Decentralized saddle-point problems with different constants of strong convexity and strong concavity

Dmitry Metelev1 · Alexander Rogozin1,2 · Alexander Gasnikov1,2,3 · Dmitry Kovalev4

Received: 30 June 2023 / Accepted: 13 October 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract
Large-scale saddle-point problems arise in such machine learning tasks as GANs and linear models with affine constraints. In this paper, we study distributed saddle-point problems with strongly-convex–strongly-concave smooth objectives that have different strong convexity and strong concavity parameters of composite terms, which correspond to min and max variables, and bilinear saddle-point part. We consider two types of first-order oracles: deterministic (returns gradient) and stochastic (returns unbiased stochastic gradient). Our method works in both cases and takes several consensus steps between oracle calls.

Keywords Decentralized optimization · Time-varying graphs · Saddle-point problem · Stochastic optimization · Consensus subroutine · Inexact oracle

1 Introduction
In this paper we study saddle-point problems (SPP) with two composite terms and a bilinear part

Dmitry Metelev
metelev.ds@phystech.edu

Alexander Rogozin
aleksandr.rogozin@phystech.edu

Alexander Gasnikov
gasnikov@yandex.ru

Dmitry Kovalev
dakovalev1@gmail.com

1 Moscow Institute of Physics and Technology, Moscow, Russia
2 HSE University, Moscow, Russia
3 ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia
4 King Abdullah University of Science and Technology, Abu Dhabi, United Arab Emirates

Published online: 06 November 2023
where function \( f \) is \( \mu_x \)-strongly convex (\( \mu_x > 0 \)) and \( L_x \)-smooth and function \( g \) is \( \mu_y \)-strongly convex (\( \mu_y > 0 \)) and \( L_y \)-smooth, \( A \) is a coupling matrix of dimensions \( d_y \times d_x \). Interest in this class of problems has grown in the last few years (Zhang et al. 2019; Alkousa et al. 2020; Lin et al. 2020; Wang and Li 2020; Tominin et al. 2021; Kovalev and Gasnikov 2022) due to the general growth of interest in SPP in ML (machine learning) community. The following lower complexity bound was obtained in Zhang et al. (2019):

\[
\mathcal{O} \left( \left( \sqrt{\frac{L_x}{\mu_x}} + \sqrt{\frac{\lambda_{\text{max}}(A^TA)}{\mu_x \mu_y}} + \sqrt{\frac{L_y}{\mu_y}} \right) \log \frac{1}{\epsilon} \right),
\]

where \( \lambda_{\text{max}}(A^TA) \) denotes the largest eigenvalue of the matrix \( A^TA \), and \( \epsilon \) is the accuracy at which we want to solve the optimization problem. Optimal methods that work according to the lower bound were independently obtained and almost simultaneously in Kovalev et al. (2021), Thekumparampil et al. (2022), Jin et al. (2022). For the case \( \mu_x = \mu_y, L_x = L_y \), the lower bound and optimal methods were known much earlier (Nemirovski and Yudin 1983; Tseng 2000; Lan 2020).

The case of the stochastic oracle, in which we have access to unbiased stochastic gradients of \( f \) and \( g \), has been studied less extensively in the literature. For example, in Zhang et al. (2021), a non-bilinear SPP was considered and composite terms \( f \), \( g \) were assumed to be proximal-friendly. We study a general case where \( f \) and \( g \) may not be proximal-friendly.

The problem (1) often arises in decentralized optimization (Boyd et al. 2011; Yarmoshik et al. 2022) and can be considered as a particular case (when \( g \equiv 0 \)) of decentralized convex optimization problems with affine constraints (Gorbunov et al. 2020). From these applications, it follows that we have \( L_y \gg \mu_x = \mu_y \sim \epsilon \), where \( \epsilon \) is the desired accuracy in the duality gap.\(^1\) The state-of-the-art results for decentralized SPP (1) proposed in papers (Beznosikov et al. 2020; Rogozin et al. 2021a; Kovalev et al. 2022; Luo and Ye 2022) (both for deterministic and stochastic oracles) require \( \mu_x = \mu_y \) and \( L_x = L_y \).

This paper describes a generalized method for transferring a set of non-distributed optimization algorithms to a decentralized setting. The idea of the method itself was first described in Rogozin et al. (2021a), but was applied only in the deterministic case. In this paper, we show that this method can also be applied to fairly complex stochastic algorithms, including those designed for SPP.

The contributions of this paper are summarized as follows.

\(^1\) Strictly speaking, we must set \( \mu_y = 0 \), since \( g \equiv 0 \). But without limiting generality, we can consider \( \mu_y \) to be as small as \( \epsilon \) due to the regularization technique (Wang and Li 2020; Rogozin et al. 2021a).
1. (Sensitivity) we generalize the optimal algorithm for (1) from Kovalev et al. (2021) to the case where oracle returns inexact gradients of $f$ and $g$ (Devolder et al. 2014);

2. (Stochasticity) by using sensitivity analysis and standard batch-technique (e.g. see Gasnikov et al. (2022)), we generalize (Kovalev et al. 2021) to the case where oracle returns stochastic gradients of $f$ and $g$ (with different variances);

3. (Decentralization) by using consensus-projection technique from Rogozin et al. (2021a), Rogozin et al. (2021b) and sensitivity analysis, we generalize stochastic version of (1) for decentralized setup.

Note that we could try to implement the above plan starting with an arbitrary optimal method from Kovalev et al. (2021), Thekumparampil et al. (2022), Jin et al. (2022) that has a complexity bound (2). However, we definitely prefer (Kovalev et al. 2021), because the results of Kovalev et al. (2021) also include the situation where $\mu_x = \mu_y = 0$, but we still have a linear rate of convergence (Ibrahim et al. 2020; Alkousa et al. 2020). This gives us a stochastic generalization in non-convex-non-concave setup. But at the same time, we can only get results for the strongly-convex-strongly-concave setup in decentralized case. This will be discussed in more detail below.

The drawbacks of the proposed approach are as follows:

- (Lack of overparametrization) Based on the proposed batch-technique we do not know how to replace the variance determined across the whole space to the variance determined only at the solution point, see Gorbunov et al. (2021), Beznosikov et al. (2022) for $\mu_x = \mu_y$, $L_x = L_y$ and non-distributed setup.

- (Extra logarithmic multiplier) Consensus-projection procedure leads to the addition of an extra logarithmic multiplier (on a desired accuracy) in comparison with direct approaches, which was clearly demonstrated in the case $\mu_x = \mu_y$, $L_x = L_y$ in Beznosikov et al. (2021) and Kovalev et al. (2022).

The advantages of the proposed approach are as follows:

- (Universality) The idea of proposing a general scheme that allows the construction of optimal decentralized stochastic methods based on non-distributed deterministic ones seems to be quite attractive (Gasnikov et al. 2022). To our knowledge, this has only been done for standard Nesterov’s accelerated (momentum) method (Rogozin et al. 2021b). The acceleration from Kovalev et al. (2021) is much more difficult. So the starting point of the plan (sensitivity analysis) required significant generalization of the results from Devolder et al. (2014), which was used in Rogozin et al. (2021b). The results obtained in the sensitivity

---

It seems that Kovalev et al. (2021) is the most tricky approach among (Kovalev et al. 2021; Thekumparampil et al. 2022; Jin et al. 2022), which uses significantly new ideas of acceleration unlike standard ones (Nesterov 2018; Lin et al. 2020).
part of this paper build a bridge to a much wider class of optimal modern non-distributed non-stochastic methods. It is crucial to note that the proposed scheme preserves optimality (up to a logarithmic factor) of the input method for the output one (see Rogozin et al. 2021a for the lower bound in the deterministic setup).

- (Average constants) As in the works (Rogozin et al. 2021a, b; Beznosikov et al. 2021) the complexity bounds (communication steps, oracle calls) are determined by the average (among all nodes) smoothness, strong convexity (concavity), and variance constants rather than the worst ones, which is typical for any other approach (that does not use a dual oracle) (Gorbunov et al. 2020).

- (Time-varying networks) As in the works (Rogozin et al. 2021a, b; Beznosikov et al. 2021), our results can be easily generalized to time-varying networks. This contrasts with the almost optimal loop-less approaches that are currently much trickier, as they have been specifically developed for certain optimization problems, rather than for general application (Kovalev et al. 2021a, b; Li and Lin 2021; Song et al. 2021).

An alternative approach that could be used is based on the decentralized Catalyst envelope (Tian et al. 2021). In this approach, we could construct an optimal method for SPP (1) based on a decentralized accelerated method for decentralized problem (Kovalev et al. 2020, 2021a, b; Li and Lin 2021; Song et al. 2021). We could try to build an optimal (up to a logarithmic factors) method in the same way as was done in the non-distributed setup (Lin et al. 2020; Tominin et al. 2021). Unfortunately, this approach: a) can only handle deterministic oracles; b) is characterized by worst-case constants (not the average ones); and c) leads to at least a third-degree logarithmic factor (Lin et al. 2020; Tominin et al. 2021), which is worse than in the previously described approach.  

Before conducting our research, we were not aware of papers that considered decentralized strongly-convex-strongly-concave SPP with different constants of strong convexity and strong concavity, even in deterministic case. The most competitive paper to ours is Kovalev et al. (2022) which considers equal constants of strong convexity and strong concavity. Therefore, the result of Kovalev et al. (2022) would contain $1 / \min \{\mu_x, \mu_y\}$, while our result is $1 / \sqrt{\mu_x \mu_y}$, which may be much better.

### 1.1 Decentralized setting

This paper primarily focuses on the decentralized saddle point problems. Specifically, we aim at solving

$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^b} \frac{1}{n} \sum_{i=1}^{n} f_i(x) + y^T A x - g_i(y),$$

(3)

---

3 Maybe this drawback could be eliminated over time, like it was done in Kovalev and Gasnikov (2022) for non-distributed setup.
where $f_i$ and $g_i$ are certain functions stored in the local memory of the corresponding agent $i$, we will refer to these functions as local functions. It is assumed that there are multiple computational agents, some of which have the ability to exchange information with each other, while others cannot. The set of agents and their ability to communicate are represented through a graph, with vertices representing the agents and edges indicating their ability to communicate with each other. We further assume that the network can change over time - meaning that some communication links can appear or disappear. The entire history of network changes is expressed as a sequence of graphs $G_k = (V, E_k)_{k=1}^\infty$, all with a common set of vertices $V$. We denote the number of vertices as $n = |V|$. Each agent stores its corresponding pair of functions $f_i$ and $g_i$ in its local memory, and has access to their values and gradients at any point. Furthermore, each agent can obtain the values $A^T x$ and $A^T y$ from arbitrary vectors $x$ and $y$. We assume that the local functions $f_i$ and $g_i$ are both strongly convex and smooth.

1.2 Basic definitions and assumptions

This section will provide the basic definitions and assumptions used later in the paper. These points refer mostly to the problem (3).

Further we assume that vector norm and matrix norm $\| \cdot \| = \| \cdot \|_2$.

Definition 1.1 Function $h(x) : \mathbb{R}^d \rightarrow \mathbb{R}$ is called $L$-smooth if for any $x, y \in \mathbb{R}^d$ it holds

$$\| \nabla h(y) - \nabla h(x) \| \leq L \| y - x \|.$$  

Definition 1.2 Function $h(x) : \mathbb{R}^d \rightarrow \mathbb{R}$ is called $\mu$-strongly convex if for any $x, y \in \mathbb{R}^d$ it holds

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

Assumption 1.3 Function $f_i(x) : \mathbb{R}^d_i \rightarrow \mathbb{R}$ is $\mu_{x,i}$-strongly convex and $L_{x,i}$-smooth function, $\mu_{x,i} \geq 0$, $L_{x,i} > 0$. There exists $i$ such that $\mu_{x,i} > 0$.

Assumption 1.4 Function $g_i(y) : \mathbb{R}^d_i \rightarrow \mathbb{R}$ is $\mu_{y,i}$-strongly convex and $L_{y,i}$-smooth function, $\mu_{y,i} \geq 0$, $L_{y,i} > 0$. There exists $i$ such that $\mu_{y,i} > 0$.

Every agent’s oracle has access only to stochastic gradients of $f_i(x)$ and $g_i(y)$, we denote them as $\nabla f_i(x, \xi_x)$ and $\nabla g_i(y, \xi_y)$ respectively, where $\xi_x$ and $\xi_y$ are random variables.

Assumption 1.5 For each $i$ there exists $\sigma_{f,i}^2$ such that

$${\mathbb{E}_{\xi_x}} \| \nabla f_i(x, \xi_x) - \nabla f_i(x) \|^2 \leq \sigma_{f,i}^2.$$

Assumption 1.6 For each $i$ there exists $\sigma_{g,i}^2$ such that
Assumption 1.7 There exist constants $L_{xy} > 0$, $\mu_{xy}$, $\mu_{yx} \geq 0$ such that

$$L_{xy}^2 \geq \lambda_{\max}(A^TA) = \lambda_{\max}(AA^T),$$

$$\mu_{xy}^2 \leq \begin{cases} \lambda_{\min}^+(AA^T), & \text{if } \nabla g_i(y, \xi) \in \text{range } A \text{ for all } \xi, i \text{ and } y \in \mathbb{R}^{d_y} \\ \lambda_{\min}(AA^T), & \text{otherwise} \end{cases}$$

$$\mu_{yx}^2 \leq \begin{cases} \lambda_{\min}^+(A^TA), & \text{if } \nabla f_i(x, \xi) \in \text{range } A^T \text{ for all } \xi, i \text{ and } x \in \mathbb{R}^{d_x} \\ \lambda_{\min}(A^TA), & \text{otherwise} \end{cases}$$

where $\lambda_{\min}(\cdot)$, $\lambda_{\min}^+(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest, smallest positive and largest eigenvalue of a matrix, respectively, and range(·) denotes the range space of a matrix.

Each node holds its own copy of global variables $x$ and $y$, and we introduce matrices $X = [x_1 x_2 \ldots x_n] \in \mathbb{R}^{d_x \times n}$ and $Y = [y_1 y_2 \ldots y_n] \in \mathbb{R}^{d_y \times n}$. We also denote $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{X} = [\bar{x} \ldots \bar{x}] \in \mathbb{R}^{d_x \times n}$, $\bar{Y} = [\bar{y} \ldots \bar{y}] \in \mathbb{R}^{d_y \times n}$.

