depends on $\xi^{k-2+1/3}$, then we apply (14) with $c = \beta_1$ and get

$$
\begin{aligned}
M \cdot \mathbb{E}\left[\mathrm{Err}(k)\right] \le (1+\beta_1)\mathbb{E}\Bigg[\Bigg\| & X^{h\tau} \prod_{i=k-1}^{h\tau} W^i - \bar{X}^{h\tau} - \left(\bar{X}^{h\tau} \prod_{i=k-1}^{h\tau} W^i - \bar{X}^{h\tau}\right) \\
& - \gamma \sum_{j=h\tau}^{k-2} (\Phi^{j+1/3} - \bar{\Phi}^{j+1/3}) \prod_{i=k-1}^{j} W^i \\
& - \gamma \sum_{j=h\tau}^{k-2} (G^{j+1/3} - \Phi^{j+1/3} - \bar{G}^{j+1/3} + \bar{\Phi}^{j+1/3}) \prod_{i=k-1}^{j} W^i \Bigg\|_F^2\Bigg] \\
& + (1+\beta_1^{-1})\gamma^2 \mathbb{E}\left[\left\|\Phi^{k-1+1/3} - \bar{\Phi}^{k-1+1/3}\right\|_F^2\right] \\
& + \gamma^2 \mathbb{E}\left[\left\|G^{k-1+1/3} - \Phi^{k-1+1/3} - \bar{G}^{k-1+1/3} + \bar{\Phi}^{k-1+1/3}\right\|_F^2\right].
\end{aligned}
$$

We also use (16) in last two lines. Now, similarly, we split terms that depend on $X^{k-2+1/3}$ with $c = \beta_2$.

$$
\begin{aligned}
M \cdot \mathbb{E}\left[\mathrm{Err}(k)\right] \le (1+\beta_1)(1+\beta_2)\mathbb{E}\Bigg[\Bigg\| & X^{h\tau} \prod_{i=k-1}^{h\tau} W^i - \bar{X}^{h\tau} - \left(\bar{X}^{h\tau} \prod_{i=k-1}^{h\tau} W^i - \bar{X}^{h\tau}\right) \\
& - \gamma \sum_{j=h\tau}^{k-3} (\Phi^{j+1/3} - \bar{\Phi}^{j+1/3}) \prod_{i=k-1}^{j} W^i \\
& - \gamma \sum_{j=h\tau}^{k-3} (G^{j+1/3} - \Phi^{j+1/3} - \bar{G}^{j+1/3} + \bar{\Phi}^{j+1/3}) \prod_{i=k-1}^{j} W^i \Bigg\|_F^2\Bigg] \\
& + (1+\beta_1^{-1})\gamma^2 \mathbb{E}\left[\left\|\Phi^{k-1+1/3} - \bar{\Phi}^{k-1+1/3}\right\|_F^2\right] \\
& + (1+\beta_1)(1+\beta_2^{-1})\gamma^2 \mathbb{E}\left[\left\|\Phi^{k-2+1/3} - \bar{\Phi}^{k-2+1/3}\right\|_F^2\right] \\
& + (1+\beta_1)\gamma^2 \mathbb{E}\left[\left\|G^{k-2+1/3} - \Phi^{k-2+1/3} - \bar{G}^{k-2+1/3} + \bar{\Phi}^{k-2+1/3}\right\|_F^2\right] \\
& + \gamma^2 \mathbb{E}\left[\left\|G^{k-1+1/3} - \Phi^{k-1+1/3} - \bar{G}^{k-1+1/3} + \bar{\Phi}^{k-1+1/3}\right\|_F^2\right].
\end{aligned}
$$

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

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

Note that $k - 1 - h\tau \leq 2\tau$, hence for all $i = 0, \ldots, (k - 1 - h\tau)$

$$(1 + \beta_1)(1 + \beta_2) \ldots (1 + \beta_i) \leq (1 + \beta_1)(1 + \beta_2) \ldots (1 + \beta_{k-1-h\tau}) \leq \frac{\alpha}{\alpha - 2\tau} \leq 2.$$

Additionally, $1 + \beta_i^{-1} \leq \alpha$, then

$$\begin{aligned}
M \cdot \mathbb{E}\left[\mathrm{Err}(k)\right] \leq \frac{\alpha}{\alpha - 2\tau} \mathbb{E}\left[\left\| X^{h\tau} \prod_{i=k-1}^{h\tau} W^i - \bar{X}^{h\tau} - \left(\bar{X}^{h\tau} \prod_{i=k-1}^{h\tau} W^i - \bar{X}^{h\tau} \right) \right\|_F^2 \right] \\
+ 2\gamma^2 \alpha \sum_{j=h\tau}^{k-1} \mathbb{E}\left[\left\| \Phi^{j+1/3} - \bar{\Phi}^{j+1/3} \right\|_F^2 \right] \\
+ 2\gamma^2 \sum_{j=h\tau}^{k-1} \mathbb{E}\left[\left\| G^{j+1/3} - \Phi^{j+1/3} - \bar{G}^{j+1/3} + \bar{\Phi}^{j+1/3} \right\|_F^2 \right].
\end{aligned}$$

With $\alpha = 4\tau \left(1 + \frac{2}{p}\right)$ we get

$$\begin{aligned}
M \cdot \mathbb{E}\left[\mathrm{Err}(k)\right] \leq \left(1 + \frac{1}{1 + \frac{4}{p}}\right) \mathbb{E}\left[\left\| X^{h\tau} \prod_{i=k-1}^{h\tau} W^i - \bar{X}^{h\tau} - \left(\bar{X}^{h\tau} \prod_{i=k-1}^{h\tau} W^i - \bar{X}^{h\tau} \right) \right\|_F^2 \right] \\
+ \frac{24\gamma^2\tau}{p} \sum_{j=h\tau}^{k-1} \mathbb{E}\left[\left\| \Phi^{j+1/3} - \bar{\Phi}^{j+1/3} \right\|_F^2 \right] \\
+ 2\gamma^2 \sum_{j=h\tau}^{k-1} \mathbb{E}\left[\left\| G^{j+1/3} - \Phi^{j+1/3} - \bar{G}^{j+1/3} + \bar{\Phi}^{j+1/3} \right\|_F^2 \right].
\end{aligned}$$

Next, one can note that $\|A - \bar{A}\|_F^2 \leq \|A\|_F^2$ and hence

$$\begin{aligned}
M \cdot \mathbb{E}\left[\mathrm{Err}(k)\right] &\leq \left(1 + \frac{1}{1 + \frac{4}{p}}\right) \mathbb{E}\left[\left\| X^{h\tau} \prod_{i=k-1}^{h\tau} W^i - \bar{X}^{h\tau} \right\|_F^2 \right] \\
&\quad + \frac{24\gamma^2\tau}{p} \sum_{j=h\tau}^{k-1} \mathbb{E}\left[\left\| \Phi^{j+1/3} - \bar{\Phi}^{j+1/3} \right\|_F^2 \right] \\
&\quad + 2\gamma^2 \sum_{j=h\tau}^{k-1} \mathbb{E}\left[\left\| G^{j+1/3} - \Phi^{j+1/3} - \bar{G}^{j+1/3} + \bar{\Phi}^{j+1/3} \right\|_F^2 \right] \\
&\overset{(3)}{\leq} (1 - p) \left(1 + \frac{1}{1 + \frac{4}{p}}\right) \mathbb{E}\left[\left\| X^{h\tau} - \bar{X}^{h\tau} \right\|_F^2 \right] \\
&\quad + \frac{24\gamma^2\tau}{p} \sum_{j=h\tau}^{k-1} \mathbb{E}\left[\left\| \Phi^{j+1/3} - \bar{\Phi}^{j+1/3} \right\|_F^2 \right] \\
&\quad + 2\gamma^2 \sum_{j=h\tau}^{k-1} \mathbb{E}\left[\left\| G^{j+1/3} - \Phi^{j+1/3} - \bar{G}^{j+1/3} + \bar{\Phi}^{j+1/3} \right\|_F^2 \right]. \qquad (27)
\end{aligned}$$