Introduce functions

$$F(X) = \sum_{i=1}^n f_i(x_i), \quad G(Y) = \sum_{i=1}^n g_i(y_i). \tag{4}$$

Definition 1.8 Let $S$ be a nonempty set of solutions of saddle-point problem (3). Then we call a pair of vectors $(x, y)$ an $\epsilon$-solution to SPP for given accuracy $\epsilon > 0$ if it satisfies

$$\min_{(x, y)\in S} \max \{\|x - x^*\|^2, \|y - y^*\|^2\} \leq \epsilon. \tag{5}$$

Remark 1.9 Our analysis provides an algorithm whose complexity linearly depends on global parameters.
Local parameters are defined as
\[ L_{lx} = \max_i \{ L_{x,i} \}, \quad \mu_{lx} = \min_i \{ \mu_{x,i} \}, \quad L_{ly} = \max_i \{ L_{y,i} \}, \quad \mu_{ly} = \min_i \{ \mu_{y,i} \}. \]

### 1.3 Main idea: approximation of non-distributed algorithm via consensus subroutine

Problem (3) can be written as
\[
\min_{x \in C_{dx}} \max_{y \in C_{dy}} \frac{1}{n} \sum_{i=1}^{n} f_i(x_i) + y_i^\top A x_i - g_i(y_i),
\]
where
\[ C_{dx} = \{ X \in \mathbb{R}^{d_x \times n} : x_1 = \ldots = x_n \} \text{ and } C_{dy} = \{ Y \in \mathbb{R}^{d_y \times n} : y_1 = \ldots = y_n \}. \]

We want to extend the existing algorithm to the decentralized case of finite-sum SPP. Ideally, it would be very convenient if we could pass the information about gradients from every node to a single centre, average the gradients, and pass them back. This would be equivalent if we made one gradient iteration and averaged variables at each node. Therefore, the main idea of consensus is to make this “averaging” as precise as possible. The only thing we can do is make agents communicate with neighbours. Numerous solutions exist in the literature for achieving this “averaging”, which often translate to the properties provided in the Sect. 3.1. The relevance of such properties is discussed in more detail in “Consensus” section in Rogozin et al. (2021a).

Suppose that at the initial moment of time, each agent contains the same copies of \( x_0 \) and \( y_0 \). Also, let us calculate the gradients in each agent and obtain sets \( \{ \nabla f_i(x_0) \}_{i=1}^{n} \) and \( \{ \nabla g_i(y_0) \}_{i=1}^{n} \), where each element of the set belongs to the corresponding agent. In a fully connected network, we could average these sets in one communicational iteration and obtain the corresponding gradients of functions \( \frac{1}{n} \sum_{i=1}^{n} f_i(x) \) and \( \frac{1}{n} \sum_{i=1}^{n} g_i(x) \), and then perform a gradient-based computation locally in each agent. However, this cannot be done in general, but it is possible to pursue it.

So, we want to make our “averaging” as close as possible to the ideal one, but there are always “inexactness” in equality among agent values (in our example gradients), which can be expressed by the distance from \( Z = X \times Y \) to \( C_{dz} = C_{d_1} \times C_{d_2} \) subspace. Hence, we need the Lemma (2.1) which helps us to transform the “averaging inexactness” to \( \delta \) constant in the definition of inexact oracle, which is introduced in the section below. This allows the inexactness from the consensus to be expressed...
in the context of the inexact oracle, which is more convenient to work with. Then, adjusting the number of consensus iterations, we bound the $\delta$ constant.

We use a fixed number of consensus iterations to obtain the theoretical guarantees that the $\delta$ constant will be sufficiently small during the algorithm. We need to prove that the possible “inexactness” after a gradient iteration is bounded and that this bound can be expressed polynomially through the initial constants.

Since we work in a stochastic setup, we can only get the bound for expectation of “inexactness” after gradient iteration; therefore, we get the bound for expectation of “inexactness” after consensus, which guarantees us the bound in expectation of $\delta$ constant.

To sum up, implementation of this plan requires sensitive and stochastic (in a stochastic decentralized setup) analysis of the algorithm used in the decentralized case.

In our case in particular, we take Algorithm 1 from Kovalev et al. (2021) as a basis for our decentralized algorithm. For this purpose, we need to expand the results from Kovalev et al. (2021) to the case of stochastic inexact oracle.

2 Stochastic inexact framework

This section contains generalisations of some results from Kovalev et al. (2021) and Rogozin et al. (2021a), which themselves have a value outside the spp problem, so it is appropriate to describe them in general terms without being attached to the notations relating to decentralized spp (3). Therefore, in this section only, some notations will be redefined.

2.1 Preliminaries

We will use the definition of $(\delta, L, \mu)$-oracle. Let $h(x)$ be a convex function defined on a convex set $Q \subseteq \mathbb{R}^d$. We say that $(h_{\delta,\mu,L}(x), s_{\delta,\mu,L}(x))$ is a $(\delta, L, \mu)$-model of $h(x)$ at point $x \in Q$ if for all $y \in Q$ holds

$$\frac{\mu}{2} \|y - x\|^2 \leq h(y) - (h_{\delta,\mu,L}(x) + \langle s_{\delta,\mu,L}(x), y - x \rangle) \leq \frac{L}{2} \|y - x\|^2 + \delta.$$

With slight abuse of notation, we say that $\nabla f_\delta(x)$ is $(\delta, L, \mu)$-model of $f(x)$ at point $x$ if there exists $c$ such that $(c, \nabla f_\delta(x))$ is a $(\delta, L, \mu)$-model of $f(x)$ at point $x$. Constants $L$ and $\mu$ are derived from the context.

2.2 Inexact stochastic oracle

Consider the set $\{f_i(x)\}_{i=1}^n$, where each $f_i$ is $L_i$-smooth and $\mu_i$ strongly convex with $\mu_i > 0$. For each $i$, we have access to an unbiased stochastic oracle $\nabla f_i(x, \xi)$ such that $\mathbb{E}_\xi[\nabla f_i(x, \xi)] = \nabla f_i(x)$, and there exists $\sigma_i^2$ fulfilling $\mathbb{E}_\xi[\|\nabla f_i(x, \xi) - \nabla f_i(x)\|^2] \leq \sigma_i^2$.

We define $f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$ and $F(X) = \sum_{i=1}^n f_i(x_i)$. The local parameters $L_i$ and $\mu_i$ are given by $L_i = \max_i \{L_i\}$ and $\mu_i = \min_i \{\mu_i\}$, respectively. Meanwhile, the
global parameters $L_g$, $\mu_g$, and $\sigma^2_g$ are defined as $L_g = \frac{1}{n} \sum_{i=1}^{n} L_i$, $\mu_g = \frac{1}{n} \sum_{i=1}^{n} \mu_i$, and $\sigma^2_g = \frac{1}{n} \sum_{i=1}^{n} \sigma^2_i$, respectively.

The subsequent lemma is a generalization of Lemma 2.1 from Rogozin et al. (2021a), extending to the case of a stochastic unbiased oracle.

**Lemma 2.1** Let $X \in \mathbb{R}^{d \times n}$ and $\|X - \overline{X}\|^2 \leq \delta'$.

Define

$$
\delta = \frac{1}{2n} \left( \frac{L_i^2}{L_g} + \frac{2L_i^2}{\mu_g} + L_i - \mu_i \right) \delta',
$$

$$
f_{\delta,L,\mu}(\overline{x},X) = \frac{1}{n} \left( F(X) + \langle \nabla F(X), \overline{X} - X \rangle + \frac{1}{2} \left( \mu_i - \frac{2L_i}{\mu_g} \right) \| \overline{X} - X \|^2 \right),
$$

$$
g_{\delta,L,\mu}(\overline{x},X) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_i),
$$

$$
\tilde{g}_{\delta,L,\mu}(\overline{x},X) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{r} \sum_{j=1}^{r} \nabla f_i(x_i, \xi_i^j).
$$

Then $(f_{\delta,L,\mu}(\overline{x},X), g_{\delta,L,\mu}(\overline{x},X))$ is a $(\delta, 2L_g, \mu_g/2)$-model of $f$ at point $\overline{x}$. Moreover, we have

$$
\mathbb{E} \tilde{g}_{\delta,L,\mu}(x) = g_{\delta,L,\mu}(x),
$$

$$
\mathbb{E} \| \tilde{g}_{\delta,L,\mu}(x) - g_{\delta,L,\mu}(x) \|^2 \leq \frac{\sum_{i=1}^{n} \sigma^2_i}{n^2 r} = \frac{\sigma^2_g}{nr}.
$$

**Proof** See Lemma 4.1 from Rogozin et al. (2021b). \qed

This is a fundamental theorem for our analysis because it allows us to use an inexact stochastic oracle, through which we can use such approximations (consider the case of $\|X - \overline{X}\|^2 \leq \delta'$).

$$
\frac{\mu_g}{4} \| \overline{y} - \overline{x} \|^2 \leq f(\overline{y}) - f_{\delta,L,\mu}(\overline{x},X) - \langle g_{\delta,L,\mu}(\overline{x},X), \overline{y} - \overline{x} \rangle \leq L_g \| \overline{y} - \overline{x} \|^2 + \delta.
$$

If we could calculate the gradient directly at the point $\overline{x}$, we could get rid of $\delta$ (i.e. to set $\delta = 0$), but this is the price of decentralization, so we are trying to call the oracle at the point that is as close as possible to $C_z$. The iteration of our decentralized algorithm is performed as follows: being in the $\sqrt{\delta'}$ neighborhood of $C_z$, we iterate the basic algorithm, then make projections to the $\sqrt{\delta'}$ neighborhood of $C_z$. 
2.3 Inexact stochastic APDG (accelerated primal-dual gradient descent)

This part considers the problem (1). Function \( f \) is \( \mu_x \)-strongly convex (\( \mu_x \geq 0 \)) and \( L_x \)-smooth, function \( g \) is \( \mu_y \)-strongly convex (\( \mu_y \geq 0 \)) and \( L_y \)-smooth.

**Assumption 2.2** There exist constants \( L_{xy} > 0 \), \( \mu_{xy}, \mu_{yx} \geq 0 \) such that

\[
L_{xy}^2 \geq \lambda_{\max}(AA^T) = \lambda_{\max}(A^TA),
\]

\[
\mu_{xy}^2 \leq \begin{cases} \lambda_{\min}^+(AA^T), & \text{if } \nabla g(y, \xi) \in \text{range } A & \forall \xi, y \in \mathbb{R}^d, \\ \lambda_{\min}^-(AA^T), & \text{otherwise} \end{cases}
\]

\[
\mu_{yx}^2 \leq \begin{cases} \lambda_{\min}^+(A^TA), & \text{if } \nabla f(x, \xi) \in \text{range } A^T & \forall \xi, x \in \mathbb{R}^d, \\ \lambda_{\min}^-(A^TA), & \text{otherwise} \end{cases}
\]

where \( \lambda_{\min}^-(\cdot) \), \( \lambda_{\min}^+(\cdot) \) and \( \lambda_{\max}(\cdot) \) denote the smallest, smallest positive and largest eigenvalue of a matrix, respectively, and range(\( \cdot \)) denotes the range space of a matrix.

This assumption is slightly different from (1.7) and is used in a non-decentralized setting.

**Definition 2.3** (Stochastic process history) Let’s denote the history of the stochastic process up to iteration \( k \) as \( \xi_x^k = (\xi_x^1, \ldots, \xi_x^k, \xi_x^{k+1}, \ldots, \xi_x^k) \). \( \xi_y^k = (\xi_y^1, \ldots, \xi_y^k, \xi_y^{k+1}, \ldots, \xi_y^k) \) are sets of random variables for batches at iteration \( k \); \( r_f, r_g \) are the batch sizes for \( f \) and \( g \) respectively. All introduced random variables are independent.

Each iteration of the algorithm will use a batched gradient for the sake of being able to control the variance of the stochastic part of the oracle. We will also assume that the inexact stochastic gradients of functions \( f \) and \( g \) at the \( k \)th iteration from the random variables \( \xi_x^i \) and \( \xi_y^j \) are also depend from stochastic history \( \xi^{k-1} \). So we denote them by \( \nabla f_\delta(x, \xi_x^i, \xi_y^j) \) and \( \nabla g_\delta(y, \xi_x^i, \xi_y^j) \), where \( 1 \leq i \leq r_f \), \( 1 \leq j \leq r_g \).

**Definition 2.4** (Batched gradients) Denote the batched gradients at iteration \( k \) by

\[
\nabla f_\delta(x, \xi_x^i, \xi_y^j) = \frac{1}{r_f} \sum_{i=1}^{r_f} \nabla f_\delta(x, \xi_x^i, \xi_y^j),
\]

\[
\nabla g_\delta(y, \xi_x^i, \xi_y^j) = \frac{1}{r_g} \sum_{j=1}^{r_g} \nabla g_\delta(y, \xi_x^i, \xi_y^j).
\]

Let \( \nabla f_\delta(x, \xi_y^j) = \mathbb{E}_{\xi_x} \nabla f_\delta(x, \xi_x^i, \xi_y^j) \) and \( \nabla g_\delta(y, \xi_y^j) = \mathbb{E}_{\xi_x} \nabla g_\delta(y, \xi_x^i, \xi_y^j) \).
Decentralized saddle-point problems with different constants…

\begin{align*}
\textbf{Input: } & \eta_x, \eta_y, \alpha_x, \alpha_y, \beta_x, \beta_y > 0, \tau_x, \tau_y, \sigma_x, \sigma_y \in (0, 1], \theta \in (0, 1), \nonumber \\
x^0_f = x^0 \in \text{range } A^i, \nonumber \\
y^0_f = y^0 \in \text{range } A, \nonumber \\
\text{for } k = 0, 1, 2, \ldots \text{ do} \nonumber \\
& y^k_f = y^k + \theta \left(y^k - y^{k-1}\right) \nonumber \\
& x^k = x^k + (1 - \tau_x) x^k_f \nonumber \\
& y^k_y = y^k + \theta \left(y^k - y^{k-1}\right) \nonumber \\
& x^k+1 = x^k + \eta_x \alpha_y x^k \left(y^k - x^k\right) - \eta_x \beta_x A^T (Ax^k - \nabla g_k) - \eta_x \left(\nabla f_k + A^T y^k\right) \nonumber \\
& y^k+1 = y^k + \eta_y \alpha_y \left(y^k - y^{k-1}\right) - \eta_y \beta_y A^T y^k - \eta_y \left(\nabla g_k - A x^{k+1}\right) \nonumber \\
& x^k+1 = x^k + \sigma_x \left(x^{k+1} - x^k\right) \nonumber \\
& y^k+1 = y^k + \sigma_y \left(y^{k+1} - y^k\right) \nonumber \\
\text{end for} \nonumber 
\end{align*}

Algorithm 1 Inexact stochastic APDG

**Assumption 2.5 (Inexact oracle property)** Assume that the inexact constants (denoted as \(\delta\)) for batched gradients also depend on history, that is let \(\nabla g_\delta(y, \xi^k)\) to be the \((\delta_\xi, L_\xi, \mu_\xi)\)-model of \(g\) at point \(y\) and \(\nabla f_\delta(x, \xi^k)\) to be the \((\delta_\xi, L_\xi, \mu_\xi)\)-model of \(f\) at point \(x\), respectively. Moreover, we assume that \(\mathbb{E}_\xi \delta_\xi(\xi^k) \leq \delta_\xi\) for some constants \(\delta_\xi\).

Denote the variance for random variable \(\eta\) by \(\mathbb{D} \eta = \mathbb{E} \|\eta - \mathbb{E} \eta\|^2\).

**Assumption 2.6 (Stochastic oracle property)** It is assumed that \(\mathbb{D}_x \xi^k \nabla f_\delta(x, \xi^k, \xi_{k-1}) \leq \sigma_f^2\) and \(\mathbb{D}_y \xi^k \nabla g_\delta(y, \xi_{k-1}) \leq \sigma_g^2\) for some constants \(\sigma_f^2\) and \(\sigma_g^2\).

For brevity we write \(\nabla g_k = \nabla g_\delta(y^k, \xi^k, \xi_{k-1})\), \(\nabla f_k = \nabla f_\delta(x^k, \xi^k, \xi_{k-1})\).

**Theorem 2.7** Let \(f\) be \(\mu_x\)-strongly convex and \(L_x\) smooth, \(g\) be \(\mu_y\)-strongly convex and \(L_y\) smooth (\(\mu_x, \mu_y \geq 0\)). We have access to \((\delta_x, L_x, \mu_x)\)-stochastic oracle of \(f\) with variance upper bounded by \(\sigma_f^2\) and \((\delta_y, \xi^k, \xi_{k-1})\)-stochastic oracle of \(g\) with variance upper bounded by \(\sigma_g^2\) at iteration \(k\). Denote batch sizes for \(f\) and \(g\) by \(r_f\) and \(r_g\) respectively. Also suppose Assumption (2.2) and Assumption (2.5) hold in environment of Algorithm (1). Then there exist different sets of parameters of Algorithm (1) such that

\[\mathbb{E} \|x^k - x^*\|^2 \leq \frac{\omega}{3 L_{xy}} \left( \theta \kappa \psi^0 + \frac{4}{(1 - \theta)^2} (\delta_x + \delta_y) + \frac{\Sigma^2}{2(1 - \theta)} \right),\]

\[\mathbb{E} \|y^k - y^*\|^2 \leq \frac{1}{4 L_{xy} \omega} \left( \theta \kappa \psi^0 + \frac{4}{(1 - \theta)^2} (\delta_x + \delta_y) + \frac{\Sigma^2}{2(1 - \theta)} \right),\]

\[\Sigma^2 = \left( \frac{1}{L_x} + \frac{\omega}{L_{xy}} \right) \sigma_f^2 r_f + \left( \frac{1}{L_y} + \frac{1}{L_{xy} \omega} \right) \sigma_g^2 r_g.\]
Here is the list of possible estimations depending on different constants:

\[
\frac{1}{1 - \theta} \leq 4 + 4 \max \left\{ \sqrt{\frac{L_x}{\mu_x}}, \sqrt{\frac{L_y}{\mu_y}}, \frac{L_{xy}}{\sqrt{\mu_x \mu_y}} \right\}, \omega = \sqrt{\frac{\mu_y}{\mu_x}}, \quad (7a)
\]

\[
\frac{1}{1 - \theta} \leq 4 + 8 \max \left\{ \sqrt{\frac{L_x L_y}{\mu_{xy}}}, \sqrt{\frac{L_{xy}}{\mu_x}}, \frac{L_{xy}^2}{\mu_{xy}^2} \right\}, \omega = \sqrt{\frac{2 \mu_y L_y}{2 \mu_x L_x}}, \quad (7b)
\]

\[
\frac{1}{1 - \theta} \leq 4 + 8 \max \left\{ \sqrt{\frac{L_x L_y}{\mu_{xy}}} \frac{L_{xy}}{\mu_{yx}}, \sqrt{\frac{L_{xy}}{\mu_y}}, \frac{L_{xy}^2}{\mu_{xy}^2} \right\}, \omega = \sqrt{\frac{2 \mu_y L_y}{\mu_{xy}^2}}, \quad (7c)
\]

\[
\frac{1}{1 - \theta} \leq 2 + 8 \max \left\{ \sqrt{\frac{L_x L_y L_{xy}}{\mu_{xy} \mu_{yx}}}, \sqrt{\frac{L_{xy}^2}{\mu_{xy}^2}}, \frac{L_{xy}^2}{\mu_{xy}^2} \right\}, \omega = \frac{\mu_{xy}}{\mu_{yx}} \sqrt{\frac{L_{xy}}{L_x}}. \quad (7d)
\]

Here \( \Psi^0 \) depends polynomially on the set of initial values \( \mu_x, \mu_y, L_x, L_y, \sigma_f^2, \sigma_g^2, \delta_f, \delta_g, r_f, r_g, x_0, y_0 \).

The proof is provided in “Appendix A”, including the constants values for every estimation.

Each line in the list of possible estimations symbolises particular case in a convex-concave conditions. Firstly, \((7a)\) expresses strongly-convex-strongly-concave case. Secondly, \((7b)\) describes strongly-convex-concave and positive \( \mu_{xy} \) case. Thirdly, line \((7c)\) expresses convex-strongly-concave and positive \( \mu_{xy}^2 \) case. Forthly, \((7d)\) covers convex-concave and positive \( \mu_{xy}, \mu_{yx} \) case. Every case provides a linear convergence rate for corresponding class of problems. The last case is associated with convex-concave problem with square full-range matrix (in this case \( \mu_{xy}, \mu_{yx} > 0 \) and Assumption \((2.2)\) holds).

Speaking about deterministic side, according to Zhang et al. (2019) this upper bound is optimal in strongly-convex-strongly-concave case.

From this theorem, which describes the behavior of the algorithm in the case of an inexact stochastic oracle, a straightforward generalization for a stochastic oracle follows if we set \( \delta_f = \delta_g = 0 \). This result is of independent interest.

There is a work that also considers saddle point problems with bilinear coupling (Du et al. 2022). We begin the comparison to our paper with the deterministic setting. For bilinear games (i.e. for \( F(X) = \text{const} \) and \( G(X) = \text{const} \)) the algorithm in Du et al. (2022) meets lower bounds derived in Ibrahim et al. (2020), whereas our method is not optimal. For strongly-convex-strongly-concave setup, both the method in Du et al. (2022) and our algorithm achieve lower bounds. Considering the stochastic setup, in the aforementioned paper, the number of iterations to achieve the solution accuracy \( \varepsilon \) is \( \mathcal{O}(P \sigma^2 / \varepsilon^2) \), whereas in our paper we have \( \mathcal{O}(Q \sigma^2 / \varepsilon \log(1/\varepsilon)) \) \((P \text{ and } Q \text{ are constants independent of } \varepsilon)\), therefore in our case the dependency on...
\( \varepsilon \) in stochastic part is better. On the one hand, Du et al. (2022) assumes that the gradient of the bilinear part is stochastic, while we assume this gradient to be deterministic. On the other hand, our approach allows one of the functions \( F \) and \( G \) to be strongly-convex (strongly-concave) and the other be non-strongly concave (convex), whereas Du et al. (2022) only supports strongly-convex-strongly-concave case.

### 3 Decentralized algorithm and results

This section considers a decentralised stochastic algorithm for (3). We take Algorithm 1 from Kovalev et al. (2021) as a basis for our method. This algorithm works according to lower bounds and also covers convex-concave case, therefore we obtain desirable non-decentralized inexact stochastic generalization, which may be further generalized to decentralized case.

Every node has its own sequence of independent random variables, which is used in calls of stochastic oracle. Let \( r_{f,i} \) and \( r_{g,i} \) denote batch sizes at node \( i \) for \( f_i \) and \( g_i \), respectively, and let \( \xi^k_{x,i} \) and \( \xi^k_{y,i} \) be the sets of random variables for batches at \( i \)'th node at iteration \( k \).

Introduce a set of batches over \( x \) variable for all agents in \( k \)'th iteration as follows:

\[
\xi^k_x = (\xi^k_{x,1}, \ldots, \xi^k_{x,n})
\]

Next, we introduce a gradient matrix for the function \( F(X) \) as follows:

\[
\nabla F(X, \xi^k_x) = [\nabla f_1(x_1, \xi^k_{x,1}) \cdots \nabla f_n(x_n, \xi^k_{x,n})].
\]

Here \( \nabla f_j(x_j, \xi^k_{x,j}) \) is a batched gradient. Notations \( \xi^k_y \) and \( \nabla G(Y, \xi^k_y) \) are defined in the same way.

In the algorithm below, we write \( \nabla F_k = \nabla F(X^k, \xi^k_x) \) and \( \nabla G_k = \nabla G(Y^k, \xi^k_y) \) for brevity.

The algorithm can be executed in a decentralized way due to “decentralized” property of mixing matrices in Consensus algorithm (Sect. 3.1). One may notice that it is possible to rewrite this algorithm in terms of average values of variables at nodes (algorithm is provided in “Appendix B”). We perform Consensus iterations in order to make calls of stochastic oracle from relatively close to each other points. This process allows the average gradient across nodes to be considered as an inexact gradient at the average point, provided that \( \delta \) is sufficiently small. Consensus iterations do not change the average value due to doubly stochastic property of mixing matrices.

#### 3.1 Consensus

We consider a sequence of non-directed communication graphs \( \{(V, E^k)\}_{k=0}^\infty \) and a sequence of corresponding mixing matrices \( \{W^k\}_{k=0}^\infty \) associated with it. We impose the following assumption.

**Assumption 3.1** Mixing matrix sequence \( \{W^k\}_{k=0}^\infty \) satisfies the following properties.

(Decentralized property) If \( (i,j) \notin E^k \), then \( [W^k]_{ij} = 0 \).
(Doubly stochastic property) \( W_k = \mathbf{X}^0, \mathbf{\bar{X}}^0 \in \text{range } \mathbf{A} \)

(Contraction property) There exist \( u_1 \in \mathbb{Z}^+ \) and \( \lambda \in (0, 1) \), such that for every \( k \geq u_1 - 1 \) it holds

\[
\| W_{k+1} X - \mathbf{X} \| \leq (1 - \lambda) \| X - \mathbf{\bar{X}} \|,
\]

where \( W_{k+1} = W_k \ldots W_{k-u_1+1} \).

A sequence of graphs expresses our time-varying network (vertices are agents, edges are communication links) and a sequence of matrices expresses the rule by which agents will exchange information.

Each agent stores a set of vectors in its local memory. During the communication iteration, if the exchanged set of vectors (one vector for every agent) forms a matrix \( X \) (with vectors arranged in columns), it is replaced by \( XW \) after the iteration, where \( W \) is a gossip matrix that satisfies the decentralized property of (3.1) for the communication graph. This exchange is equivalent to the weighted averaging of all the vectors of one’s own and neighbours, which follows from the basic properties of matrix multiplication.

Define \( \chi \), the characteristic number of the network, as \( \chi = \frac{1}{\lambda} \).
3.2 Complexity results for Algorithm 2

This section uses the global and local function notations from Remark (1.9).
Our analysis shows that performance of Algorithm 2 depends on global constants.

Theorem 3.2 Let $f_i(x)$ be $\mu_{x,i}$-strongly convex and $L_{x,i}$-smooth, and let $g_i(y)$ be $\mu_{y,i}$-strongly convex and $L_{y,i}$ smooth ($\mu_{x,i}, \mu_{y,i} \geq 0$). Also assume that there exist $i_1$ and $i_2$ such that $\mu_{x,i_1}, \mu_{y,i_2} > 0$. The $i$th node has access only to unbiased stochastic gradients of function $f_i$ with variance $\sigma_{f,i}^2$ and of function $g_i$ with variance $\sigma_{g,i}^2$. Also suppose Assumption (3.1) holds, and Assumption (1.7) holds for pairs of functions $(f_i, g_i)$ at every node. Then there exist sets of constants of Algorithm 2, such that to achieve an accuracy of $\varepsilon$, Algorithm 2 exhibits the following complexities.

The number of iterations of Algorithm 2

$$N = \mathcal{O}\left(\frac{1}{1 - \theta} \log \left(\frac{D''}{\varepsilon}\right)\right).$$

The number of communications

$$N_{comm} = \mathcal{O}\left(\frac{1}{1 - \theta} \kappa \log \left(\frac{D''}{\varepsilon}\right) \log \left(\frac{D'}{\varepsilon}\right)\right).$$

The number of stochastic oracle calls at node $i$

$$N_{comp}^i = 2N + \mathcal{O}\left(\frac{\max\{\omega, \omega^{-1}\}}{nL_{xy}(1 - \theta)^2\varepsilon} \left(\left(\frac{1}{L_x} + \frac{\omega}{L_{xy}}\right)\sigma_{f,i}^2 + \left(\frac{1}{L_y} + \frac{1}{L_{xy}\omega}\right)\sigma_{g,i}^2\right) \log \left(\frac{D''}{\varepsilon}\right)\right).$$

Here is the list of possible estimations depending on different constants:

$$\frac{1}{1 - \theta} = \mathcal{O}\left(\max\left\{\sqrt{L_x}, \sqrt{L_y}, \frac{L_{xy}}{\mu_{x}}, \frac{L_{xy}}{\mu_{y}}, \frac{L_{xy}}{\sqrt{\mu_{x} \mu_{y}}}\right\}\right), \quad \omega = \sqrt{\frac{\mu_{y}}{\mu_{x}}},$$

$$\frac{1}{1 - \theta} = \mathcal{O}\left(\max\left\{\sqrt{\frac{L_xL_y}{\mu_{xy}}}, \sqrt{\frac{L_xL_{xy}}{\mu_{xy}}}, \sqrt{\frac{L_{xy}}{\mu_{y}}}, \sqrt{\frac{L_{xy}}{\mu_{x}}}, \sqrt{\frac{L_{xy}}{\mu_{x} \mu_{y}}}\right\}\right), \quad \omega = \sqrt{\frac{\mu_{xy}^2}{2\mu_{x} L_x}},$$

$$\frac{1}{1 - \theta} = \mathcal{O}\left(\max\left\{\sqrt{\frac{L_xL_y}{\mu_{xy}}}, \sqrt{\frac{L_xL_{xy}}{\mu_{xy}}}, \sqrt{\frac{L_{xy}}{\mu_{y}}}, \sqrt{\frac{L_{xy}}{\mu_{x}}}, \sqrt{\frac{L_{xy}}{\mu_{x} \mu_{y}}}\right\}\right), \quad \omega = \sqrt{\frac{2\mu_{xy}L_y}{\mu_{xy}^2}},$$

$$\frac{1}{1 - \theta} = \mathcal{O}\left(\max\left\{\sqrt{\frac{L_xL_yL_{xy}}{\mu_{xy} \mu_{yx}}}, \sqrt{\frac{L_xL_{xy}^2}{\mu_{xy}^2}}, \sqrt{\frac{L_{xy}L_{xy}^2}{\mu_{xy}^2}}, \sqrt{\frac{L_{xy}^2}{\mu_{xy}^2}}, \sqrt{\frac{L_{xy}^2}{\mu_{xy} \mu_{yx}}}\right\}\right), \quad \omega = \sqrt{\frac{\mu_{xy} L_y^2}{\mu_{xy}^2}},$$

where $D'$ and $D''$ are polynomially depend on local variables, global variables, and a pair $\overline{x}_0, \overline{y}_0$.

The proof is provided in “Appendix B”.

Springer
The result associated with the number of communication calls is provided under the time-varying graph conditions. However, in static graph setting we can use Chebyshev-Accelerated subroutine (Scaman et al. 2017) to get $\sqrt{\kappa}$ dependence from the characteristic number of the network. Therefore, our result is optimal up to the logarithmic factor according to Kovalev et al. (2022) (when $L_x = L_y = L_{xy}$ and $\mu_x = \mu_y$). Our stochastic oracle calls result is also optimal according to Beznosikov et al. (2020). The deterministic side for coupling term is optimal (up to logarithmic factor for communications) according to Rogozin et al. (2021a).

It is worth mentioning, that this result only serves the case of $\mu_x, \mu_y > 0$ due to specificity of analysis of basis algorithm. The Lemma (2.1) requires non-zero $\mu_g$, therefore if we wanted to extend our result to convex-concave case, we would need to use regularization technique, this would be possible with duality gap criterion, not in our case.