It is easy to see that $(1-p)\left(1+\frac{1}{1+\frac{4}{p}}\right) \leq (1-p)\left(1+\frac{p}{4}\right) \leq \left(1-\frac{3p}{4}\right)$. It remains to estimate

$$
\begin{aligned}
\mathbb{E}\left[\left\|\Phi^{j+1/3}-\bar{\Phi}^{j+1/3}\right\|_F^2\right] &= \sum_{m=1}^{M}\left[\mathbb{E}\left\|F_m(z_m^{j+1/3})-\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{j+1/3})\right\|^2\right] \\
&\overset{(12)}{\leq} 3\sum_{m=1}^{M}\left[\mathbb{E}\left\|F_m(z_m^{j+1/3})-F_m(\bar{z}^{j+1/3})\right\|^2+\mathbb{E}\left\|F_m(\bar{z}^{j+1/3})-\frac{1}{M}\sum_{i=1}^{M}F_i(\bar{z}^{j+1/3})\right\|^2\right. \\
&\quad \left.+\mathbb{E}\left\|\frac{1}{M}\sum_{i=1}^{M}F_i(\bar{z}^{j+1/3})-\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{j+1/3})\right\|^2\right] \\
&\overset{(6)}{\leq} 3\sum_{m=1}^{M}\left[D^2+\mathbb{E}\left\|\frac{1}{M}\sum_{i=1}^{M}F_i(\bar{z}^{j+1/3})-\frac{1}{M}\sum_{i=1}^{M}F_i(z_i^{j+1/3})\right\|^2\right. \\
&\quad \left.+\mathbb{E}\left\|F_m(z_m^{j+1/3})-F_m(\bar{z}^{j+1/3})\right\|^2\right] \\
&\overset{(4)}{\leq} 6ML^2\mathbb{E}\left[\text{Err}(j+1/3)\right]+3MD^2.
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}\left[\left\|G^{j+1/3}-\Phi^{j+1/3}-\bar{G}^{j+1/3}+\bar{\Phi}^{j+1/3}\right\|_F^2\right] \\
&= \sum_{m=1}^{M}\left[\mathbb{E}\left\|F_m(z_m^{j+1/3},\xi_m^{j+1/3})-F_m(z_m^{j+1/3})-\frac{1}{M}\sum_{i=1}^{M}\left(F_i(z_i^{j+1/3},\xi_i^{j+1/3})-F_i(z_i^{j+1/3})\right)\right\|^2\right] \\
&\overset{(12)}{\leq} 2\sum_{m=1}^{M}\left[\mathbb{E}\left\|F_m(z_m^{j+1/3},\xi_m^{j+1/3})-F_m(z_m^{j+1/3})\right\|^2\right. \\
&\quad \left.+\mathbb{E}\left\|\frac{1}{M}\sum_{i=1}^{M}\left(F_i(z_i^{j+1/3},\xi_i^{j+1/3})-F_i(z_i^{j+1/3})\right)\right\|^2\right] \\
&\overset{(5)}{\leq} 4M\sigma^2.
\end{aligned}
$$

Finally, we get

$$
\mathbb{E}\left[\text{Err}(k)\right] \leq \left(1-\frac{3p}{4}\right)\mathbb{E}[\text{Err}(h\tau)]+\frac{144\gamma^2L^2\tau}{p}\sum_{j=h\tau}^{k-1}\mathbb{E}\left[\text{Err}(j+1/3)\right]+\left(\frac{72D^2\tau}{p}+8\sigma^2\right)\sum_{j=h\tau}^{k-1}\gamma^2.
$$

The estimate for $\mathbb{E}\left[\text{Err}(k+1/3)\right]$ is done in a similar way; it is enough to note just that $M\mathbb{E}\left[\text{Err}(k+1/3)\right] = \mathbb{E}\|X^k-\gamma G^k-\bar{X}^k+\gamma\bar{G}^k\|_F^2$. In the course of the proof, we need to take $\alpha = 4\tau\left(1+\frac{2}{p}\right)-1$ and to add $\beta_0 = \frac{1}{\alpha}$ for term connecting with $G^k-\bar{G}^k$. Then $(1+\beta_0)(1+\beta_1)(1+\beta_2)\ldots(1+\beta_i) \leq (1+\beta_0)(1+\beta_1)(1+\beta_2)\ldots(1+\beta_{k-1-h\tau}) \leq \frac{\alpha+1}{\alpha-2\tau} \leq 3$, $(1+\beta_i^{-1}) \leq \alpha+1$. And we get

$$
\begin{aligned}
\mathbb{E}\left[\text{Err}(k+1/3)\right] \leq &\left(1-\frac{3p}{4}\right)\mathbb{E}[\text{Err}(h\tau)]+\frac{216\gamma^2L^2\tau}{p}\sum_{j=h\tau}^{k-1}\mathbb{E}\left[\text{Err}(j+1/3)\right]+\frac{216\gamma^2L^2\tau}{p}\mathbb{E}\left[\text{Err}(k)\right] \\
&+\left(\frac{108D^2\tau}{p}+12\sigma^2\right)\sum_{j=h\tau}^{k-1}\gamma^2+\left(\frac{108D^2\tau}{p}+12\sigma^2\right)\gamma^2.
\end{aligned}
$$

□

The previous Lemma is valid for $k \geq (h+1)\tau$. For further analysis we also need estimates for the case when $(h+1)\tau > k \geq h\tau$:

Lemma 5 *Under Assumptions 2, 4, 5, 1 it holds that for* $(h+1)\tau > k \geq h\tau$

$$\mathbb{E}\left[Err(k)\right] \leq \left(1+\frac{p}{4}\right)\mathbb{E}[Err(h\tau)] + \frac{144\gamma^2 L^2\tau}{p}\sum_{j=h\tau}^{k-1}\mathbb{E}\left[Err(j+1/3)\right] + \left(\frac{72D^2\tau}{p} + 8\sigma^2\right)\sum_{j=h\tau}^{k-1}\gamma^2 \tag{28}$$

$$\mathbb{E}\left[Err(k+1/3)\right] \leq \left(1+\frac{p}{4}\right)\mathbb{E}[Err(h\tau)] + \frac{216\gamma^2 L^2\tau}{p}\sum_{j=h\tau}^{k-1}\mathbb{E}\left[Err(j+1/3)\right] + \frac{216\gamma^2 L^2\tau}{p}\mathbb{E}\left[Err(k)\right] + \left(\frac{108D^2\tau}{p} + 12\sigma^2\right)\sum_{j=h\tau}^{k-1}\gamma^2 + \left(\frac{108D^2\tau}{p} + 12\sigma^2\right)\gamma^2. \tag{29}$$

where we define $h = \lfloor k/\tau \rfloor - 1$.

Proof: We only modify (27), because we cannot use (3) for such k.

$$\begin{aligned}
M\cdot\mathbb{E}\left[\mathrm{Err}(k)\right] \leq& \left(1+\frac{1}{1+\frac{4}{p}}\right)\mathbb{E}\left[\left\|\left(X^{h\tau}-\bar{X}^{h\tau}\right)\prod_{i=k-1}^{h\tau}W^i\right\|_F^2\right] \\
&+ \frac{24\gamma^2\tau}{p}\sum_{j=h\tau}^{k-1}\mathbb{E}\left[\left\|\Phi^{j+1/3}-\bar{\Phi}^{j+1/3}\right\|_F^2\right] \\
&+ 2\gamma^2\sum_{j=h\tau}^{k-1}\mathbb{E}\left[\left\|G^{j+1/3}-\Phi^{j+1/3}-\bar{G}^{j+1/3}+\bar{\Phi}^{j+1/3}\right\|_F^2\right] \\
\overset{(16)}{\leq}& \left(1+\frac{1}{1+\frac{4}{p}}\right)\mathbb{E}\left[\left\|X^{h\tau}-\bar{X}^{h\tau}\right\|_F^2\right] \\
&+ \frac{24\gamma^2\tau}{p}\sum_{j=h\tau}^{k-1}\mathbb{E}\left[\left\|\Phi^{j+1/3}-\bar{\Phi}^{j+1/3}\right\|_F^2\right] \\
&+ 2\gamma^2\sum_{j=h\tau}^{k-1}\mathbb{E}\left[\left\|G^{j+1/3}-\Phi^{j+1/3}-\bar{G}^{j+1/3}+\bar{\Phi}^{j+1/3}\right\|_F^2\right].
\end{aligned}$$