In conclusion, it is worth to say that one of the contributions of our work is the investigation of the algorithm’s behavior from Kovalev et al. (2021) in a stochastic setting, which led us to an accelerated bilinear algorithm with stochastic analysis, having its own pros and cons compared to the algorithm from Du et al. (2022). Additionally, this work demonstrated that the consensus subroutine technique allows for a relatively straightforward transition of standard algorithms into a decentralized setup, possessing certain advantages (for instance, convergence rate dependency on average smoothness and strong convexity constants), however introducing an extra logarithmic factor. This was further illustrated by obtaining a stochastic decentralized algorithm with average constants; to the best of our knowledge, there were no similar results even in the deterministic (non-stochastic) case prior to this work.

Our work is fundamentally based on theoretical and mathematical constructs, and as such, it does not require the examination or creation of any datasets.

The work of A. Gasnikov was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated November 2, 2021 No. 70-2021-00142.

**Appendix A: Inexact setting**

**A.1 Proof of the Theorem 2.7**

**Lemma A.1** Let us introduce several definitions.

\[ \tau_x = (\sigma_x^{-1} + 1/2)^{-1}, \]

\[ \alpha_x = \mu_x, \]

\[ (8) \]

\[ (9) \]
$$\beta_x = \min \left\{ \frac{1}{2L_y}, \frac{1}{2\eta_xL_{xy}^2} \right\}. \quad (10)$$

Then the following inequality holds

$$\frac{1}{\eta_x} \mathbb{E}_{\xi_1, \xi_2} \| x^{k+1} - x^* \|^2 \leq \left( \frac{1}{\eta_x} - \mu_x - \beta_x\mu_{yx} \right) \| x^k - x^* \|^2$$

$$+ \left( \mu_x + L_x \sigma_x - \frac{1}{2\eta_x} \right) \mathbb{E}_{\xi_1, \xi_2} \| x^{k+1} - x^k \|^2$$

$$+ B_g(y^k_x, y^*) - B_f(x^k_g, x^*) - \frac{2}{\sigma_x} \mathbb{E}_{\xi_1, \xi_2} B_f(x^{k+1}_f, x^*)$$

$$+ \left( \frac{2}{\sigma_x} - 1 \right) B_f(x^k_f, x^*) - 2 \mathbb{E}_{\xi_1, \xi_2} (A^T(y^k_m - y^*), x^{k+1} - x^*)$$

$$+ \delta_y + \left( \frac{4}{\sigma_x} + 1 \right) \delta_x + \beta_x \sigma_g^2 + 2\eta_x \sigma_f^2. \quad (11)$$

**Proof** The proof is similar to the proof of Lemma B.2 in Kovalev et al. (2020). However, we need to expand the proof to cover the inexact stochastic case. To accomplish this, we need to replace several inequalities in the proof from the referred article by the corresponding inequalities with an inexact stochastic oracle.

Let $B_f(a, b) = f(a) - f(b) - \langle \nabla f(b), a - b \rangle$, $B_g(a, b)$ is defined in the same way.

In this analysis, we will need the following inequalities (the term $\xi_{k-1}$ is omitted).

$$\frac{1}{2L_y} \mathbb{E}_{\xi_1} \| g_\phi(y^k_g, \xi^k_g) - g(y^*) \|^2 \leq B_g(y^k_g, y^*) + \delta_y + \frac{\sigma_g^2}{2L_y}, \quad (12)$$

$$\mathbb{E}_{\xi_2} \langle \nabla f_\phi(x^k_g, \xi^k_g) - \nabla f(x^*), x^{k+1} - x^* \rangle \geq \mathbb{E}_{\xi_2} \langle \nabla f_\phi(x^k_g) - \nabla f(x^*), x^{k+1} - x^* \rangle - \eta_x \sigma_f^2, \quad (13)$$

$$\langle \nabla f_\phi(x^k_g) - \nabla f(x^*), x^{k+1}_f - x^k_g \rangle \geq B_f(x^{k+1}_f, x^*) - B_f(x^k_g, x^*) - \frac{L_x}{2} \| x^{k+1}_f - x^k_g \|^2 - \delta_x, \quad (14)$$

$$2\langle \nabla f_\phi(x^k_g) - \nabla f(x^*), x^k_g - x^* \rangle \geq 2B_f(x^k_g, x^*) + \mu_g \| x^k_g - x^* \|^2 - 2\delta_x, \quad (15)$$

$$\langle \nabla f_\phi(x^k_g) - \nabla f(x^*), x^k_f - x^k_g \rangle \leq B_f(x^k_f, x^*) - B_f(x^k_g, x^*) + \delta_x. \quad (16)$$

Let us prove these inequalities.

Inequality (12). Using $\mathbb{E}_{\xi_1} \nabla g_\phi(y^k_g, \xi^k_g) = \nabla g_\phi(y^k_g)$ and according to Theorem 1 from Devolder et al. (2014), we have
\[
\frac{1}{2L_y} \mathbb{E}_{y_i^k} \| \nabla g_\delta(y_{g^k}, z_{g^k}) - \nabla g(y^*) \|^2 = \frac{1}{2L_y} \mathbb{E}_{y_i^k} \| \nabla g_\delta(y_{g^k}, z_{y_i^k}) - \nabla g_\delta(z_{y_i^k}) \|^2 \\
+ \frac{1}{2L_y} \| \nabla g_\delta(y_{g^k}) - \nabla g(y^*) \|^2 \\
\leq g(y_{g^k}^k) - g(y^*) - \langle \nabla g(y^*), y_{g^k}^k - y^* \rangle + \delta_y + \frac{\sigma_y^2}{2L_y} \\
= B_g(y_{g^k}^k, y^*) + \delta_y + \frac{\sigma_y^2}{2L_y}.
\]

We choose inexact oracle \((g_\delta, \nabla g_\delta)\) such that it is the same as \((g_\delta, \nabla g_\delta)\) at all points except \(y^*\), and at \(y^*\) it equals \((g(y^*), \nabla g(y^*))\).

Inequality (13):

\[
\mathbb{E}_{x_i^k} \langle \nabla f_\delta(x_{g^k}^k, z_{x_i^k}^k) - \nabla f(x^*), x_{g^k}^{k+1} - x^* \rangle = \mathbb{E}_{x_i^k} \langle \nabla f_\delta(x_{g^k}^k, z_{x_i^k}^k) - \nabla f_\delta(x_{g^k}^k), x_{g^k}^{k+1} - x^* \rangle \\
+ \langle \nabla f_\delta(x_{g^k}^k) - \nabla f(x^*), x_{g^k}^{k+1} - x^* \rangle.
\]

Using Line (8) of the Algorithm 1 we obtain

\[
\mathbb{E}_{x_i^k} \langle \nabla f_\delta(x_{g^k}^k, z_{x_i^k}^k) - \nabla f(x^*), x_{g^k}^{k+1} - x^* \rangle = \mathbb{E}_{x_i^k} \langle \nabla f_\delta(x_{g^k}^k, z_{x_i^k}^k) - \nabla f_\delta(x_{g^k}^k), -\eta_x \nabla f_\delta(x_{g^k}^k, z_{x_i^k}^k) \rangle \\
+ \mathbb{E}_{x_i^k} \langle \nabla f_\delta(x_{g^k}^k) - \nabla f(x^*), x_{g^k}^{k+1} - x^* \rangle \\
= \mathbb{E}_{x_i^k} \langle \nabla f_\delta(x_{g^k}^k) - \nabla f(x^*), x_{g^k}^{k+1} - x^* \rangle \\
- \eta_x \mathbb{E}_{x_i^k} \| \nabla f_\delta(x_{g^k}^k, z_{x_i^k}^k) - \nabla f_\delta(x_{g^k}^k) \|^2 \\
\geq \mathbb{E}_{x_i^k} \langle \nabla f_\delta(x_{g^k}^k) - \nabla f(x^*), x_{g^k}^{k+1} - x^* \rangle - \eta_x \sigma_f^2.
\]

Inequality (14):

\[
\langle \nabla f_\delta(x_{g^k}^k) - \nabla f(x^*), x_{g^k}^{k+1} - x_{g^k}^k \rangle \geq f(x_{g^k}^{k+1}) - f_\delta(x_{g^k}^k) - \frac{L_x}{2} \| x_{g^k}^{k+1} - x_{g^k}^k \|^2 \\
- \delta_x - \langle \nabla f(x^*), x_{g^k}^{k+1} - x_{g^k}^k \rangle \\
\geq f(x_{g^k}^{k+1}) - f(x_{g^k}^k) - \frac{L_x}{2} \| x_{g^k}^{k+1} - x_{g^k}^k \|^2 \\
- \delta_x - \langle \nabla f(x^*), x_{g^k}^{k+1} - x_{g^k}^k \rangle \\
= B_f(x_{g^k}^{k+1}, x^*) - B_f(x_{g^k}^k, x^*) - \frac{L_x}{2} \| x_{g^k}^{k+1} - x_{g^k}^k \|^2 - \delta_x.
\]

Inequality (15):

\[
2(f(x^*) - f(x_{g^k}^k)) - 2\langle \nabla f_\delta(x_{g^k}^k), x^* - x_{g^k}^k \rangle + 2\delta_x \geq 2(f(x^*) - f_\delta(x_{g^k}^k)) - 2\langle \nabla f_\delta(x_{g^k}^k), x^* - x_{g^k}^k \rangle \\
\geq \mu_x \| x_{g^k}^k - x^* \|^2.
\]

Inequality (16):
\[
\langle \nabla f(x^*) - \nabla f(x^*_j), x_j^k - x^*_j \rangle \leq f(x_j^k) - f(x^*_j) - \langle \nabla f(x^*_j), x_j^k - x^*_j \rangle \\
\leq f(x_j^k) - f(x^*_j) - \langle \nabla f(x^*_f), x_j^k - x^*_f \rangle + \delta_x \\
= B_f(x_j^k, x^*_{f}) - B_f(x^*_f) + \delta_x.
\]

Using Line (8) of the Algorithm 1 we get
\[
\frac{1}{\eta_x} \left\| x^{k+1} - x^* \right\|^2 = \frac{1}{\eta_x} \left\| x^k - x^* \right\|^2 + \frac{2}{\eta_x} \langle x^{k+1} - x^k, x^{k+1} - x^* \rangle - \frac{1}{\eta_x} \left\| x^{k+1} - x^k \right\|^2 \\
= \frac{1}{\eta_x} \left\| x^k - x^* \right\|^2 + 2\alpha_s \langle x^k - x^*, x^{k+1} - x^* \rangle \\
- 2\beta_s \langle \mathbf{A}^T (\mathbf{A}x^k - \nabla g_0(y^k, \zeta^k, \zeta^{k-1}), x^{k+1} - x^*) \rangle \\
- 2 \langle \nabla f_0(x^k, \zeta^k, \zeta^{k-1}) + \mathbf{A}^T y^k_m, x^{k+1} - x^* \rangle - \frac{1}{\eta_x} \left\| x^{k+1} - x^k \right\|^2.
\]

Using the parallelogram rule we get
\[
\frac{1}{\eta_x} \left\| x^{k+1} - x^* \right\|^2 = \frac{1}{\eta_x} \left\| x^k - x^* \right\|^2 \\
+ \alpha_s \left( \left\| x^k - x^* \right\|^2 - \left\| x^k - x^{k+1} \right\|^2 - \left\| x^k - x^* \right\|^2 + \left\| x^{k+1} - x^k \right\|^2 \right) \\
- 2\beta_s \langle \mathbf{A}^T (\mathbf{A}x^k - \nabla g_0(y^k, \zeta^k, \zeta^{k-1}), x^{k+1} - x^*) \rangle \\
- 2 \langle \nabla f_0(x^k, \zeta^k, \zeta^{k-1}) + \mathbf{A}^T y^k_m, x^{k+1} - x^* \rangle - \frac{1}{\eta_x} \left\| x^{k+1} - x^k \right\|^2.
\]

Using the optimality condition \( \nabla g(y^*) = \mathbf{A}x^* \), which follows from \( \nabla_j f(x^*, y^*) = 0 \), and the parallelogram rule we get
\[
\frac{1}{\eta_x} \left\| x^{k+1} - x^* \right\|^2 = \frac{1}{\eta_x} \left\| x^k - x^* \right\|^2 \\
+ \alpha_s \left( \left\| x^k - x^* \right\|^2 - \left\| x^k - x^{k+1} \right\|^2 - \left\| x^k - x^* \right\|^2 + \left\| x^{k+1} - x^k \right\|^2 \right) \\
+ \beta_s \left( \left\| \mathbf{A}(x^{k+1} - x^*) \right\|^2 - \left\| \mathbf{A}(x^{k+1} - x^*) \right\|^2 \right) \\
+ \beta_s \left( \left\| \nabla g_0(y^k, \zeta^k, \zeta^{k-1}) - \nabla g(y^*) \right\|^2 - \left\| \nabla g_0(y^k, \zeta^k, \zeta^{k-1}) - \mathbf{A}(x^{k+1}) \right\|^2 \right) \\
- 2 \langle \nabla f_0(x^k, \zeta^k, \zeta^{k-1}) + \mathbf{A}^T y^k_m, x^{k+1} - x^* \rangle - \frac{1}{\eta_x} \left\| x^{k+1} - x^k \right\|^2.
\]

Using Assumption (1.7) and Eq. (12), we get
Using the optimality condition \( \nabla f(x^*) + A^T y^* = 0 \), which follows from \( \nabla_x F(x^*, y^*) = 0 \) and Eq. (13), we get

\[
\frac{1}{\eta_x} \mathbb{E}_{\xi^g_t, \xi^x_t} \left\| x^{k+1} - x^* \right\|^2 
\leq \frac{1}{\eta_x} \left\| x^k - x^* \right\|^2 + \alpha_x \mathbb{E}_{\xi^g_t, \xi^x_t} \left\| x^{k+1} - x^k \right\|^2 + \beta_x L^2_{\xi^y} \mathbb{E}_{\xi^g_t, \xi^x_t} \left\| x^{k+1} - x^k \right\|^2

+ \alpha_x \mathbb{E}_{\xi^g_t, \xi^x_t} \left\| x^k - x^* \right\|^2 + \beta_x L^2_{\xi^y} \mathbb{E}_{\xi^g_t, \xi^x_t} \left\| x^{k+1} - x^k \right\|^2

- \beta_x \mu_{x}^2 \left\| x^k - x^* \right\|^2

+ 2 \beta_x L_y B_g (y_g^k, y^*) + \beta_x \sigma_{g}^2

- 2 \mathbb{E}_{\xi^g_t, \xi^x_t} (\nabla f(x^k, \xi_x^k, \xi^k_m) + A^T y^k, x^{k+1} - x^*)

- \frac{1}{\eta_x} \mathbb{E}_{\xi^g_t, \xi^x_t} \left\| x^{k+1} - x^* \right\|^2

= \left( \frac{1}{\eta_x} - \alpha_x - \beta_x \mu_{x}^2 \right) \left\| x^k - x^* \right\|^2

+ \left( \beta_x L^2_{\xi^y} + \alpha_x - \frac{1}{\eta_x} \right) \mathbb{E}_{\xi^g_t, \xi^x_t} \left\| x^{k+1} - x^k \right\|^2 + 2 \beta_x L_y B_g (y_g^k, y^*)

+ \alpha_x \left\| x^k - x^* \right\|^2 - 2 \mathbb{E}_{\xi^g_t, \xi^x_t} (\nabla f(x^k, \xi_x^k, \xi^k_m) - \nabla f(x^k), x^{k+1} - x^*)

- 2 \mathbb{E}_{\xi^g_t, \xi^x_t} (A^T (y^k - y^*), x^{k+1} - x^*) + 2 \beta_x L_y \delta_j (\xi^k_m) + \beta_x \sigma_{g}^2