□

Further proof is reduced to solving recurrent (24), (25), (26), (28) and (29). Note that in the general case $\mathbb{E}\left[\mathrm{Err}(k+1/3)\right]$ may be less than $\mathbb{E}\left[\mathrm{Err}(k)\right]$, but since recurrent (26) is stronger than (25), we assume for simplicity that $\mathbb{E}\left[\mathrm{Err}(k+1/3)\right] \geq \mathbb{E}\left[\mathrm{Err}(k)\right]$. Then the resulting recurrences are written as follows (Here we additionally use that $\gamma \leq \frac{p}{120\tau L}$):

$$\begin{aligned}
\mathbb{E}\left[\|\bar{z}^{k+1}-z^*\|^2\right] &\leq \left(1-\frac{\gamma\mu}{2}\right)\mathbb{E}\left[\|\bar{z}^k-z^*\|^2\right] + \left(\frac{\gamma L^2}{\mu}+10\gamma^2L^2\right)\mathbb{E}\left[\mathrm{Err}(k+1/3)\right] + \frac{10\gamma^2\sigma^2}{M} \\
&\leq \left(1-\frac{\gamma\mu}{2}\right)\mathbb{E}\left[\|\bar{z}^k-z^*\|^2\right] + \frac{2\gamma L^2}{\mu}\mathbb{E}\left[\mathrm{Err}(k+1/3)\right] + \frac{10\gamma^2\sigma^2}{M}.
\end{aligned}$$

$$\left(1-\frac{216\gamma^2L^2\tau}{p}\right)\mathbb{E}\left[\mathrm{Err}(k+1/3)\right] \leq \left(1-\frac{3p}{4}\right)\mathbb{E}[\mathrm{Err}(h\tau+1/3)] + \frac{216\gamma^2L^2\tau}{p}\sum_{j=h\tau}^{k-1}\mathbb{E}\left[\mathrm{Err}(j+1/3)\right] + \left(\frac{216D^2\tau}{p}+24\sigma^2\right)\sum_{j=h\tau}^{k-1}\gamma^2.$$

$$\left(1-\frac{p}{64}\right)\mathbb{E}\left[\mathrm{Err}(k+1/3)\right] \leq \left(1-\frac{3p}{4}\right)\mathbb{E}[\mathrm{Err}(h\tau+1/3)] + \frac{p}{64\tau}\sum_{j=h\tau}^{k-1}\mathbb{E}\left[\mathrm{Err}(j+1/3)\right]$$
$$+\left(\frac{216D^2\tau}{p}+24\sigma^2\right)\sum_{j=h\tau}^{k-1}\gamma^2.$$

$$\mathbb{E}\left[\mathrm{Err}(k+1/3)\right] \leq \left(1-\frac{p}{2}\right)\mathbb{E}[\mathrm{Err}(h\tau+1/3)] + \frac{p}{64\tau}\sum_{j=h\tau}^{k-1}\mathbb{E}\left[\mathrm{Err}(j+1/3)\right]$$
$$+\left(\frac{225D^2\tau}{p}+25\sigma^2\right)\sum_{j=h\tau}^{k-1}\gamma^2.$$

In the last inequality, we took into account that $0 < p \leq 1$, in particular, $\left(1-\frac{3p}{4}\right)\left(1-\frac{p}{64}\right)^{-1} \leq 1-\frac{p}{2}$. Analogically,

$$\mathbb{E}\left[\mathrm{Err}(k+1/3)\right] \leq \left(1+\frac{p}{2}\right)\mathbb{E}[\mathrm{Err}(h\tau)] + \frac{216\gamma^2L^2\tau}{p}\sum_{j=h\tau}^{k-1}\mathbb{E}\left[\mathrm{Err}(j+1/3)\right] + \frac{216\gamma^2L^2\tau}{p}\mathbb{E}\left[\mathrm{Err}(k)\right]$$
$$+\left(\frac{108D^2\tau}{p}+12\sigma^2\right)\sum_{j=h\tau}^{k-1}\gamma^2 + \left(\frac{108D^2\tau}{p}+12\sigma^2\right)\gamma^2.$$

With $r_k = \mathbb{E}\left[\|\bar{z}^k - z^*\|^2\right]$, $e_k = \mathbb{E}\left[\mathrm{Err}(k+1/3)\right]$, $a = \frac{\mu}{2}$, $B = \frac{2L^2}{\mu}$ and $C = \frac{10\sigma^2}{M}$ and $A = \frac{225D^2\tau}{p} + 25\sigma^2$ we get

$$r_{k+1} \leq (1-\gamma a)\, r_k + \gamma B e_k + \gamma^2 C. \tag{30}$$

$$e_k \leq \left(1-\frac{p}{2}\right)e_{h\tau} + \frac{p}{64\tau}\sum_{j=h\tau}^{k-1}e_j + A\sum_{j=h\tau}^{k-1}\gamma^2, \quad k \geq (h+1)\tau. \tag{31}$$

$$e_k \leq \left(1+\frac{p}{2}\right)e_{h\tau} + \frac{p}{64\tau}\sum_{j=h\tau}^{k-1}e_j + A\sum_{j=h\tau}^{k-1}\gamma^2, \quad h\tau \leq k < (h+1)\tau. \tag{32}$$

For such sequences, we can apply the following lemma:

Lemma 6 *If non-negative sequence $\{e_k\}$ satisfy* (31) *and* (32) *with some constants $0 < \tilde{p} \leq 1$, $\tau \geq 1$, $A \geq 0$. Then for non-negative sequence $\{w_k\}$ it holds that*

$$e_k \leq \frac{8\gamma^2 A\tau}{p}.$$

Proof: We start from (31) and substitute all e_j for $j \geq (h+1)\tau$ from $k-1$ to $(h+1)\tau$:

$$e_k \leq \left(1-\frac{p}{2}\right)\cdot\left(1+\frac{p}{64\tau}\right)e_{h\tau} + \frac{p}{64\tau}\left(1+\frac{p}{64\tau}\right)\sum_{j=h\tau}^{k-2}e_j + A\sum_{j=h\tau}^{k-1}\gamma^2 + \frac{p}{64\tau}\cdot A\sum_{j=h\tau}^{k-2}\gamma^2$$
$$\leq \left(1-\frac{p}{2}\right)\cdot\left(1+\frac{p}{64\tau}\right)^{\tau}e_{h\tau} + \frac{p}{64\tau}\left(1+\frac{p}{64\tau}\right)^{\tau}\sum_{j=h\tau}^{(h+1)\tau-1}e_j$$
$$+ A\left(1+\frac{p}{64\tau}\right)^{k-(h+1)\tau}\sum_{j=h\tau}^{(h+1)\tau-1}\gamma^2 + A\sum_{j=(h+1)\tau}^{k-1}\left(1+\frac{p}{64\tau}\right)^{k-1-j}\gamma^2.$$

Then we substitute all e_j for $h\tau \leq k < (h+1)\tau$ using (32):

$$e_k \leq \left(1-\frac{p}{2}+\frac{p}{64\tau}\left(1+\frac{p}{2}\right)\right)\cdot\left(1+\frac{p}{64\tau}\right)^{\tau}e_{h\tau} + \frac{p}{64\tau}\left(1+\frac{p}{64\tau}\right)^{\tau+1}\sum_{j=h\tau}^{(h+1)\tau-2}e_j$$
$$+ A\left(1+\frac{p}{64\tau}\right)^{k-(h+1)\tau+1}\sum_{j=h\tau}^{(h+1)\tau-2}\gamma^2 + A\sum_{j=(h+1)\tau-1}^{k-1}\left(1+\frac{p}{64\tau}\right)^{k-1-j}\gamma^2.$$

With $\frac{p}{64\tau}\left(1+\frac{p}{2}\right) \leq \frac{p}{16\tau}\left(1-\frac{p}{2}\right)$ we get

$$e_k \leq \left(1-\frac{p}{2}\right)\left(1+\frac{p}{16\tau}\right)\left(1+\frac{p}{64\tau}\right)^{\tau} e_{h\tau} + \frac{p}{64\tau}\left(1+\frac{p}{64\tau}\right)^{\tau+1}\sum_{j=h\tau}^{(h+1)\tau-2} e_j$$
$$+ A\left(1+\frac{p}{64\tau}\right)^{k-(h+1)\tau+1}\sum_{j=h\tau}^{(h+1)\tau-2}\gamma^2 + A\sum_{j=(h+1)\tau-1}^{k-1}\left(1+\frac{p}{64\tau}\right)^{k-1-j}\gamma^2.$$

Making the same way for the rest e_j, we have

$$e_k \leq \left(1-\frac{p}{2}\right)\left(1+\frac{p}{16\tau}\right)^{2\tau} e_{h\tau} + A\sum_{j=h\tau}^{k-1}\left(1+\frac{p}{64\tau}\right)^{k-1-j}\gamma^2.$$

Then one can note that $\left(1+\frac{p}{64\tau}\right)^{k-1-j} \leq \left(1+\frac{p}{16\tau}\right)^{2\tau} \leq \exp(p/8) \leq 1+\frac{p}{4}$ for $p \leq 1$ and then

$$e_k \leq \left(1-\frac{p}{4}\right) e_{h\tau} + 2A\sum_{j=h\tau}^{k-1}\gamma^2.$$

It remains to run recursion for $e_{h\tau}$:

$$e_k \leq 2A\gamma^2\sum_{j=0}^{k-1}\left(1-\frac{p}{4}\right)^{\lfloor (k-j)/\tau \rfloor}.$$

For $p \leq 1$ it holds that $\left(1-\frac{p}{4}\right)^{1/\tau} \leq \exp(-p/4\tau) \leq 1-\frac{p}{4\tau}$, hence

$$e_k \leq 2A\gamma^2\sum_{j=0}^{k-1}\left(1-\frac{p}{4\tau}\right)^{k-j} \leq \frac{8\gamma^2 A\tau}{p}.$$

□

Substitute the estimate for e_k in (30):

$$r_{k+1} \leq (1-\gamma a)\, r_k + \frac{8\gamma^3 AB\tau}{p} + \gamma^2 C.$$

Running the recursion from 0 to K gives

$$r_{K+1} \leq (1-\gamma a)^K r_0 + \frac{8\gamma^2 AB\tau}{ap} + \frac{\gamma C}{a} \leq \exp\left(-\gamma a K\right) r_0 + \frac{8\gamma^2 AB\tau}{ap} + \frac{\gamma C}{a}.$$

Finally, we need tuning of $\gamma \leq \frac{1}{d} = \frac{p}{120L\tau}$:

- If $\frac{1}{d} \geq \frac{\ln(\max\{2, a r_0 K/C\})}{aK}$ then $\gamma = \frac{\ln(\max\{2, a r_0 K/C\})}{aK}$ gives

$$\tilde{\mathcal{O}}\left(\exp\left(-\ln\left(\max\{2, a r_0 K/C\}\right)\right) r_0 + \frac{AB\tau}{a^3 p K^2} + \frac{C}{a^2 K}\right) \leq \tilde{\mathcal{O}}\left(\exp\left(-\frac{aK}{d}\right) r_0 + \frac{AB\tau}{a^3 p K^2} + \frac{C}{a^2 K}\right).$$

- If $\frac{1}{d} \leq \frac{\ln(\max\{2, a r_0 K/C\})}{aK}$ then $\gamma = \frac{1}{d}$ gives

$$\tilde{\mathcal{O}}\left(\exp\left(-\frac{aK}{d}\right) r_0 + \frac{AB\tau}{ad^2 p} + \frac{C}{ad}\right) \leq \tilde{\mathcal{O}}\left(\exp\left(-\frac{aK}{d}\right) r_0 + \frac{AB\tau}{a^3 p K^2} + \frac{C}{a^2 K}\right).$$

What in the end gives that

$$r_{k+1} = \tilde{\mathcal{O}}\left(\exp\left(-\frac{aK}{d}\right) r_0 + \frac{AB\tau}{a^3 p K^2} + \frac{C}{a^2 K}\right).$$

This completes the proof of the strongly-convex–strongly-concave case.

□

C.3 Proof of Theorem 1, monotone case

Note that in the proof of inequality (20) we can take an arbitrary z instead of z^*. Rearranging terms, we obtain for an arbitrary z:

$$2\gamma\mathbb{E}\left[\langle \bar{g}^{k+1/3}, \bar{z}^{k+1/3} - z\rangle\right] = \mathbb{E}\left[\|\bar{z}^k - z\|^2\right] - \mathbb{E}\left[\|\bar{z}^{k+1} - z\|^2\right] - \mathbb{E}\left[\|\bar{z}^{k+1/3} - \bar{z}^k\|^2\right] + \gamma^2\mathbb{E}\left[\|\bar{g}^{k+1/3} - \bar{g}^k\|^2\right]. \tag{33}$$

Next we need two bounds: a lower bound for the l.h.s. that relates it with the true operator F, and an upper bound for the last term in the r.h.s. that is given by Lemma 3.

The lower bound is given by the following Lemma.

Lemma 7 *Let the operator F satisfy Assumption 4. Then, for any fixed z we have*

$$\mathbb{E}\left[\langle \bar{g}^{k+1/3}, \bar{z}^{k+1/3} - z\rangle\right] \geq \mathbb{E}\left[\left\langle F(\bar{z}^{k+1/3}), \bar{z}^{k+1/3} - z\right\rangle\right] \tag{34}$$

$$-\frac{\gamma L^2}{2}\mathbb{E}\left[\|\bar{z}^{k+1/3} - \bar{z}^k\|^2\right] - \frac{1}{2\gamma}\mathbb{E}\text{Err}(k+1/3) - L\sqrt{\mathbb{E}\text{Err}(k+1/3)}\sqrt{\mathbb{E}\|\bar{z}^k - z\|^2} \tag{35}$$

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 $\mathbb{E}_{\xi^{k+1/3}}$ on vector $\xi^{k+1/3}$

$$\begin{aligned}
\mathbb{E}\left[\langle \bar{g}^{k+1/3}, \bar{z}^{k+1/3} - z\rangle\right] &= \mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^M \mathbb{E}_{\xi^{k+1/3}}[F_m(z_m^{k+1/3}, \xi_m^{k+1/3})], \bar{z}^{k+1/3} - z\right\rangle\right] \\
&\overset{(5)}{=} \mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^M F_m(z_m^{k+1/3}), \bar{z}^{k+1/3} - z\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^M F_m(\bar{z}^{k+1/3}), \bar{z}^{k+1/3} - z\right\rangle\right] \\
&\quad -\mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^M [F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})], \bar{z}^{k+1/3} - z\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle F(\bar{z}^{k+1/3}), \bar{z}^{k+1/3} - z\right\rangle\right] \\
&\quad -\mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^M [F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})], \bar{z}^{k+1/3} - z\right\rangle\right].
\end{aligned}$$

Next we estimate from below the last term in the r.h.s. Since, for any $\kappa > 0$, it is true that $-2\langle a, b\rangle \geq -\frac{1}{2\kappa}\|a\|^2 - \frac{\kappa}{2}\|b\|^2$, taking $\kappa = \gamma L^2$ and using the Cauchy-Schwarz, we obtain

$$\begin{aligned}
&-\mathbb{E}\left\langle \frac{1}{M}\sum_{m=1}^M [F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})], \bar{z}^{k+1/3} - \bar{z}^k + \bar{z}^k - z\right\rangle \\
\geq\; & -\frac{\gamma L^2}{2}\mathbb{E}\|\bar{z}^{k+1/3} - \bar{z}^k\|^2 - \frac{1}{2\gamma L^2}\mathbb{E}\left\|\frac{1}{M}\sum_{m=1}^M [F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})]\right\|^2 \\
& -\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M [F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})]\right\|\|\bar{z}^k - z\|\right] \\
\overset{(4)}{\geq}\; & -\frac{\gamma L^2}{2}\mathbb{E}\left[\|\bar{z}^{k+1/3} - \bar{z}^k\|^2\right] - \frac{L^2}{2M\gamma L^2}\mathbb{E}\left[\sum_{m=1}^M \left\|\bar{z}^{k+1/3} - z_m^{k+1/3}\right\|^2\right] \\
& -\frac{L}{M}\mathbb{E}\left[\sum_{m=1}^M \left\|\bar{z}^{k+1/3} - z_m^{k+1/3}\right\|\|\bar{z}^k - z\|\right] \\
\overset{(19)}{\geq}\; & -\frac{\gamma L^2}{2}\mathbb{E}\left[\|\bar{z}^{k+1/3} - \bar{z}^k\|^2\right] - \frac{1}{2\gamma}\mathbb{E}\text{Err}(k+1/3) - L\sqrt{\mathbb{E}\text{Err}(k+1/3)}\sqrt{\mathbb{E}\|\bar{z}^k - z\|^2},
\end{aligned}$$

where in the last inequality we used also that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M}\left\|\bar{z}^{k+1/3}-z_m^{k+1/3}\right\|\|\bar{z}^k-z\|\right] &\le \sqrt{\mathbb{E}\left(\frac{1}{M}\sum_{m=1}^{M}\left\|\bar{z}^{k+1/3}-z_m^{k+1/3}\right\|\right)^2}\sqrt{\mathbb{E}\|\bar{z}^k-z\|^2}\\ &\le \sqrt{\frac{1}{M}\mathbb{E}\sum_{m=1}^{M}\left\|\bar{z}^{k+1/3}-z_m^{k+1/3}\right\|^2}\sqrt{\mathbb{E}\|\bar{z}^k-z\|^2}\\ &\overset{(19)}{=} \sqrt{\mathbb{E}\mathrm{Err}(k+1/3)}\sqrt{\mathbb{E}\|\bar{z}^k-z\|^2}\end{aligned}$$

Combining the above, we obtain the statement of the Lemma.