Using \( \mu_g \)-strong convexity of \( f \) and Lines (6) and (10) of the Algorithm 1 and Eq. (15) we get
Using Eq. (16), we get

\[
\frac{1}{\eta_k} E_{\xi_k, z_k} \|x^{k+1} - x^*\|^2 \leq \left( \frac{1}{\eta_k} - \alpha_x - \beta_x \mu_{yx}^2 \right) \|x^k - x^*\|^2 \\
+ \left( \beta_x L_{xy}^2 + \alpha_x - \frac{1}{\eta_k} \right) E_{\xi_k, z_k} \|x^{k+1} - x^k\|^2 \\
+ 2 \beta_x L_x B_{xy}(y^k, y^*) \\
+ \alpha_x \|x_k^* - x^*\|^2 - \frac{2}{\sigma_x} E_{\xi_k, \gamma_k} (\nabla f_\delta(x_k^*, \xi_{k-1}^*) - \nabla f(x^*), x_{f}^{k+1} - x_g^k) \\
+ \frac{2(1 - \tau_x)}{\tau_x} (\nabla f_\delta(x_k^*, \xi_{k-1}^*) - \nabla f(x^*), x_{f}^k - x_g^k) \\
- \mu_{x} \|x_k^* - x^*\|^2 + 2 \delta_x(x_k^* - x^*) \\
+ 2 \beta_x L_x \delta_{\gamma}(\xi_{k-1}^*) + \beta_x \sigma_g^2 + 2 \eta_x \sigma_f^2 + 2 \delta_x(\xi_{k-1}^*). 
\]
Using Eq. (14), we get
\[
\frac{1}{\eta_x} \mathbb{E}_{\xi_t, \zeta_t} \| x^{k+1} - x^* \|^2 \leq \left( \frac{1}{\eta_x} - \alpha_s - \beta_s \mu_{y_x}^2 \right) \| x^k - x^* \|^2 \\
+ \left( \beta_s L_{xy} + \alpha_s - \frac{1}{\eta_x} \right) \mathbb{E}_{\xi_t, \zeta_t} \| x^{k+1} - x^k \|^2 + (\alpha_s - \mu_s) \| x^k - x^* \|^2 \\
+ 2 \beta_s L_y B_y(y_g^k, y^*) - 2B_f(x_g^k, x^*) \\
- \frac{2}{\sigma_x} \mathbb{E}_{\xi_t, \zeta_t} \left( B_f(x_{y_{jk}}^{k+1}, x^*) - B_f(x_{y_{jk}}^k, x^*) - \frac{L}{2} \| x_{y_{jk}}^{k+1} - x_{y_{jk}}^k \|^2 \right) \\
+ \frac{2(1 - \tau_x)}{\tau_x} \left( B_f(x_{y_{jk}}^k, x^*) - B_f(x_{y_{jk}}^{k-1}, x^*) \right) \\
- 2 \mathbb{E}_{\xi_t, \zeta_t} (A^T(y^k - y^*), x^{k+1} - x^*) \\
+ 2 \beta_s L_y \delta_y(\xi^{k-1}) + \beta_s \sigma_y^2 + 2 \eta_s \sigma_f^2 + \left( \frac{2}{\tau_x} + \frac{2}{\sigma_x} \right) \delta_s(\xi^{k-1}).
\]

Using Line (10) of the Algorithm 1 we get
\[
\frac{1}{\eta_x} \mathbb{E}_{\xi_t, \zeta_t} \| x^{k+1} - x^* \|^2 \leq \left( \frac{1}{\eta_x} - \alpha_s - \beta_s \mu_{y_x}^2 \right) \| x^k - x^* \|^2 \\
+ \left( \beta_s L_{xy} + \alpha_s - \frac{1}{\eta_x} \right) \mathbb{E}_{\xi_t, \zeta_t} \| x^{k+1} - x^k \|^2 + (\alpha_s - \mu_s) \| x^k - x^* \|^2 \\
+ 2 \beta_s L_y B_y(y_g^k, y^*) - 2B_f(x_g^k, x^*) \\
- \frac{2}{\sigma_x} \mathbb{E}_{\xi_t, \zeta_t} \left( B_f(x_{y_{jk}}^{k+1}, x^*) - B_f(x_{y_{jk}}^k, x^*) - \frac{L}{2} \| x_{y_{jk}}^{k+1} - x_{y_{jk}}^k \|^2 \right) \\
+ \frac{2(1 - \tau_x)}{\tau_x} \left( B_f(x_{y_{jk}}^k, x^*) - B_f(x_{y_{jk}}^{k-1}, x^*) \right) \\
- 2 \mathbb{E}_{\xi_t, \zeta_t} (A^T(y^k - y^*), x^{k+1} - x^*) \\
+ 2 \beta_s L_y \delta_y(\xi^{k-1}) + \beta_s \sigma_y^2 + 2 \eta_s \sigma_f^2 + \left( \frac{2}{\tau_x} + \frac{2}{\sigma_x} \right) \delta_s(\xi^{k-1}).
\]
Transforming this inequality we get

\[
\frac{1}{\eta_x} \mathbb{E}_{\xi_x, \eta_x} \left\| x^{k+1} - x^* \right\|^2 \leq \left( \frac{1}{\eta_x} - \alpha_x - \beta_x \mu_x^2 \right) \left\| x^k - x^* \right\|^2 \\
+ \left( \beta_x L_x^2 + \alpha_x + L_x \sigma_x - \frac{1}{\eta_x} \right) \mathbb{E}_{\xi_x, \eta_x} \left\| x^{k+1} - x^k \right\|^2 \\
+ (\alpha_x - \mu_x) \left\| x_g^k - x^* \right\|^2 + 2\beta_x L_x B_g(y_g^k, y^*) + \left( \frac{2}{\sigma_x} - \frac{2}{\tau_x} \right) B_f(x^k_g, x^*) \\
- \frac{2}{\sigma_x} \mathbb{E}_{\xi_x, \eta_x} B_f(x^{k+1}_g, x^*) + \left( \frac{2}{\sigma_x} - 2 \right) B_f(x^k_g, x^*) \\
- 2\mathbb{E}_{\xi_x, \eta_x} \langle A^T(y^k_m - y^*), x^{k+1} - x^* \rangle \\
+ 2\beta_x L_y \delta_g(s^{k-1}) + \beta_x \sigma_g^2 + 2\eta_x \sigma_f^2 + \left( \frac{2}{\tau_x} + \frac{2}{\sigma_x} \right) \delta_x(s^{k-1}).
\]

Using the definition of \( \tau_x, \alpha_x \) and \( \beta_x \) we get

\[
\frac{1}{\eta_x} \mathbb{E}_{\xi_x, \eta_x} \left\| x^{k+1} - x^* \right\|^2 \leq \left( \frac{1}{\eta_x} - \mu_x - \beta_x \mu_x^2 \right) \left\| x^k - x^* \right\|^2 \\
+ \left( \mu_x + L_x \sigma_x - \frac{1}{2\eta_x} \right) \mathbb{E}_{\xi_x, \eta_x} \left\| x^{k+1} - x^k \right\|^2 \\
+ B_g(y_g^k, y^*) - B_f(x^k_g, x^*) \\
- \frac{2}{\sigma_x} \mathbb{E}_{\xi_x, \eta_x} B_f(x^{k+1}_g, x^*) + \left( \frac{2}{\sigma_x} - 1 \right) B_f(x^k_g, x^*) \\
- 2\mathbb{E}_{\xi_x, \eta_x} \langle A^T(y^k_m - y^*), x^{k+1} - x^* \rangle \\
+ \delta_g(s^{k-1}) + \left( \frac{4}{\sigma_x} + 1 \right) \delta_x(s^{k-1}) + \beta_x \sigma_g^2 + 2\eta_x \sigma_f^2.
\]

\( \square \)

**Lemma A.2** Let us introduce several definitions.

\[
\tau_y = (\sigma_y^{-1} + 1/2)^{-1}, \quad (17)
\]

\[
\alpha_y = \mu_y, \quad (18)
\]

\[
\beta_y = \min \left\{ \frac{1}{2L_x}, \frac{1}{2\eta_y L_y^2} \right\}, \quad (19)
\]

\( \copyright \) Springer
Then the following inequality holds

\[
\frac{1}{\eta_y} E_{\xi_k, \xi_y} \|y^{k+1} - y^*\|^2 \leq \left( \frac{1}{\eta_y} - \mu_y - \beta_y \mu_{xy}^2 \right) \|y^k - y^*\|^2 \\
+ \left( \mu_y + L_y \sigma_y - \frac{1}{2\eta_y} \right) E_{\xi_k, \xi_y} \|y^{k+1} - y^k\|^2 \\
+ B_f(x^k, y^k) - B_f(y^k, y^k) - \frac{2}{\sigma_y} E_{\xi_k, \xi_y} B_g(y^{k+1}, y^k) \\
+ \left( \frac{2}{\sigma_y} - 1 \right) B_g(y^k, y^k) + 2 E_{\xi_k, \xi_y} (A(x^{k+1} - x^k), y^{k+1} - y^k) \\
+ \delta_x(\xi^{k-1}) + \left( \frac{4}{\sigma_y} + 1 \right) \delta_y(\xi^{k-1}) + \beta_y \sigma_f^2 + 2\eta_y \sigma_g^2.
\]

(20)

**Proof** The proof is similar to the proof of the previous lemma.

**Lemma A.3** Let \( \eta_y \) be defined as

\[
\eta_y = \min \left\{ \frac{1}{4(\mu_y + L_y \sigma_y)}, \frac{\omega}{4L_{xy}} \right\}.
\]

and let \( \eta_y \) be defined as

\[
\eta_y = \min \left\{ \frac{1}{4(\mu_y + L_y \sigma_y)}, \frac{1}{4L_{xy} \omega} \right\}.
\]

where \( \omega > 0 \) is a parameter. Let \( \theta \) be defined as

\[
\theta = \theta(\omega, \sigma_x, \sigma_y) = 1 - \max\{ \rho_a(\omega, \sigma_x, \sigma_y), \rho_b(\omega, \sigma_x, \sigma_y), \rho_c(\omega, \sigma_x, \sigma_y), \rho_d(\omega, \sigma_x, \sigma_y) \},
\]

where

\[
\rho_a(\omega, \sigma_x, \sigma_y) = \max \left\{ \frac{4(\mu_y + L_y \sigma_y)}{\mu_y}, \frac{2}{\sigma_x}, \frac{4(\mu_y + L_y \sigma_y)}{\mu_y}, \frac{2}{\sigma_y}, \frac{4L_{xy}}{\mu_y}, \frac{4L_{xy} \omega}{\mu_y} \right\}^{-1},
\]

\[
\rho_b(\omega, \sigma_x, \sigma_y) = \max \left\{ \frac{4(\mu_y + L_y \sigma_y)}{\mu_y}, \frac{2}{\sigma_x}, \frac{8L_y(\mu_y + L_y \sigma_y)}{\mu_y}, \frac{2}{\sigma_y}, \frac{2L_{xy}}{\mu_y}, \frac{8L_{xy} \omega}{\mu_y} \right\}^{-1},
\]

\[
\rho_c(\omega, \sigma_x, \sigma_y) = \max \left\{ \frac{4(\mu_y + L_y \sigma_y)}{\mu_y}, \frac{2}{\sigma_x}, \frac{8L_y(\mu_y + L_y \sigma_y)}{\mu_y}, \frac{2}{\sigma_y}, \frac{2L_{xy}^2}{\mu_y}, \frac{8L_{xy} \omega}{\mu_y} \right\}^{-1},
\]

\[
\rho_d(\omega, \sigma_x, \sigma_y) = \max \left\{ \frac{8L_y(\mu_y + L_y \sigma_y)}{\mu_y}, \frac{2}{\sigma_x}, \frac{8L_y(\mu_y + L_y \sigma_y)}{\mu_y}, \frac{2}{\sigma_y}, \frac{8L_{xy} \omega}{\mu_y}, \frac{2L_{xy}^2}{\mu_y}, \frac{2L_{xy}^2}{\mu_y} \right\}^{-1}.
\]

Let \( \Psi^k \) be the following Lyapunov function:
Using Line (6) of the Algorithm 1 we get

\[ \frac{\eta_x}{\eta_y} \geq \frac{1}{\eta_x} \| x^k - x^* \|^2 + \frac{1}{\eta_y} \| y^k - y^* \|^2 + \frac{2}{\sigma_x} B_f (x^k_j, x^*), \]

Then, the following inequalities hold

\[ \psi^k \geq \frac{3}{4 \eta_x} \| x^k - x^* \|^2 + \frac{1}{\eta_y} \| y^k - y^* \|^2, \quad \text{(22)} \]

\[ \mathbb{E} \psi^{k+1} \leq \theta \mathbb{E} \psi^k + \frac{4}{1 - \theta} (\delta_x + \delta_y) + \frac{1}{2} \left( \frac{1}{L_x} + \frac{\omega}{L_{xy}} \right) \sigma_f^2 + \frac{1}{2} \left( \frac{1}{L_y} + \frac{1}{L_{xy} \omega} \right) \sigma_g^2. \quad \text{(23)} \]

**Proof** The proof of this lemma is similar to the proof of Lemma B.4. in Kovalev et al. (2021).

After adding up (A.3) and (A.2) we get

(LHS) \leq \left( \frac{1}{\eta_x} - \mu_x - \beta_x \mu_{y, x}^2 \right) \| x^k - x^* \|^2 + \left( \frac{1}{\eta_y} - \mu_y - \beta_y \mu_{y, y}^2 \right) \| y^k - y^* \|^2

\[ + \left( \frac{2}{\sigma_x} - 1 \right) B_g (y^k, y^*) + 2 \mathbb{E} \xi^k \mathbb{E} \xi^k \| y^{k+1} - y^* \|^2 \]

\[ + \left( 2 + \frac{4}{\sigma_y} \right) \delta_x (x^{k-1}) + \left( 2 + \frac{4}{\sigma_y} \right) \delta_y (y^{k-1}) + (\beta_x + 2 \eta_x) \sigma_f^2 + (\beta_x + 2 \eta_x) \sigma_g^2, \]

where (LHS) is given as

\[ \text{(LHS)} = \frac{1}{\eta_x} \mathbb{E} \xi^k \mathbb{E} \xi^k \| x^{k+1} - x^* \|^2 + \frac{1}{\eta_y} \mathbb{E} \xi^k \mathbb{E} \xi^k \| y^{k+1} - y^* \|^2 \]

\[ + \frac{2}{\sigma_x} \mathbb{E} \xi^k \mathbb{E} \xi^k B_f (x^{k+1}_j, x^*), \quad \frac{2}{\sigma_y} \mathbb{E} \xi^k \mathbb{E} \xi^k B_g (y^{k+1}, y^*). \]

Using Line (6) of the Algorithm 1 we get
Using Assumption (1.7) we get

\[
\begin{align*}
\text{(LHS)} & \leq \left( \frac{1}{\eta_x} - \mu_x - \beta_x \mu_{y,x}^2 \right) \left\| x^k - x^* \right\|^2 + \left( \frac{1}{\eta_y} - \mu_y - \beta_y \mu_{x,y}^2 \right) \left\| y^k - y^* \right\|^2 \\
& \quad + \left( \mu_x + L_x \sigma_x - \frac{1}{2\eta_x} \right) E_{S^{\xi_1,\xi_y}} \left\| x^{k+1} - x^k \right\|^2 \\
& \quad + \left( \mu_y + L_y \sigma_y - \frac{1}{2\eta_y} \right) E_{S^{\xi_1,\xi_y}} \left\| y^{k+1} - y^k \right\|^2 \\
& \quad + \left( \frac{2}{\sigma_x} - 1 \right) B_f (x^k, x^*) + \left( \frac{2}{\sigma_y} - 1 \right) B_g (y^k, y^*) \\
& \quad + 2E_{S^{\xi_1,\xi_y}} (y^{k+1} - y^k, \Lambda (x^{k+1} - x^*)) - 2\theta (y^k - y^{k-1}, \Lambda (x^k - x^*)) \\
& \quad + \left( 2 + \frac{4}{\sigma_x} \right) \delta_x (\bar{\varepsilon}^{k-1}) + \left( 2 + \frac{4}{\sigma_y} \right) \delta_y (\bar{\varepsilon}^{k-1}) + (\beta_x + 2\eta_x) \sigma_f^2 + (\beta_y + 2\eta_y) \sigma_g^2.
\end{align*}
\]
Using the definition of $\eta_x$ and $\eta_y$ and the definition of $\theta < 1$ we get

\[
\begin{align*}
\text{(LHS)} & \leq \left( \frac{1}{\eta_x} - \mu_x - \beta_x \mu_{xy}^2 \right) \left\| x^k - x^* \right\|^2 + \left( \frac{1}{\eta_y} - \mu_y - \beta_y \mu_{xy}^2 \right) \left\| y^k - y^* \right\|^2 \\
& \quad - \frac{1}{4 \eta_x} \mathbb{E}_{\xi_x, \xi_y} \left\| x^{k+1} - x^k \right\|^2 - \frac{1}{4 \eta_y} \mathbb{E}_{\xi_x, \xi_y} \left\| y^{k+1} - y^k \right\|^2 \\
& \quad + \left( \frac{2}{\sigma_x} - 1 \right) B_f(x_f^k, x^*) + \left( \frac{2}{\sigma_y} - 1 \right) B_g(y_f^k, y^*) \\
& \quad + 2 \mathbb{E}_{\xi_x, \xi_y} (y^{k+1} - y^k, A(x^{k+1} - x^*)) - 2 \theta (y^k - y^{k-1}, A(x^k - x^*)) \\
& \quad + \frac{\theta}{2 \sqrt{\eta_x \eta_y}} \mathbb{E}_{\xi_x, \xi_y} \left\| y^k - y^{k-1} \right\| \left\| x^{k+1} - x^k \right\| \\
& \quad + \frac{4}{1 - \theta} (\delta_x(\xi^{k-1}) + \delta_y(\xi^{k-1})) + \left( \beta_y + \frac{\omega}{2L_{xy}} \right) \sigma_f^2 + \left( \beta_x + \frac{1}{2L_{xy} \omega} \right) \sigma_g^2.
\end{align*}
\]

Transforming this inequality we get

\[
\begin{align*}
\text{(LHS)} & \leq \left( \frac{1}{\eta_x} - \mu_x - \beta_x \mu_{xy}^2 \right) \left\| x^k - x^* \right\|^2 + \left( \frac{1}{\eta_y} - \mu_y - \beta_y \mu_{xy}^2 \right) \left\| y^k - y^* \right\|^2 \\
& \quad - \frac{1}{4 \eta_x} \mathbb{E}_{\xi_x, \xi_y} \left\| x^{k+1} - x^k \right\|^2 - \frac{1}{4 \eta_y} \mathbb{E}_{\xi_x, \xi_y} \left\| y^{k+1} - y^k \right\|^2 \\
& \quad + \left( \frac{2}{\sigma_x} - 1 \right) B_f(x_f^k, x^*) + \left( \frac{2}{\sigma_y} - 1 \right) B_g(y_f^k, y^*) \\
& \quad + 2 \mathbb{E}_{\xi_x, \xi_y} (y^{k+1} - y^k, A(x^{k+1} - x^*)) - 2 \theta (y^k - y^{k-1}, A(x^k - x^*)) \\
& \quad + \frac{\theta}{4 \eta_x} \mathbb{E}_{\xi_x, \xi_y} \left\| x^{k+1} - x^k \right\|^2 + \frac{\theta}{4 \eta_y} \left\| y^k - y^{k-1} \right\|^2 \\
& \quad + \frac{4}{1 - \theta} (\delta_x(\xi^{k-1}) + \delta_y(\xi^{k-1})) + \left( \beta_y + \frac{\omega}{2L_{xy}} \right) \sigma_f^2 + \left( \beta_x + \frac{1}{2L_{xy} \omega} \right) \sigma_g^2.
\end{align*}
\]

Transforming further

\[
\begin{align*}
\text{(LHS)} & \leq \left( \frac{1}{\eta_x} - \mu_x - \beta_x \mu_{xy}^2 \right) \left\| x^k - x^* \right\|^2 + \left( \frac{1}{\eta_y} - \mu_y - \beta_y \mu_{xy}^2 \right) \left\| y^k - y^* \right\|^2 \\
& \quad + \frac{\theta}{4 \eta_y} \left\| y^k - y^{k-1} \right\|^2 - \frac{1}{4 \eta_y} \mathbb{E}_{\xi_x, \xi_y} \left\| y^{k+1} - y^k \right\|^2 \\
& \quad + \left( \frac{2}{\sigma_x} - 1 \right) B_f(x_f^k, x^*) + \left( \frac{2}{\sigma_y} - 1 \right) B_g(y_f^k, y^*) \\
& \quad + 2 \mathbb{E}_{\xi_x, \xi_y} (y^{k+1} - y^k, A(x^{k+1} - x^*)) - 2 \theta (y^k - y^{k-1}, A(x^k - x^*)) \\
& \quad + \frac{4}{1 - \theta} (\delta_x(\xi^{k-1}) + \delta_y(\xi^{k-1})) + \left( \beta_y + \frac{\omega}{2L_{xy}} \right) \sigma_f^2 + \left( \beta_x + \frac{1}{2L_{xy} \omega} \right) \sigma_g^2.
\end{align*}
\]
Using the definition of $\beta_x$ and $\beta_y$ we get

$$(LHS) \leq \left( 1 - \eta_x \mu_x - \min \left\{ \frac{\eta_x \mu_{yx}^2}{2L_y}, \frac{\mu_{yx}^2}{2L_{xy}} \right\} \right) \frac{1}{\eta_x} \| x^k - x^* \|^2$$

$$+ \left( 1 - \eta_y \mu_y - \min \left\{ \frac{\eta_y \mu_{xy}^2}{2L_x}, \frac{\mu_{xy}^2}{2L_{xy}} \right\} \right) \frac{1}{\eta_y} \| y^k - y^* \|^2$$

$$+ \frac{\theta}{4\eta_y} \| y^k - y^{k-1} \|^2 - \frac{1}{4\eta_y} \mathbb{E}_{s_i, s_i'} \| y^{k+1} - y^k \|^2$$

$$+ \left( \frac{2}{\sigma_x} - 1 \right) B_j(x^*_j, x^*) + \left( \frac{2}{\sigma_y} - 1 \right) B_g(y^*_j, y^*)$$

$$+ 2\mathbb{E}_{s_i, s_i'} (y^{k+1} - y^k, A(x^{k+1} - x^*)) - 2\theta (y^k - y^{k-1}, A(x^k - x^*))$$

$$+ \frac{4}{1 - \theta} (\delta_x(\xi^{k-1}) + \delta_y(\zeta^{k-1})) + \frac{1}{2} \left( \frac{1}{L_x} + \frac{\omega}{L_{xy}} \right) \sigma_f^2 + \frac{1}{2} \left( \frac{1}{L_y} + \frac{1}{L_{xy} \omega} \right) \sigma_g^2$$

Transforming this inequality

$$(LHS) \leq \left( 1 - \max \left\{ \eta_x \mu_x, \min \left\{ \frac{\eta_x \mu_{yx}^2}{2L_y}, \frac{\mu_{yx}^2}{2L_{xy}} \right\} \right\} \right) \frac{1}{\eta_x} \| x^k - x^* \|^2$$

$$+ \left( 1 - \max \left\{ \eta_y \mu_y, \min \left\{ \frac{\eta_y \mu_{xy}^2}{2L_x}, \frac{\mu_{xy}^2}{2L_{xy}} \right\} \right\} \right) \frac{1}{\eta_y} \| y^k - y^* \|^2$$

$$+ \frac{\theta}{4\eta_y} \| y^k - y^{k-1} \|^2 - \frac{1}{4\eta_y} \mathbb{E}_{s_i, s_i'} \| y^{k+1} - y^k \|^2$$

$$+ \left( \frac{2}{\sigma_x} - 1 \right) B_j(x^*_j, x^*) + \left( \frac{2}{\sigma_y} - 1 \right) B_g(y^*_j, y^*)$$

$$+ 2\mathbb{E}_{s_i, s_i'} (y^{k+1} - y^k, A(x^{k+1} - x^*)) - 2\theta (y^k - y^{k-1}, A(x^k - x^*))$$

$$+ \frac{4}{1 - \theta} (\delta_x(\xi^{k-1}) + \delta_y(\zeta^{k-1})) + \frac{1}{2} \left( \frac{1}{L_x} + \frac{\omega}{L_{xy}} \right) \sigma_f^2 + \frac{1}{2} \left( \frac{1}{L_y} + \frac{1}{L_{xy} \omega} \right) \sigma_g^2.$$
Using the definition of $\theta$ we get

\[
\text{(LHS)} \leq \theta \left( \frac{1}{\eta_x} \| x^k - x^* \|^2 + \frac{1}{\eta_y} \| y^k - y^* \|^2 + \frac{1}{4\eta_y} \| y^k - y^{k-1} \|^2 \right) \\
+ \theta \left( -2(y^k - y^{k-1}, A(x^k - x^*)) + \frac{2}{\sigma_x} B_f(x_j^k, y^k) + \frac{2}{\sigma_y} B_g(y_j^k, y^*) \right) \\
- \frac{1}{4\eta_y} \mathbb{E}_{z_i, z_i'} \| y^{k+1} - y^k \|^2 + 2\mathbb{E}_{z_i, z_i'} (y^{k+1} - y^k, A(x^{k+1} - x^*)) \\
+ \frac{4}{1 - \theta} \left( \delta_x(\xi^{k-1}) + \delta_y(\xi^{k-1}) \right) + \frac{1}{2} \left( \frac{1}{L_x} + \frac{\omega}{L_{xy}} \right) \sigma_j^2 + \frac{1}{2} \left( \frac{1}{L_y} + \frac{1}{L_y \omega} \right) \sigma_g^2.
\]

After taking the expectation over all random variables, rearranging and using the definition of $\Psi^k$, using the fact that $\mathbb{E} \delta_x(\xi^{k-1}) \leq \delta_x, \mathbb{E} \delta_y(\xi^{k-1}) \leq \delta_y$ we get

\[
\mathbb{E} \Psi^{k+1} \leq \theta \mathbb{E} \Psi^k + \frac{4}{1 - \theta} \left( \delta_x + \delta_y \right) + \frac{1}{2} \left( \frac{1}{L_x} + \frac{\omega}{L_{xy}} \right) \sigma_j^2 + \frac{1}{2} \left( \frac{1}{L_y} + \frac{1}{L_y \omega} \right) \sigma_g^2.
\]

Finally, using the definition of $\Psi^k$, $\eta_x$ and $\eta_y$ we get

\[
\Psi^k \geq \frac{1}{\eta_x} \| x^k - x^* \|^2 + \frac{1}{\eta_y} \| y^k - y^* \|^2 + \frac{1}{4\eta_y} \| y^k - y^{k-1} \|^2 - 2(y^k - y^{k-1}, A(x^k - x^*)) \\
\geq \frac{1}{\eta_x} \| x^k - x^* \|^2 + \frac{1}{\eta_y} \| y^k - y^* \|^2 + \frac{1}{4\eta_y} \| y^k - y^{k-1} \|^2 - 2L_{xy} \| y^k - y^{k-1} \| \| x^k - x^* \| \\
\geq \frac{1}{\eta_x} \| x^k - x^* \|^2 + \frac{1}{\eta_y} \| y^k - y^* \|^2 + \frac{1}{4\eta_y} \| y^k - y^{k-1} \|^2 - \frac{1}{2\sqrt{\eta_x \eta_y}} \| y^k - y^{k-1} \| \| x^k - x^* \| \\
\geq \frac{3}{4\eta_x} \| x^k - x^* \|^2 + \frac{1}{\eta_y} \| y^k - y^* \|^2.
\]

□

Back to proof of the Theorem (2.7).

Let $\Sigma^2 \triangleq \left( \frac{1}{L_x} + \frac{\omega}{L_{xy}} \right) \sigma_j^2 + \left( \frac{1}{L_y} + \frac{1}{L_y \omega} \right) \sigma_g^2$. Then

\[
\mathbb{E} \Psi^k \leq \theta^k \Psi^0 + \left( \frac{4}{1 - \theta} (\delta_x + \delta_y) + \frac{\Sigma^2}{2} \right) (1 + \theta + \theta^2 + \ldots) \\
\leq \theta^k \Psi^0 + \frac{4}{(1 - \theta)^2} (\delta_x + \delta_y) + \frac{\Sigma^2}{2(1 - \theta)}.
\]
\[ \theta^k \Psi^0 + \frac{4}{(1 - \theta)^2} (\delta_x + \delta_y) + \frac{\Sigma^2}{2(1 - \theta)} \geq \mathbb{E} \Psi^k \geq \frac{3}{4\eta_x} \mathbb{E} \| x^k - x^* \|^2 + \frac{1}{\eta_y} \mathbb{E} \| y^k - y^* \|^2. \]

Using the definitions of \( \eta_x \) and \( \eta_y \), we get

\[ \mathbb{E} \| x^k - x^* \|^2 \leq \frac{\omega}{3L_{xy}} \left( \theta^k \Psi^0 + \frac{4}{(1 - \theta)^2} (\delta_x + \delta_y) + \frac{\Sigma^2}{2(1 - \theta)} \right), \quad (24) \]

\[ \mathbb{E} \| y^k - y^* \|^2 \leq \frac{1}{4L_{xy}} \omega \left( \theta^k \Psi^0 + \frac{4}{(1 - \theta)^2} (\delta_x + \delta_y) + \frac{\Sigma^2}{2(1 - \theta)} \right). \quad (25) \]