□

Combining inequality (33) with Lemma 3 and Lemma 7, rearranging the terms, and using the monotonicity of the operator F, i.e. (SM) with $\mu = 0$, we obtain, for any z

$$\begin{aligned}2\gamma\mathbb{E}\left[\left\langle F(z),\bar{z}^{k+1/3}-z\right\rangle\right] &\le 2\gamma\mathbb{E}\left[\left\langle F(\bar{z}^{k+1/3}),\bar{z}^{k+1/3}-z\right\rangle\right]\\ &\le \mathbb{E}\left[\|\bar{z}^k-z\|^2\right]-\mathbb{E}\left[\|\bar{z}^{k+1}-z\|^2\right]\\ &\quad -\mathbb{E}\left[\|\bar{z}^{k+1/3}-\bar{z}^k\|^2\right]+5\gamma^2L^2\mathbb{E}\left[\|\bar{z}^{k+1/3}-\bar{z}^k\|^2\right]\\ &\quad +\frac{10\sigma^2\gamma^2}{M}+5L^2\gamma^2\mathbb{E}\left[\mathrm{Err}(k+1/3)\right]+5L^2\gamma^2\mathbb{E}\left[\mathrm{Err}(k)\right]\\ &\quad +L^2\gamma^2\mathbb{E}\left[\|\bar{z}^{k+1/3}-\bar{z}^k\|^2\right]+\mathrm{Err}(k+1/3)\\ &\quad +2\gamma L\sqrt{\mathbb{E}\mathrm{Err}(k+1/3)}\sqrt{\mathbb{E}\|\bar{z}^k-z\|^2}\\ &\le \mathbb{E}\left[\|\bar{z}^k-z\|^2\right]-\mathbb{E}\left[\|\bar{z}^{k+1}-z\|^2\right]+\frac{10\sigma^2\gamma^2}{M}\\ &\quad +5L^2\gamma^2\mathbb{E}\left[\mathrm{Err}(k+1/3)\right]+5L^2\gamma^2\mathbb{E}\left[\mathrm{Err}(k)\right]\\ &\quad +\mathrm{Err}(k+1/3)+2\gamma L\sqrt{\mathbb{E}\mathrm{Err}(k+1/3)}\sqrt{\mathbb{E}\|\bar{z}^k-z\|^2}\\ &\le \mathbb{E}\left[\|\bar{z}^k-z\|^2\right]-\mathbb{E}\left[\|\bar{z}^{k+1}-z\|^2\right]+\frac{10\sigma^2\gamma^2}{M}+(1+5\gamma^2L^2)\mathbb{E}\left[\mathrm{Err}(k+1/3)\right]\\ &\quad +5\gamma^2L^2\mathbb{E}\left[\mathrm{Err}(k)\right]+2\gamma L\sqrt{\mathbb{E}\mathrm{Err}(k+1/3)}\sqrt{\mathbb{E}\|\bar{z}^k-z\|^2},\end{aligned} \tag{36}$$

where in the last but one inequality we used the choice $\gamma \le \frac{1}{L\sqrt{6}}$. Further, by Lemma 6, we have, for any z,

$$\begin{aligned}2\gamma\mathbb{E}\left[\left\langle F(z),\bar{z}^{k+1/3}-z\right\rangle\right] &\le \mathbb{E}\left[\|\bar{z}^k-z\|^2\right]-\mathbb{E}\left[\|\bar{z}^{k+1}-z\|^2\right]\\ &\quad +\frac{10\sigma^2\gamma^2}{M}+(1+5\gamma^2L^2)\cdot\frac{8\gamma^2\tau}{p}\cdot\left(\frac{225D^2\tau}{p}+25\sigma^2\right)\\ &\quad +5\gamma^2L^2\cdot\frac{8\gamma^2\tau}{p}\cdot\left(\frac{225D^2\tau}{p}+25\sigma^2\right)\\ &\quad +2\gamma L\sqrt{\frac{8\gamma^2\tau}{p}\cdot\left(\frac{225D^2\tau}{p}+25\sigma^2\right)}\sqrt{\mathbb{E}\|\bar{z}^k-z\|^2}\\ &\le \mathbb{E}\left[\|\bar{z}^k-z\|^2\right]-\mathbb{E}\left[\|\bar{z}^{k+1}-z\|^2\right]+\frac{10\sigma^2\gamma^2}{M}\\ &\quad +(1+10\gamma^2L^2)\cdot\frac{8\gamma^2\tau}{p}\cdot\left(\frac{225D^2\tau}{p}+25\sigma^2\right)\\ &\quad +\gamma L\sqrt{\frac{32\gamma^2\tau}{p}\cdot\left(\frac{225D^2\tau}{p}+25\sigma^2\right)}\sqrt{\mathbb{E}\|\bar{z}^k-z\|^2}\\ &\le \mathbb{E}\left[\|\bar{z}^k-z\|^2\right]-\mathbb{E}\left[\|\bar{z}^{k+1}-z\|^2\right]+\xi\\ &\quad +\sqrt{\eta}\sqrt{\mathbb{E}\|\bar{z}^k-z\|^2},\end{aligned} \tag{37}$$

where we denote $\Delta := 32 \cdot \frac{\tau}{p} \cdot \left(\frac{225D^2\tau}{p} + 25\sigma^2\right)$, $\xi := (1+10\gamma^2L^2)\gamma^2\Delta + \frac{10\sigma^2\gamma^2}{M}$, $\eta = \gamma^4 L^2 \Delta$.

Unbounded iterates First, we consider the general case when the iterates $\bar{z}^k$ are not assumed to be bounded. We carefully analyze this sequence and prove that this sequence can not go too far from any solution to the variational inequality. This allows us to obtain the final convergence rate bound. Let us denote $r_k(z) := \sqrt{\mathbb{E}\|\bar{z}^k - z\|^2}$ and let z^* be a solution to the variational inequality. Then, we have

$$\begin{aligned} r_k(z) &\leq \sqrt{2\mathbb{E}\|\bar{z}^k - z^*\|^2 + 2\|z - z^*\|^2} \leq \sqrt{2\mathbb{E}\|\bar{z}^k - z^*\|^2} + \sqrt{2\|z - z^*\|^2} \\ &= \sqrt{2} r_k(z^*) + \sqrt{2}\|z - z^*\|, \\ (r_k(z))^2 &\leq 2\mathbb{E}\|\bar{z}^k - z^*\|^2 + \|z - z^*\|^2 = 2(r_k(z^*))^2 + 2\|z - z^*\|^2. \end{aligned}$$

Thus, from (37), we have, for any z and any $k \geq 0$,

$$2\gamma\mathbb{E}\left[\left\langle F(z), \bar{z}^{k+1/3} - z\right\rangle\right] \leq r_k(z)^2 - r_{k+1}(z)^2 + \xi + \sqrt{\eta} r_k(z), \tag{38}$$

and summing these inequalities from $k = 0$ to K, we obtain, for any z,

$$\begin{aligned} 2\gamma(K+1)\mathbb{E}\left[\langle F(z), \hat{z}^K - z\rangle\right] &\leq r_0(z)^2 + (K+1)\xi + \sqrt{\eta}\sum_{k=0}^{K} r_k(z) \\ &\leq 2r_0(z^*)^2 + 2\|z - z^*\|^2 + (K+1)\xi \\ &\quad + \sqrt{\eta}\left(\sqrt{2}(K+1)\|z - z^*\| + \sqrt{2}\sum_{k=0}^{K} r_k(z^*)\right), \end{aligned} \tag{39}$$

where $\hat{z}^K = \frac{1}{K+1}\sum\limits_{k=0}^{K} \bar{z}^{k+1/3}$.

Our next goal is to bound from above

$$r_0(z^*)^2 + (K+1)\xi + \sqrt{2\eta}\sum_{k=0}^{K} r_k(z^*).$$

Taking $z = z^*$ in (38) and using the fact that z^* is a solution to the variational inequality, we obtain, for any $k \geq 0$

$$0 \leq 2\gamma\mathbb{E}\left[\left\langle F(z^*), \bar{z}^{k+1/3} - z^*\right\rangle\right] \leq r_k(z^*)^2 - r_{k+1}(z^*)^2 + \xi + \sqrt{\eta} r_k(z^*).$$

Thus, for all $k \geq 0$,

$$r_{k+1}(z^*)^2 \leq r_k(z^*)^2 + \xi + \sqrt{\eta} r_k(z^*).$$

Summing these inequalities from $k = 0$ to K, we obtain

$$r_{K+1}(z^*)^2 \leq r_0(z^*)^2 + (K+1)\xi + \sqrt{\eta}\sum_{k=0}^{K} r_k(z^*).$$

Note that this inequality holds for arbitrary $K \geq 0$. We next use the following

Lemma 8 (Lemma B.2 in (Gorbunov, Dvurechensky, and Gasnikov 2018)) *Let $\alpha, a_0, \ldots, a_{N-1}, b, R_1, \ldots, R_{N-1}$ be non-negative numbers and*

$$R_l \leqslant \sqrt{2} \cdot \sqrt{\left(\sum_{k=0}^{l-1} a_k + b\alpha\sum_{k=1}^{l-1} R_k\right)} \quad l = 1, \ldots, N.$$

Then, for $l = 1, \ldots, N$,

$$\sum_{k=0}^{l-1} a_k + b\alpha\sum_{k=1}^{l-1} R_k \leqslant \left(\sqrt{\sum_{k=0}^{l-1} a_k} + \sqrt{2} b\alpha l\right)^2.$$

Choosing $\alpha = 1$, $b = \sqrt{\eta}$, $a_0 = r_0(z^*)^2 + \xi$, $a_k = \xi$, $k = 1, ..., K-1$, $R_k = r_k(z^*)$, we obtain

$$\begin{aligned} r_0(z^*)^2 + (K+1)\xi + \sqrt{\eta}\sum_{k=0}^{K} r_k(z^*) &\le \left(\sqrt{r_0(z^*)^2 + (K+1)\xi} + (K+1)\sqrt{2\eta}\right)^2 \\ &\le 2r_0(z^*)^2 + 2(K+1)\xi + 4(K+1)^2\eta \end{aligned}$$

Combining the last inequality with (39), we obtain

$$\begin{aligned} 2\gamma(K+1)\mathbb{E}\left[\left\langle F(z), \widehat{z}^K - z\right\rangle\right] &\le r_0(z^*)^2 + 2\|z - z^*\|^2 + \sqrt{2\eta}(K+1)\|z - z^*\| \\ &\quad + \left(2r_0(z^*)^2 + 2(K+1)\xi + 4(K+1)^2\eta\right) \\ &\le 3r_0(z^*)^2 + 2\|z - z^*\|^2 + 2(K+1)\xi \\ &\quad + \sqrt{2\eta}(K+1)\|z - z^*\| + 6(K+1)^2\eta, \end{aligned}$$

Dividing both sides of the inequality by $2\gamma(K+1)$ and using the definitions $\Delta := 32 \cdot \frac{\tau}{p} \cdot \left(\frac{225D^2\tau}{p} + 25\sigma^2\right)$, $\xi := (1 + 10\gamma^2L^2)\gamma^2\Delta + \frac{10\sigma^2\gamma^2}{M}$, $\eta = \gamma^4L^2\Delta$, we obtain, for all $z \in \mathcal{C}$

$$\begin{aligned} \mathbb{E}\left[\left\langle F(z), \widehat{z}^K - z\right\rangle\right] &\le 2\frac{\|z^0 - z^*\|^2 + \|z - z^*\|^2}{\gamma(K+1)} + \frac{\xi}{\gamma} + \|z - z^*\|\sqrt{\frac{\eta}{2\gamma^2}} + 3(K+1)\frac{\eta}{\gamma} \\ &\le 2\frac{\|z^0 - z^*\|^2 + \|z - z^*\|^2}{\gamma(K+1)} + \frac{10\sigma^2\gamma}{M} + (1 + 10\gamma^2L^2)\gamma\Delta \\ &\quad + \gamma L\|z - z^*\|\sqrt{\Delta} + 3(K+1)\gamma^3L^2\Delta \\ &\le \frac{4\Omega_{\mathcal{C}}^2}{\gamma(K+1)} + \frac{10\sigma^2\gamma}{M} + \gamma\Delta \\ &\quad + \gamma L\Omega_{\mathcal{C}}\sqrt{\Delta} + 8(K+1)\gamma^3L^2\Delta, \end{aligned}$$

where in the last inequality we used that $z^0, z, z^* \in \mathcal{C}$ and $\max_{z,z' \in \mathcal{C}} \|z - z'\| \le \Omega_{\mathcal{C}}$ and that $K \ge 1$.

Choosing

$$\gamma = \min\left\{\frac{1}{3L}, \left(\frac{2\Omega_{\mathcal{C}}^2 M}{5(K+1)\sigma^2}\right)^{\frac{1}{2}}, \left(\frac{\Omega_{\mathcal{C}}^2}{6(K+1)^2L^2\Delta}\right)^{\frac{1}{4}}\right\},$$

which implies

$$\frac{4\Omega_{\mathcal{C}}^2}{\gamma(K+1)} = \mathcal{O}\left(\frac{L\Omega_{\mathcal{C}}^2}{K} + \frac{\sigma\Omega_{\mathcal{C}}}{\sqrt{MK}} + \frac{\sqrt{L\Omega_{\mathcal{C}}^3\sqrt{\Delta}}}{\sqrt{K}}\right),$$

we obtain

$$\sup_{z \in \mathcal{C}} \mathbb{E}\left[\left\langle F(z), \widehat{z}^K - z\right\rangle\right] = \mathcal{O}\left(\frac{L\Omega_{\mathcal{C}}^2}{K} + \frac{\sigma\Omega_{\mathcal{C}}}{\sqrt{MK}} + \frac{\sqrt{L\Omega_{\mathcal{C}}^3\sqrt{\Delta}}}{\sqrt{K}} + \sqrt{\frac{(\Delta + L^2\Omega_{\mathcal{C}}^2)\Omega_{\mathcal{C}}\sqrt{\Delta}}{KL}}\right).$$

Bounded iterates Let us now consider the situation under the additional assumption that for all k the iterations of the algorithm satisfy $\|\bar{z}^k\| \le \Omega$. In this case, summing (37) from $k = 0$ to K, we obtain, for any z,

$$\begin{aligned} 2\gamma(K+1)\mathbb{E}\left[\left\langle F(z), \widehat{z}^K - z\right\rangle\right] &\le \|z^0 - z\|^2 + (K+1)\xi + \sqrt{\eta}\sum_{k=0}^{K}\sqrt{\mathbb{E}\|\bar{z}^k - z\|^2} \\ &\le \|z^0 - z\|^2 + (K+1)\xi + 2(K+1)\sqrt{\eta}(\Omega + \|z\|). \end{aligned}$$

Dividing both sides of this inequality by $2\gamma(K+1)$ and using the definitions $\Delta := 32 \cdot \frac{\tau}{p} \cdot \left(\frac{225D^2\tau}{p} + 25\sigma^2\right)$, $\xi := (1 + 10\gamma^2L^2)\gamma^2\Delta + \frac{10\sigma^2\gamma^2}{M}$, $\eta = \gamma^4L^2\Delta$, we obtain, for all $z \in \mathcal{C}$

$$
\begin{aligned}
\mathbb{E}\left[\langle F(z), \widehat{z}^K - z\rangle\right] &\leq \frac{\|z^0 - z\|^2}{2\gamma(K+1)} + \frac{\xi}{\gamma} + (\Omega + \|z\|)\sqrt{\frac{\eta}{\gamma^2}} \\
&\leq \frac{\|z^0 - z\|^2}{2\gamma(K+1)} + \frac{10\sigma^2\gamma}{M} + (1 + 10\gamma^2L^2)\gamma\Delta \\
&\quad + (\Omega + \|z\|)\gamma L\sqrt{\Delta} \\
&\leq \frac{\Omega_{\mathcal{C}}^2}{2\gamma(K+1)} + \frac{10\sigma^2\gamma}{M} + 10\gamma^3L^2\Delta \\
&\quad + \gamma((\Omega + \Omega_{\mathcal{C}})L\sqrt{\Delta} + \Delta),
\end{aligned}
$$

where in the last inequality we used that $z^0, z, z^* \in \mathcal{C}$ and $\max_{z,z'\in\mathcal{C}} \|z - z'\| \leq \Omega_{\mathcal{C}}$ and that $K \geq 1$.