Also for such definitions we know from Kovalev et al. (2021)

\[ \frac{1}{\rho_a} \leq 4 + 4 \max \left\{ \sqrt{\frac{L_x}{\mu_x}}, \frac{L_y}{\mu_y}, \sqrt{\frac{L_{xy}}{\mu_x \mu_y}} \right\} \]

for \( \omega = \frac{\mu_y}{\mu_x}, \sigma_x = \sqrt{\frac{\mu_x}{2L_x}}, \sigma_y = \frac{\mu_y}{2L_y}, \)

\[ \frac{1}{\rho_b} \leq 4 + 8 \max \left\{ \frac{\sqrt{L_x L_y}}{\mu_{xy}}, \frac{L_{xy}}{\mu_{xy}}, \frac{L_y^2}{\mu_y^2} \right\} \]

for \( \omega = \frac{\mu_x^2}{2\mu_x L_x}, \sigma_x = \sqrt{\frac{\mu_x}{2L_x}}, \sigma_y = \min \left\{ 1, \sqrt{\frac{\mu_{xy}^2}{4L_x L_y}} \right\}, \)

\[ \frac{1}{\rho_c} \leq 4 + 8 \max \left\{ \frac{\sqrt{L_x L_y}}{\mu_{xy}}, \frac{L_y^2}{\mu_y^2}, \frac{L_{xy}^2}{\mu_x^2} \right\} \]

for \( \omega = \frac{2\mu_x L_y}{\mu_{xy}^2}, \sigma_x = \sqrt{\frac{\mu_x^2}{4L_x L_y}}, \sigma_y = \sqrt{\frac{\mu_y}{2L_y}}, \)

\[ \frac{1}{\rho_d} \leq 2 + 8 \max \left\{ \frac{\sqrt{L_x L_y L_{xy}}}{\mu_{xy} \mu_{xy}}, \frac{L_{xy}^2}{\mu_{xy}^2}, \frac{L_{xy}^2}{\mu_x^2} \right\} \]

for \( \omega = \frac{\mu_{xy}}{\mu_{xy}}, \sigma_x = \frac{L_y}{L_x}, \sigma_y = \min \left\{ 1, \sqrt{\frac{\mu_{xy}^2}{4L_x L_y}} \right\}, \)

\[ \frac{1}{1 - \theta} = \min \{ \rho_a^{-1}, \rho_b^{-1}, \rho_c^{-1}, \rho_d^{-1} \}. \]

Note, that adding up batches and choosing \( \omega = \sqrt{\frac{\mu_x}{\mu}}, \sigma_x = \sqrt{\frac{\mu_x}{2L_x}}, \sigma_y = \sqrt{\frac{\mu_y}{2L_y}} \) proves the Theorem (2.7).
Rewriting inequalities in batch setting and assuming $\delta_x = \delta_y = 0$ we get

\[
\mathbb{E}\|x^k - x^*\|^2 \leq \frac{\omega}{3L_{xy}} \left( \theta^k \psi^0 + \frac{1}{2(1 - \theta)} \left( \left( \frac{1}{L_x} + \frac{\omega}{L_{xy}} \right) \sigma_f^2 + \left( \frac{1}{L_y} + \frac{1}{L_{xy} \alpha} \right) \sigma_g^2 \right) \right),
\]

\[
\mathbb{E}\|y^k - y^*\|^2 \leq \frac{1}{4L_{xy} \omega} \left( \theta^k \psi^0 + \frac{1}{2(1 - \theta)} \left( \left( \frac{1}{L_x} + \frac{\omega}{L_{xy}} \right) \sigma_f^2 + \left( \frac{1}{L_y} + \frac{1}{L_{xy} \alpha} \right) \sigma_g^2 \right) \right).
\]

Therefore, we can estimate the number of algorithm iterations $N = \mathcal{O}\left( \frac{1}{1 - \theta} \log \frac{C}{\varepsilon} \right)$, where $C$ is polynomial and not depend on $\varepsilon$. Rewriting it we obtain $N = \mathcal{O}\left( \min\{\rho_a^{-1}, \rho_b^{-1}, \rho_c^{-1}, \rho_d^{-1}\} \log \frac{C}{\varepsilon} \right)$.

It is sufficient to take such batch sizes $r_f = \left[ \frac{\max\{\alpha, \alpha^{-1}\}}{2L_{xy}(1 - \theta) \varepsilon} \left( \frac{1}{L_x} + \frac{\omega}{L_{xy}} \right) \sigma_f^2 \right]$, $r_g = \left[ \frac{\max\{\alpha, \alpha^{-1}\}}{2L_{xy}(1 - \theta) \varepsilon} \left( \frac{1}{L_x} + \frac{1}{L_{xy} \alpha} \right) \sigma_g^2 \right]$. Rewriting it with the selected constants

\[
r_f = \left[ \max \left\{ \frac{L_x}{\mu_x}, \sqrt{\frac{L_x}{\mu_x}}, \frac{L_{xy}}{\mu_x \mu_y} \right\} \frac{\mu}{2L_{xy} \sqrt{\mu_x \mu_y} \varepsilon} \left( \frac{1}{L_x} + \frac{1}{L_{xy} \sqrt{\mu_x} \mu_y} \right) \sigma_f^2 \right],
\]

\[
r_g = \left[ \max \left\{ \frac{L_x}{\mu_x}, \sqrt{\frac{L_x}{\mu_x}}, \frac{L_{xy}}{\mu_x \mu_y} \right\} \frac{\mu}{2L_{xy} \sqrt{\mu_x \mu_y} \varepsilon} \left( \frac{1}{L_y} + \frac{1}{L_{xy} \sqrt{\mu_x} \mu_y} \right) \sigma_g^2 \right],
\]

where $\mu = \max\{\mu_x, \mu_y\}$.

### Appendix B: Decentralized setting

Let us get an Algorithm 4 from the Algorithm (2), by multiplying every line by $\frac{1}{n}$, where $1$ is a column of $I$.

Algorithm 4 shows what happens to the average values at the nodes.

Supporting values $X$ and $Y$ to be in the neighborhood of $C(d_x)$ and $C(d_y)$ and using Lemma (2.1), conditions of Theorem (2.7) hold.

Using Assumption (1.5), Assumption (1.6), we get

\[
\mathbb{E}\|f_\delta(\bar{x}_k^k, \bar{x}_x^k) - \nabla f_\delta(\bar{x}_x^k)\|^2 \leq \frac{\sum_{i=1}^n \sigma_f^2 / f_{f,i}}{n^2} \leq \frac{\sigma_{F,t}^2}{n},
\]

\[
\mathbb{E}\|g_\delta(\bar{y}_k^k, \bar{y}_{y,k}^k) - \nabla g_\delta(\bar{y}_k^k)\|^2 \leq \frac{\sum_{i=1}^n \sigma_g^2 / g_{g,i}}{n^2} \leq \frac{\sigma_{G,t}^2}{n}.
\]

Let us support the number of iterations of Consensus to be sufficiently big to guarantee $\mathbb{E}\|X^k - \bar{X}^k\| \leq \sqrt{\delta^t}$ and $\mathbb{E}\|Y^k - \bar{Y}^k\| \leq \sqrt{\delta^t}$.

Introducing some definitions, which correspond to Lemma (2.1)
\[ \hat{\mu}_x = \frac{\mu_x}{2}, \quad \hat{\mu}_y = \frac{\mu_y}{2}. \]

Consider the iteration \( k \geq 1 \). Assuming that \( \|X^t - X^t\| \leq \sqrt{\delta'} \) and \( \|Y^t - Y^t\| \leq \sqrt{\delta'} \) for \( t = 0, 1, \ldots, k \), we are going to prove it for \( t = k + 1 \), using constant number of consensus iterations.

Using Line (10) and (6) of Algorithm 2, we get

\[ X^k_g = \tau_x X^k + (1 - \tau_x) X^k_f = (\tau_x + (1 - \tau_x) \sigma_x) X^k - (1 - \tau_x) \sigma_x X^{k-1} + (1 - \tau_x) X^{k-1}_g. \]

Define \( V^k = X^k_g - \sigma_x X^k \). Using \( X^0_g = X^0 \), we get

\[ V^0 = (1 - \sigma_x) X^0, \quad \|V^0 - V^0\| \leq (1 - \sigma_x) \sqrt{\delta'}. \]

\[ V^k = (1 - \sigma_x) \tau_x X^k + (1 - \tau_x) V^{k-1}, \quad \|V^k - V^k\| \leq (1 - \sigma_x) \tau_x \sqrt{\delta'} + (1 - \tau_x) \sqrt{\delta'} = (1 - \sigma_x) \sqrt{\delta'}. \]

Let us now estimate \( X^k_f, k \geq 1 \). Using Line (10), we get

\[ X^k_f = V^{k-1} + \sigma_x \overline{X^k}. \]

Let us now estimate \( X^k_g \) and \( Y^k_m \). Using Line (6) and (5), we get...
\[ \mathbb{E}\|X^k_g - X^k\| \leq \sqrt{\delta'}, \]
\[ \mathbb{E}\|Y^k_m - Y^k\| \leq (1 + 2\theta)\sqrt{\delta'}. \]

The estimations for \( Y^k_g \), \( Y^k_f \) are similar.

Let us now estimate \( \mathbb{E}\|U^{k+1} - \overline{U}^{k+1}\| \). Using Line (8), we get
\[
U^{k+1} - \overline{U}^{k+1} = (1 - \eta_x\alpha_x)(X^k - X^k) + \eta_x\alpha_x(X^k_g - \overline{X}^k_g)
- \eta_x\beta_x A^T\left( A(X^k - X^k) - \left( \nabla^r G(Y^k_g, \xi, k) - \overline{\nabla^r G}(Y^k_g, \xi, k) \right) \right)
- \eta_x\left( \nabla^r F(X^k_g, \xi, k) - \overline{\nabla^r F}(X^k_g, \xi, k) \right) + A^T\left( Y^k_m - \overline{Y}^k_m \right).
\]

Using that \( \eta_x\alpha_x \leq 1 \) and previous estimations, we get
\[
\mathbb{E}\|U^{k+1} - \overline{U}^{k+1}\| \leq (1 - \eta_x\alpha_x)\sqrt{\delta'} + \eta_x\alpha_x\sqrt{\delta'} + \eta_x\beta_x L_{xy}^2 \sqrt{\delta'}
+ \eta_x\beta_x L_{xy} \mathbb{E}\|\nabla^r G(Y^k_g, \xi, k)\| + \eta_x \mathbb{E}\|\nabla^r F(X^k_g, \xi, k)\| + \eta_x L_{xy} (1 + 2\theta) \sqrt{\delta'}
= (1 + \eta_x\beta_x L_{xy}^2 + \eta_x L_{xy} (1 + 2\theta)) \sqrt{\delta'} + \eta_x\beta_x L_{xy} \mathbb{E}\|\nabla^r G(Y^k_g, \xi, k)\|
+ \eta_x \mathbb{E}\|\nabla^r F(X^k_g, \xi, k)\|.
\]

Getting estimations for \( \mathbb{E}\|\nabla^r G(Y^k_g, \xi, k)\| \) and \( \mathbb{E}\|\nabla^r F(X^k_g, \xi, k)\| \):
\[
\mathbb{E}\|\nabla^r F(X^k_g, \xi, k)\| \leq \mathbb{E}\|\nabla^r F(X^k_g, \xi, k) - \nabla F(X^k_g)\| + \mathbb{E}\|\nabla F(X^k_g) - \nabla F(\overline{X}^k_g)\|
+ \mathbb{E}\|\nabla F(\overline{X}^k_g) - \nabla F(X^*)\| + \|\nabla F(X^*)\|
\leq \left( \mathbb{E}\|\nabla^r F(X^k_g, \xi, k) - \nabla F(X^k_g)\|^2 \right)^{\frac{1}{2}} + L_{lx} \mathbb{E}\|X^k_g - \overline{X}^k_g\|
+ L_{x} \mathbb{E}\|\overline{X}^k_g - X^*\| + \|\nabla F(X^*)\|
\leq \left( \sum_{i=1}^{n} \sigma_{f,i}^2 / r_{f,i} \right)^{\frac{1}{2}} + L_{lx} \sqrt{\delta'} + L_{x} \sqrt{n} \mathbb{E}\|X^k_g - X^*\| + \|\nabla F(X^*)\|.
\]

Let us define \( M_x \)
\[
M_x^2 = \frac{\omega}{3L_{xy}} \left( \frac{4}{(1 - \theta)^2} (\delta_x + \delta_y) + \frac{\Sigma^2}{2(1 - \theta)} \right),
\]
\[
\Sigma^2 = \left( \frac{1}{2L_x} + \frac{\omega}{L_{xy}} \right) \sigma_{f,r}^2 + \left( \frac{1}{2L_y} + \frac{1}{L_{xy} \omega} \right) \sigma_{G,r}^2.
\]

We choose constants the same as in Eqs. (24) and (25) for Algorithm 4.
Now we are going to estimate \( \|\bar{x}^k_g - x^*\| \). As we know from Eq. (24) and from Eq. (26)

\[
\begin{align*}
\mathbb{E}\|x^k - x^*\|^2 &\leq M_x^2, \\
\mathbb{E}\|x^k - x^*\| &\leq \sqrt{\mathbb{E}\|x^k - x^*\|^2} \leq M_x.
\end{align*}
\]

Let \( k \geq 1 \). Using Line (10) and (6) of Algorithm 4, we get

\[
\bar{x}^k_g = \tau_x^k \bar{x}^k + (1 - \tau_x^k) \bar{x}^k_f = (\tau_x + (1 - \tau_x) \sigma_x) \bar{x}^k - (1 - \tau_x) \sigma_x x^{k-1} + (1 - \tau_x) x_g^{k-1}.
\]

Let’s define \( \bar{v}^k = x^k_g - \sigma_x x^k \) and \( v^* = (1 - \sigma_x) x^* \). \( \bar{v}^0 = (1 - \sigma_x) x^0 \), therefore \( \|\bar{v}^0 - v^*\| \leq (1 - \sigma_x) M_x \).

\[
\bar{v}^k = \tau_x (1 - \sigma_x) \bar{x}^k + (1 - \tau_x) x_g^{k-1}.
\]

Firstly, we want to estimate \( \mathbb{E}\|\bar{v}^k - v^*\| \).

\[
\begin{align*}
\mathbb{E}\|\bar{v}^k - v^*\| &\leq \tau_x (1 - \sigma_x) \mathbb{E}\|\bar{x}^k - x^*\| + (1 - \tau_x) \mathbb{E}\|\bar{x}^{k-1} - v^*\| \\
&\leq (\tau_x (1 - \sigma_x) + (1 - \tau_x)(1 - \sigma_x)) M_x = (1 - \sigma_x) M_x.
\end{align*}
\]

Using Line (10), we get

\[
\begin{align*}
\bar{x}^k_f = \bar{v}^{k-1} + \sigma_x \bar{x}^k, \\
\mathbb{E}\|x^k_f - x^*\| &\leq \mathbb{E}\|\bar{v}^{k-1} - v^*\| + \sigma_x \mathbb{E}\|\bar{x}^k - x^*\| \leq (1 - \sigma_x) M_x + \sigma_x M_x = M_x.
\end{align*}
\]

Let’s estimate \( \mathbb{E}\|\bar{x}^k_g - x^*\| \). Using Line (5), we get

\[
\begin{align*}
\mathbb{E}\|\bar{x}^k_g - x^*\| &\leq \tau_x \mathbb{E}\|\bar{x}^k_g - x^*\| + (1 - \tau_x) \mathbb{E}\|x^k_f - x^*\| \leq M_x.
\end{align*}
\]

Returning to \( \mathbb{E}\|\nabla^r F(X_g^k, \xi_{x,k})\| \)

\[
\mathbb{E}\|\nabla^r F(X_g^k, \xi_{x,k})\| \leq \sqrt{\frac{n \sigma_{F,r}^2}{2} + L_x \sqrt{\delta'} + L_x \sqrt{n M_x} + \|\nabla F(x^*)\|}.
\]

Let’s define \( M_y \)

\[
M_y^2 = \frac{1}{4L_y \omega} \left( \Psi^0 + \frac{4}{(1 - \theta)^2} (\delta_x + \delta_y) + \frac{\Sigma^2}{2(1 - \theta)} \right).
\]