Similar to the above case, choosing

$$
\gamma = \min\left\{\frac{1}{3L}, \left(\frac{\Omega_{\mathcal{C}}^2 M}{20(K+1)\sigma^2}\right)^{\frac{1}{2}}, \left(\frac{\Omega_{\mathcal{C}}^2}{60(K+1)^2L^2\Delta}\right)^{\frac{1}{4}}, \left(\frac{\Omega_{\mathcal{C}}^2}{(K+1)((\Omega + \Omega_{\mathcal{C}})L\sqrt{\Delta} + \Delta)}\right)^{\frac{1}{2}}\right\},
$$

we obtain

$$
\sup_{z\in\mathcal{C}} \mathbb{E}\left[\langle F(z), \widehat{z}^K - z\rangle\right] = \mathcal{O}\left(\frac{L\Omega_{\mathcal{C}}^2}{K} + \frac{\sigma\Omega_{\mathcal{C}}}{\sqrt{MK}} + \frac{\sqrt{L\Omega_{\mathcal{C}}^3\sqrt{\Delta}}}{K^{3/4}} + \sqrt{\frac{((\Omega + \Omega_{\mathcal{C}})L\sqrt{\Delta} + \Delta)\Omega_{\mathcal{C}}^2}{K}}\right). \tag{40}
$$

□

C.4 Proof of Theorem 1, non-monotone case

The proof starts very similar to the strongly-monotone case. In particular, we can get (20). Lemma 3 does not need modification, but we will change Lemma 2:

Lemma 9 *Under Assumptions 2, 4 it holds:*

$$
\begin{aligned}
-2\gamma\mathbb{E}\left[\langle \bar{g}^{k+1/3}, \bar{z}^{k+1/3} - z^*\rangle\right] &\leq 2\gamma L\sqrt{\mathbb{E}\left[\|\bar{z}^{k+1/3} - z^*\|^2\right]}\sqrt{\mathbb{E}\left[Err(k+1/3)\right]} \\
&\quad + \gamma L\mathbb{E}\left[\|\bar{z}^{k+1/3} - \bar{z}^k\|^2\right] + \gamma L\mathbb{E}\left[Err(k+1/3)\right].
\end{aligned} \tag{41}
$$

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 $\mathbb{E}_{\xi^{k+1/3}}$ on vector $\xi^{k+1/3}$ and get the following chain of inequalities:

$$
\begin{aligned}
-2\gamma\mathbb{E}\left[\langle \bar{g}^{k+1/3}, \bar{z}^{k+1/3} - z^*\rangle\right] &= -2\gamma\mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{\xi^{k+1/3}}[F_m(z_m^{k+1/3}, \xi_m^{k+1/3})], \bar{z}^{k+1/3} - z^*\right\rangle\right] \\
&\overset{(5)}{=} -2\gamma\mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M} F_m(z_m^{k+1/3}), \bar{z}^{k+1/3} - z^*\right\rangle\right] \\
&= -2\gamma\mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M} F_m(\bar{z}^{k+1/3}), \bar{z}^{k+1/3} - z^*\right\rangle\right] \\
&\quad + 2\gamma\mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M} [F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})], \bar{z}^{k+1/3} - z^*\right\rangle\right] \\
&= -2\gamma\mathbb{E}\left[\left\langle F(\bar{z}^{k+1/3}), \bar{z}^{k+1/3} - z^*\right\rangle\right] \\
&\quad + 2\gamma\mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M} [F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})], \bar{z}^{k+1/3} - z^*\right\rangle\right] \\
&\overset{\text{(NM)}}{\leq} 2\gamma\mathbb{E}\left[\left\langle \frac{1}{M}\sum_{m=1}^{M} [F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})], \bar{z}^{k+1/3} - z^*\right\rangle\right] \\
&\leq 2\gamma\mathbb{E}\left[\|\bar{z}^{k+1/3} - z^*\| \cdot \left\|\frac{1}{M}\sum_{m=1}^{M} F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})\right\|\right] \\
&\leq 2\gamma\mathbb{E}\left[\|\bar{z}^{k+1/3} - z^*\| \cdot \frac{1}{M}\sum_{m=1}^{M} \left\|F_m(\bar{z}^{k+1/3}) - F_m(z_m^{k+1/3})\right\|\right] \\
&\overset{(4)}{\leq} 2\gamma L\mathbb{E}\left[\|\bar{z}^{k+1/3} - z^*\| \cdot \frac{1}{M}\sum_{m=1}^{M} \left\|z_m^{k+1/3} - \bar{z}^{k+1/3}\right\|\right] \\
&\leq 2\gamma L\mathbb{E}\left[\|\bar{z}^{k} - z^*\| \cdot \frac{1}{M}\sum_{m=1}^{M} \left\|z_m^{k+1/3} - \bar{z}^{k+1/3}\right\|\right] \\
&\quad + 2\gamma L\mathbb{E}\left[\|\bar{z}^{k+1/3} - \bar{z}^{k}\| \cdot \frac{1}{M}\sum_{m=1}^{M} \left\|z_m^{k+1/3} - \bar{z}^{k+1/3}\right\|\right] \\
&\overset{(15),(13)}{\leq} 2\gamma L\sqrt{\mathbb{E}\left[\|\bar{z}^{k} - z^*\|^2\right]} \cdot \sqrt{\mathbb{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M} \left\|z_m^{k+1/3} - \bar{z}^{k+1/3}\right\|\right)^2\right]} \\
&\quad + \gamma L\mathbb{E}\left[\|\bar{z}^{k+1/3} - \bar{z}^{k}\|^2\right] + \gamma L\mathbb{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M} \|\bar{z}^{k+1/3} - z_m^{k+1/3}\|\right)^2\right].
\end{aligned}
$$

By (12) it is easy to see that

$$
\mathbb{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M} \|\bar{z}^{k+1/3} - z_m^{k+1/3}\|\right)^2\right] \leq \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} \|\bar{z}^{k+1/3} - z_m^{k+1/3}\|^2\right].
$$

This completes the proof.

□

As a result, we have an analogue of (23):

$$\mathbb{E}\left[\|\bar{z}^{k+1}-z^*\|^2\right] \leq \mathbb{E}\left[\|\bar{z}^{k}-z^*\|^2\right] - \mathbb{E}\left[\|\bar{z}^{k+1/3}-\bar{z}^{k}\|^2\right]$$
$$+ 2\gamma L\sqrt{\mathbb{E}\left[\|\bar{z}^{k}-z^*\|^2\right]}\sqrt{\mathbb{E}\left[\text{Err}(k+1/3)\right]}$$
$$+ \gamma L\mathbb{E}\left[\|\bar{z}^{k+1/3}-\bar{z}^{k}\|^2\right] + \gamma L\mathbb{E}\left[\text{Err}(k+1/3)\right]$$
$$+ \gamma^2\left(5L^2\mathbb{E}\left[\|\bar{z}^{k+1/3}-\bar{z}^{k}\|^2\right] + \frac{10\sigma^2}{M} + 5L^2\mathbb{E}\left[\text{Err}(k+1/3)\right] + 5L^2\mathbb{E}\left[\text{Err}(k)\right]\right).$$

Choosing $\gamma \leq \frac{1}{5L}$ gives

$$\frac{1}{2}\mathbb{E}\left[\|\bar{z}^{k+1/3}-\bar{z}^{k}\|^2\right] \leq \mathbb{E}\left[\|\bar{z}^{k}-z^*\|^2\right] - \mathbb{E}\left[\|\bar{z}^{k+1}-z^*\|^2\right]$$
$$+ 2\gamma L\sqrt{\mathbb{E}\left[\|\bar{z}^{k}-z^*\|^2\right]}\sqrt{\mathbb{E}\left[\text{Err}(k+1/3)\right]}$$
$$+ (5\gamma^2L^2+\gamma L)\mathbb{E}\left[\text{Err}(k+1/3)\right] + 5\gamma^2L^2\mathbb{E}\left[\text{Err}(k)\right] + \frac{10\gamma^2\sigma^2}{M}.$$

Next we work with

$$\mathbb{E}\left[\|\bar{z}^{k+1/3}-\bar{z}^{k}\|^2\right]$$
$$= \gamma^2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} F_m(z_m^k,\xi_m^k) - F_m(z_m^k) + F_m(z_m^k) - F_m(\bar{z}^k) + F_m(\bar{z}^k)\right\|^2\right]$$
$$\overset{(14)}{\geq} \frac{\gamma^2}{2}\mathbb{E}\left\|F(\bar{z}^k)\right\|^2 - \gamma^2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} F_m(z_m^k,\xi_m^k) - F_m(z_m^k) + F_m(z_m^k) - F_m(\bar{z}^k)\right\|^2\right]$$
$$\overset{(14)}{\geq} \frac{\gamma^2}{2}\mathbb{E}\left\|F(\bar{z}^k)\right\|^2 - 2\gamma^2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} F_m(z_m^k,\xi_m^k) - F_m(z_m^k)\right\|^2\right] - 2\gamma^2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} F_m(z_m^k) - F_m(\bar{z}^k)\right\|^2\right]$$
$$\overset{(4)}{\geq} \frac{\gamma^2}{2}\mathbb{E}\left\|F(\bar{z}^k)\right\|^2 - \frac{2\gamma^2\sigma^2}{M} - \frac{2\gamma^2L^2}{M}\sum_{m=1}^{M}\mathbb{E}\left[\left\|z_m^k-\bar{z}^k\right\|^2\right]$$
$$= \frac{\gamma^2}{2}\mathbb{E}\left\|F(\bar{z}^k)\right\|^2 - \frac{2\gamma^2\sigma^2}{M} - 2\gamma^2L^2\mathbb{E}\left[\text{Err}(k)\right].$$

Connecting with previous gives

$$\frac{\gamma^2}{4}\mathbb{E}\left[\|F(\bar{z}^k)\|^2\right] \leq \mathbb{E}\left[\|\bar{z}^{k}-z^*\|^2\right] - \mathbb{E}\left[\|\bar{z}^{k+1}-z^*\|^2\right]$$
$$+ 2\gamma L\sqrt{\mathbb{E}\left[\|\bar{z}^{k}-z^*\|^2\right]}\sqrt{\mathbb{E}\left[\text{Err}(k+1/3)\right]}$$
$$+ (\gamma L+5\gamma^2L^2)\mathbb{E}\left[\text{Err}(k+1/3)\right] + 6\gamma^2L^2\mathbb{E}\left[\text{Err}(k)\right] + \frac{11\gamma^2\sigma^2}{M}.$$

With result of Lemma 6 we get

$$\frac{\gamma^2}{4}\mathbb{E}\left[\|F(\bar{z}^k)\|^2\right] \leq \mathbb{E}\left[\|\bar{z}^{k}-z^*\|^2\right] - \mathbb{E}\left[\|\bar{z}^{k+1}-z^*\|^2\right]$$
$$+ 2\gamma L\sqrt{\mathbb{E}\left[\|\bar{z}^{k}-z^*\|^2\right]}\sqrt{\frac{8\gamma^2\tau}{p}\cdot\left(\frac{225D^2\tau}{p}+25\sigma^2\right)}$$
$$+ \gamma^2\left(\frac{11\sigma^2}{M} + \frac{8(\gamma L+11\gamma^2L^2)\tau}{p}\cdot\left(\frac{225D^2\tau}{p}+25\sigma^2\right)\right).$$

Summing over all k from 0 to K and averaging gives:

$$\mathbb{E}\left[\frac{1}{K+1}\sum_{k=0}^{K}\|F(\bar{z}^k)\|^2\right] \leq \frac{4\|z^0-z^*\|^2}{\gamma^2(K+1)} - \frac{4\mathbb{E}\left[\|z^{K+1}-z^*\|^2\right]}{\gamma^2(K+1)} + \frac{44\sigma^2}{M}$$
$$+ \sqrt{\frac{32L^2\tau}{p}\cdot\left(\frac{225D^2\tau}{p}+25\sigma^2\right)}\cdot\frac{1}{K+1}\sum_{k=0}^{K}\sqrt{\mathbb{E}\left[\|\bar{z}^k-z^*\|^2\right]}$$
$$+ \frac{8(\gamma L+11\gamma^2L^2)\tau}{p}\cdot\left(\frac{225D^2\tau}{p}+25\sigma^2\right). \tag{42}$$

Unbounded iterates We rewrite (42) as follows

$$\mathbb{E}\left[\|\bar{z}^{K+1}-z^*\|^2\right] \leq \|z^0-z^*\|^2 + \frac{11\gamma^2(K+1)\sigma^2}{M}$$
$$+ \sqrt{\frac{2\gamma^4L^2\tau}{p}\cdot\left(\frac{225D^2\tau}{p}+25\sigma^2\right)}\cdot\sum_{k=0}^{K}\sqrt{\mathbb{E}\left[\|\bar{z}^k-z^*\|^2\right]}$$
$$+ \frac{\gamma^2(K+1)(\gamma L+11\gamma^2L^2)\tau}{2p}\cdot\left(\frac{225D^2\tau}{p}+25\sigma^2\right).$$

Then we can use Lemma 8 with $R_k = \sqrt{\mathbb{E}\left[\|\bar{z}^k-z^*\|^2\right]}$, $b = \sqrt{\frac{2\gamma^4L^2\tau}{p}\cdot\left(\frac{225D^2\tau}{p}+25\sigma^2\right)}$, for $k\geq 1$ $a_k = \frac{\gamma^2(\gamma L+11\gamma^2L^2)\tau}{2p}\cdot\left(\frac{225D^2\tau}{p}+25\sigma^2\right)+\frac{11\gamma^2\sigma^2}{M}$, $a_0 = \|z^0-z^*\|^2 + \frac{\gamma^2(\gamma L+11\gamma^2L^2)\tau}{2p}\cdot\left(\frac{225D^2\tau}{p}+25\sigma^2\right)+\frac{11\gamma^2\sigma^2}{M}$ and get

$$\sum_{k=0}^{K}a_k + b\sum_{k=1}^{K}R_k \leq \left(\sqrt{\sum_{k=0}^{K}a_k}+\sqrt{2}b(K+1)\right)^2 \leq 2\sum_{k=0}^{K}a_k + 4b^2(K+1)^2,$$

which gives

$$\sum_{k=1}^{K}R_k \leq \frac{1}{b}\sum_{k=0}^{K}a_k + 4b(K+1)^2.$$

Substituting this in (42) with the same notation, we have

$$\frac{\gamma^2(K+1)}{4}\mathbb{E}\left[\frac{1}{K+1}\sum_{k=0}^{K}\|F(\bar{z}^k)\|^2\right] \leq \sum_{k=0}^{K}a_k + b\left(\frac{1}{b}\sum_{k=0}^{K}a_k + 4b(K+1)^2\right).$$

and

$$\mathbb{E}\left[\frac{1}{K+1}\sum_{k=0}^{K}\|F(\bar{z}^k)\|^2\right] \leq \frac{8}{\gamma^2(K+1)}\sum_{k=0}^{K}a_k + \frac{16b^2(K+1)}{\gamma^2}.$$

Finally, we get

$$\mathbb{E}\left[\frac{1}{K+1}\sum_{k=0}^{K}\|F(\bar{z}^k)\|^2\right] = \mathcal{O}\Bigg(\frac{\|z^0-z^*\|^2}{\gamma^2(K+1)} + \frac{(\gamma L+\gamma^2L^2)\tau}{p}\cdot\left(\frac{D^2\tau}{p}+\sigma^2\right)$$
$$+ \frac{\sigma^2}{M} + \frac{(K+1)\gamma^2L^2\tau}{p}\cdot\left(\frac{D^2\tau}{p}+\sigma^2\right)\Bigg).$$

As before, we denote $\Delta := 32\cdot\frac{\tau}{p}\cdot\left(\frac{225D^2\tau}{p}+25\sigma^2\right)$. Choosing $\gamma = \min\left\{\frac{1}{4L}; \left(\frac{\|z^0-z^*\|^2}{(K+1)^2L^2\Delta}\right)^{1/4}\right\}$, we obtain

$$\mathbb{E}\left[\frac{1}{K+1}\sum_{k=0}^{K}\|F(\bar{z}^k)\|^2\right] = \mathcal{O}\Bigg(\frac{L^2\|z^0-z^*\|^2}{K} + L\|z^0-z^*\|\sqrt{\Delta}$$
$$+ \frac{\sigma^2}{M} + \frac{\sqrt{L\|z^0-z^*\|\Delta^{3/4}}}{\sqrt{K}}\Bigg),$$

which completes the proof of (10).

Bounded iterates Under the additional assumption that $\|z^*\| \leq \Omega$ and $\|\bar{z}^k\| \leq \Omega$, from (42), we obtain

$$\mathbb{E}\left[\frac{1}{K+1}\sum_{k=0}^{K}\|F(\bar{z}^k)\|^2\right] = \mathcal{O}\Bigg(\frac{\|z^0 - z^*\|^2}{\gamma^2(K+1)} + \frac{(\gamma L + \gamma^2 L^2)\tau}{p}\cdot\left(\frac{D^2\tau}{p} + \sigma^2\right) + \frac{\sigma^2}{M} + \sqrt{\frac{L^2\Omega^2\tau}{p}\cdot\left(\frac{D^2\tau}{p} + \sigma^2\right)}\Bigg).$$

With $\gamma = \min\left\{\frac{1}{4L}; \left(\frac{\Omega^2}{(K+1)L\Delta}\right)^{1/3}\right\}$ we have

$$\mathbb{E}\left[\frac{1}{K+1}\sum_{k=0}^{K}\|F(\bar{z}^k)\|^2\right] = \mathcal{O}\left(\frac{L^2\Omega^2}{K} + \frac{\sigma^2}{M} + \frac{(L\Omega\Delta)^{2/3}}{K^{1/3}} + L\Omega\sqrt{\Delta}\right).$$

□