Then we can estimate \( \mathbb{E}\|\nabla^r G(Y_g^k, \xi_{y,k})\| \) in a similar way

\[
\mathbb{E}\|\nabla^r G(Y_g^k, \xi_{y,k})\| \leq \sqrt{\frac{n \sigma_{G,r}^2}{2} + L_y \sqrt{\delta'} + L_y \sqrt{h M_y} + \|\nabla G(y^*)\|}.
\]
Lemma B.1

\[
\max \left\{ \mathbb{E}\|U^{k+1} - U^{k+1}\|, \mathbb{E}\|W^{k+1} - W^{k+1}\| \right\} \leq D,
\]

where

\[
D = \max \left\{ D_{x,1} \sqrt{\delta' + D_{x,2}}, D_{y,1} \sqrt{\delta' + D_{y,2}} \right\},
\]

(27)

\[
D_{y,2} = \frac{L_{xy}}{2\mu_y} D_{x,2} + \frac{1}{2L_{xy}} \left( \sqrt{n\sigma^2_{F,r} + L_x \sqrt{n} M_x + \|\nabla F(X^*)\|} \right)
\]

\[
+ \frac{1}{2\mu_y} \left( \sqrt{n\sigma^2_{G,r} + L_y \sqrt{n} M_y + \|\nabla G(Y^*)\|} \right),
\]

(28)

\[
D_{y,1} = \frac{3}{2} + \frac{L_{xy}}{2\mu_y} D_{x,1} + \frac{L_{ly}}{2L_{xy}} + \frac{L_{by}}{2\mu_y},
\]

(29)

\[
D_{x,2} = \frac{1}{2L_{xy}} \left( \sqrt{n\sigma^2_{G,r} + L_y \sqrt{n} M_y + \|\nabla G(Y^*)\|} \right)
\]

\[
+ \frac{1}{2\mu_x} \left( \sqrt{n\sigma^2_{F,r} + L_x \sqrt{n} M_x + \|\nabla F(X^*)\|} \right),
\]

(30)

\[
D_{x,1} = \frac{3}{2} + \frac{L_{xy}}{2\mu_x} (1 + 2\theta) + \frac{L_{lx}}{2L_{xy}} + \frac{L_{lx}}{2\mu_x},
\]

(31)

\[
M^2_x = \frac{\omega}{3L_{xy}} \left( \Psi^0 + \frac{4}{(1 - \theta)^2} (\delta_x + \delta_y) + \frac{\Sigma^2}{2(1 - \theta)} \right),
\]

(32)

\[
M^2_y = \frac{1}{4L_{xy} \omega} \left( \Psi^0 + \frac{4}{(1 - \theta)^2} (\delta_x + \delta_y) + \frac{\Sigma^2}{2(1 - \theta)} \right),
\]

\[
\Sigma^2 = \left( \frac{1}{2L_x} + \frac{\omega}{L_{xy}} \right) \frac{\sigma^2_{F,r}}{n} + \left( \frac{1}{2L_y} + \frac{1}{L_{xy} \omega} \right) \frac{\sigma^2_{G,r}}{n}.
\]

(33)
Proof

\[
\mathbb{E}\|U^{k+1} - \overline{U}^{k+1}\| \leq (1 + \eta_x \beta_x L_{xy}^2 + \eta_x L_{xy}(1 + 2\theta)) \sqrt{\delta'}
\]
\[
+ \eta_x \beta_x L_{xy} \mathbb{E}\|\nabla^r G(Y^k_g, \xi_{y,k})\| + \eta_x \mathbb{E}\|\nabla^r F(X^k_g, \xi_{x,k})\|.
\]

Using the definition of \( \eta_x, \beta_x \) and estimations on gradients, we get

\[
\mathbb{E}\|U^{k+1} - \overline{U}^{k+1}\| \leq \left( 1 + \frac{1}{2} + \frac{L_{xy}}{4\mu_x}(1 + 2\theta) \right) \sqrt{\delta'} + \frac{1}{2L_{xy}} \mathbb{E}\|\nabla^r G(Y^k_g, \xi_{y,k})\|
\]
\[
+ \frac{1}{4\mu_x} \mathbb{E}\|\nabla^r F(X^k_g, \xi_{x,k})\|
\]
\[
\leq \left( \frac{3}{2} + \frac{L_{xy}}{4\mu_x}(1 + 2\theta) + \frac{L_{ly}}{2L_{xy}} + \frac{L_{lx}}{4\mu_x} \right) \sqrt{\delta'}
\]
\[
+ \frac{1}{2L_{xy}} \left( \sqrt{n\sigma_{G,r}^2} + L_y \sqrt{nM_y} + \|\nabla G(Y^*)\| \right)
\]
\[
+ \frac{1}{4\mu_x} \left( \sqrt{n\sigma_{F,r}^2} + L_x \sqrt{nM_x} + \|\nabla F(X^*)\| \right) = D_{x,1} \sqrt{\delta'} + D_{x,2}.
\]

Let’s estimate \( \mathbb{E}\|W^{k+1} - \overline{W}^{k+1}\| \)

\[
W^{k+1} - \overline{W}^{k+1} = (1 - \eta_y \alpha_y) \left( Y^k - \overline{Y}^k \right) + \eta_y \alpha_y \left( Y^k_g - \overline{Y}_g^k \right)
\]
\[
- \eta_y \beta_y A \left( A^T \left( Y^k - \overline{Y}^k \right) + \left( \nabla^r F(X^k_g, \xi_{x,k}) - \overline{\nabla^r F}(X^k_g, \xi_{x,k}) \right) \right)
\]
\[
- \eta_y \left( \nabla^r G(Y^k_g, \xi_{y,k}) - \overline{\nabla^r G}(Y^k_g, \xi_{y,k}) \right) - A \left( U^{k+1} - \overline{U}^{k+1} \right).
\]

Using that \( \eta_y \alpha_y \leq 1 \) and previous estimations, we get

\[
\mathbb{E}\|W^{k+1} - \overline{W}^{k+1}\| \leq (1 - \eta_y \alpha_y) \sqrt{\delta'} + \eta_y \alpha_y \sqrt{\delta'} + \eta_y \beta_y L_{xy}^2 \sqrt{\delta'}
\]
\[
+ \eta_y \beta_y L_{xy} \mathbb{E}\|\nabla^r F(X^k_g, \xi_{x,k})\|
\]
\[
+ \eta_y \mathbb{E}\|\nabla^r G(Y^k_g, \xi_{y,k})\| + \eta_y L_{xy} \mathbb{E}\|U^{k+1} - \overline{U}^{k+1}\|
\]
\[
= (1 + \eta_y \beta_y L_{xy}^2) \sqrt{\delta'} + \eta_y L_{xy} \mathbb{E}\|U^{k+1} - \overline{U}^{k+1}\|
\]
\[
+ \eta_y \beta_y L_{xy} \mathbb{E}\|\nabla^r F(X^k_g, \xi_{x,k})\| + \eta_y \mathbb{E}\|\nabla^r G(Y^k_g, \xi_{y,k})\|.
\]

Using the definition of \( \beta_y, \eta_y \) and gradient estimations, we get
Now let us estimate the number of communication steps $T$.

$$\left(1 - \lambda \right)^{T/\tau} \max \{ \mathbb{E} \| W^{k+1} - W^{k+1} \|, \mathbb{E} \| U^{k+1} - U^{k+1} \| \} \leq \delta'.$$

It would we sufficient to guarantee

$$\left(1 - \lambda \right)^{T/\tau} D \leq \delta'. $$

Above inequality leads from this

$$T \geq \frac{\tau}{\lambda} \log \left( \frac{D}{\delta'} \right).$$

**Putting the proof together**

Using Lemma (2.1), we get

$$\mathbb{E} \left\| x^k - x^\ast \right\|^2 \leq \frac{\omega}{3L_{xy}} \left( \theta^k \Psi^0 + \frac{4}{(1 - \theta)^2} (\delta_x + \delta_y) + \frac{\Sigma^2}{2(1 - \theta)} \right),$$

$$\mathbb{E} \left\| y^k - y^\ast \right\|^2 \leq \frac{1}{4L_{xy} \omega} \left( \theta^k \Psi^0 + \frac{4}{(1 - \theta)^2} (\delta_x + \delta_y) + \frac{\Sigma^2}{2(1 - \theta)} \right).$$

Define several notations

$$\nu = \max \left\{ \frac{1}{3L_{xy} \omega}, \frac{1}{4L_{xy} \omega^{-1}} \right\}.$$
\[ \psi^0 = \frac{1}{\eta_x} \| x^0 - x^\ast \| + \frac{1}{\eta_y} \| x^0 - y^\ast \| + \frac{2}{\sigma_x} B_f(x^0, x^\ast) + \frac{2}{\sigma_y} B_g(y^0, y^\ast), \]

where \( \eta_x = \min \left\{ \frac{1}{4(\hat{\mu}_x + L_x \sigma_x)}, \frac{\omega}{4L_{xy}} \right\}, \eta_y = \min \left\{ \frac{1}{4(\hat{\mu}_y + L_y \sigma_y)}, \frac{1}{4L_{xy} \omega} \right\}. \)

Rewriting it in terms of \( L_{lx}, L_x, L_{ly}, L_y, \mu_{lx}, \mu_y, \mu_{lx}, \mu_{ly}, \mu_y. \)

\[ \eta_x = \min \left\{ \frac{1}{2\hat{\mu}_x + 8L_x \sigma_x}, \frac{\omega}{4L_{xy}} \right\}, \eta_y = \min \left\{ \frac{1}{2\hat{\mu}_y + 8L_y \sigma_y}, \frac{1}{4L_{xy} \omega} \right\}. \]

Finally, let us estimate the right part

\[ \delta_x, \delta_y \leq \frac{(1 - \theta)^2 \varepsilon}{24 \nu}. \]

Define \( E \) as

\[ E = \frac{1}{2n} \max \left\{ \frac{L^2_x}{\mu_x} + \frac{2L^2_{xy}}{\mu_y} + L_{lx} - \mu_{lx}, \frac{L^2_y}{\mu_y} + \frac{2L^2_{xy}}{\mu_x} - L_{ly} - \mu_{lx} \right\}. \]

Using definition of \( \delta_x \) and \( \delta_y, \) we get

\[ \delta' = \frac{(1 - \theta)^2 \varepsilon}{24 E \nu}. \]

Define \( F_x \) and \( F_y \) as

\[ F_x = \frac{\nu}{2n(1 - \theta)} \left( \frac{1}{L_x} + \frac{\omega}{L_{xy}} \right), F_y = \frac{\nu}{2n(1 - \theta)} \left( \frac{1}{L_y} + \frac{1}{L_{xy} \omega} \right). \]

Using the definitions of \( \Sigma^2, \sigma^2_{F_x}, \sigma^2_{F_y}, \sigma^2_{G_x}, \sigma^2_{G_y} \) we get, that it would be sufficient to take

\[ r_{f,i} = \left[ \frac{6F_x \sigma^2_{f,i}}{\varepsilon} \right] \text{ and } r_{g,i} = \left[ \frac{6F_y \sigma^2_{g,i}}{\varepsilon} \right]. \]

Finally
Decentralized saddle-point problems with different constants…

\[ N_{\text{comm}} = NT = O\left( \frac{1}{1 - \theta} \sqrt{\frac{\log \left( \psi_{\nu}^{D^t} \right)}{\epsilon}} \log \left( \frac{D^t}{\epsilon} \right) \right), \]

\[ N_{\text{comp}}^t = N(r_{i,t} + r_{i,s}) = 2N + O\left( \max \{ 1, \frac{1}{nL_{xy}(1 - \theta)^2 \epsilon} \left( \left( \frac{1}{L_x} + \frac{\omega}{L_{xy}} \right) \sigma_{f,i}^2 + \left( \frac{1}{L_y} + \frac{1}{L_{xy} \omega} \right) \sigma_{g,i}^2 \right) \log \left( \frac{\psi_{\nu}^{D^t}}{\epsilon} \right) \right). \]

\[ \frac{1}{1 - \theta} = O\left( \max \left\{ \sqrt{\frac{L_x}{\mu_x}}, \sqrt{\frac{L_y}{\mu_y}}, \sqrt{\frac{L_{xy}}{\sqrt{\mu_x \mu_y}}} \right\} \right), \]

\[ \omega = \sqrt{\frac{\mu_y}{\mu_x}}, \sigma_x = \sqrt{\frac{\mu_x}{8L_x}}, \sigma_y = \sqrt{\frac{\mu_y}{8L_y}}, \]

\[ \frac{1}{1 - \theta} = O\left( \max \left\{ \sqrt{\frac{L_{xy}}{\mu_{xy}}}, \frac{L_{xy}}{\mu_{xy}}, \frac{L_x L_y}{\mu_{xy} L_{xy}} \right\} \right), \]

\[ \omega = \sqrt{\frac{2\mu_y L_y}{\mu_{xy}^2}}, \sigma_x = \min \left\{ 1, \sqrt{\frac{\mu_{xy}^2}{16L_x L_y}} \right\}, \sigma_y = \sqrt{\frac{\mu_y}{8L_y}}, \]

\[ \frac{1}{1 - \theta} = O\left( \max \left\{ \sqrt{\frac{L_x L_y}{\mu_{xy}^2}}, \frac{L_x L_y}{\mu_{xy}}, \frac{L_{xy}^2}{\mu_{xy} \mu_{xy}} \right\} \right), \]

\[ \omega = \frac{\mu_{xy}}{L_x L_y}, \sigma_x = \min \left\{ 1, \sqrt{\frac{\mu_{xy}^2}{16L_x L_y}} \right\}, \sigma_y = \min \left\{ 1, \sqrt{\frac{\mu_{xy}^2}{16L_x L_y}} \right\}. \]

References

Alkousa MS, Gasnikov AV, Dvinskikh DM, Kovalev DA, Stonyakin FS (2020) Accelerated methods for saddle-point problem. Computat Math Math Phys 60(11):1787–1809

Beznosikov A, Gorbunov E, Berard H, Loizou N (2022) Stochastic gradient descent-ascent: unified theory and new efficient methods. arXiv preprint arXiv:2202.07262

Beznosikov A, Rogozin A, Kovalev D, Gasnikov A (2021) Near-optimal decentralized algorithms for saddle point problems over time-varying networks. In: International conference on optimization and applications. Springer, pp 246–257

Beznosikov A, Samokhin V, Gasnikov A (2020) Distributed saddle-point problems: lower bounds, optimal and robust algorithms. arXiv preprint arXiv:2010.13112

Boyd S, Parikh N, Chu E, Peleato B, Eckstein J (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach Learn 3(1):1–122

Devolder O, Glineur F, Nesterov Y (2014) First-order methods of smooth convex optimization with inexact oracle. Math Program 146(1):37–75
Wang Y, Li J (2020) Improved algorithms for convex–concave minimax optimization. Adv Neural Inf Process Syst 33:4800–4810
Yarmoshik D, Rogozin A, Khamisov O, Dvurechensky P, Gasnikov A et al (2022) Decentralized convex optimization under affine constraints for power systems control. arXiv preprint arXiv:2203.16686
Zhang X, Aybat NS, Gurbuzbalaban M (2021) Robust accelerated primal-dual methods for computing saddle points. arXiv preprint arXiv:2111.12743
Zhang J, Hong M, Zhang S (2019) On lower iteration complexity bounds for the saddle point problems. arXiv preprint arXiv:1912.07481

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.