Communication Acceleration of Local Gradient Methods via an Accelerated Primal-Dual Algorithm with Inexact Prox

Abdurakhmon Sadiev[*]
KAUST[†], Saudi Arabia
MIPT[‡], Russia
ISP RAS[§], Russia
abdurakhmon.sadiev@kaust.edu.sa

Dmitry Kovalev
KAUST , Saudi Arabia
dmitry.kovalev@kaust.edu.sa

Peter Richtárik
KAUST, Saudi Arabia
peter.richtarik@kaust.edu.sa

Abstract

Inspired by a recent breakthrough of Mishchenko et al. [2022], who for the first time showed that local gradient steps can lead to provable communication acceleration, we propose an alternative algorithm which obtains the same communication acceleration as their method (ProxSkip). Our approach is very different, however: it is based on the celebrated method of Chambolle and Pock [2011], with several nontrivial modifications: i) we allow for an inexact computation of the prox operator of a certain smooth strongly convex function via a suitable gradient-based method (e.g., GD, Fast GD or FSFOM), ii) we perform a careful modification of the dual update step in order to retain linear convergence. Our general results offer the new state-of-the-art rates for the class of strongly convex-concave saddle-point problems with bilinear coupling characterized by the absence of smoothness in the dual function. When applied to federated learning, we obtain a theoretically better alternative to ProxSkip: our method requires fewer local steps ($\mathcal{O}(\kappa^{1/3})$ or $\mathcal{O}(\kappa^{1/4})$, compared to $\mathcal{O}(\kappa^{1/2})$ of ProxSkip), and performs a deterministic number of local steps instead. Like ProxSkip, our method can be applied to optimization over a connected network, and we obtain theoretical improvements here as well.

1 Introduction

Communication efficiency of distributed stochastic gradient descent (SGD) can be improved, often dramatically, via a simple trick: instead of synchronizing the parameters across the parallel workers after every SGD step, let the workers perform multiple optimization steps using their local loss and data only before synchronizing. This trick dates back at least to two or three decades ago [Mangasarian, 1995], and may be much older. Due to its simplicity, it has been repeatedly rediscovered [Povey et al., 2014, Moritz et al., 2016, Ma et al., 2017]. It is the basis of the famous federated averaging (FedAvg)

[*]This work was written while A. Sadiev was a research intern at KAUST during the last semester of his MS studies at the Moscow Institute of Physics and Technology, Dolgoprudny, Russia.

[†]King Abdullah University of Science and Technology

[‡]Moscow Institute of Physics and Technology

[§]The Ivannikov Institute for System Programming of the Russian Academy of Sciences, Research Center for Trusted Artificial Intelligence

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

algorithm of McMahan et al. [2016, 2017], which is the workhorse of federated learning [Konečný et al., 2016b,a]; see also the recent surveys on federated learning [Li et al., 2020a, Kairouz et al., 2019] and federated optimization [Wang et al., 2021b].

1.1 Towards provable communication acceleration via delayed parameter synchronization

Until recently, this simple trick resisted all attempts at an appropriate theoretical justification. Through collective effort of the federated learning community, the bounds of various local SGD methods were progressively getting better [Haddadpour and Mahdavi, 2019, Li et al., 2019a,b, Khaled et al., 2019a,b, 2020, Li et al., 2020b, Woodworth et al., 2020, Stich, 2020, Gorbunov et al., 2020a, Malinovsky et al., 2020, Pathak and Wainwright, 2020, Karimireddy et al., 2020, Malinovsky et al., 2021, Mishchenko et al., 2021a, Wang and Joshi, 2021, Horváth et al., 2022], and the assumptions required to achieve them weaker. A brief overview of the progress is provided in [Mishchenko et al., 2022]. However, all known theoretical rates are worse than the rate of gradient descent, which synchronizes after every gradient step. In a recent breakthrough, Mishchenko et al. [2022] developed a novel local SGD method, called ProxSkip, which performs a random number of local gradient (or stochastic gradient) steps before synchronization, and proved that it enjoys strong communication acceleration properties. In particular, while the method needs $\mathcal{O}(\kappa \log \frac{1}{\varepsilon})$ iterations, only $\mathcal{O}(\sqrt{\kappa} \log \frac{1}{\varepsilon})$ of them involve communication.

1.2 Problem formulation

In this paper, we consider the composite optimization problem

$$\min_{x \in \mathbb{R}^{d_x}} G(x) + F(Kx), \tag{1}$$

where $G : \mathbb{R}^{d_x} \to \mathbb{R}$ is a smooth and strongly convex function, $F : \mathbb{R}^{d_y} \to \mathbb{R} \cup \{+\infty\}$ is a proper, closed and convex function, and $K : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ is a linear map. Let us define

$$L_{xy} \overset{\text{def}}{=} \max \left\{ \|Kx\| : \ x \in \mathbb{R}^{d_x}, \|x\| = 1 \right\}, \tag{2}$$

where $\|\cdot\|$ refers to the standard Euclidean norm. Note that $L_{xy}^2 \geq \lambda_{\max}(KK^\top) = \lambda_{\max}(K^\top K)$.

It will be useful to formalize our assumptions at this point as we will refer to the various constants involved in them throughout the text.

Assumption 1. *Function $G : \mathbb{R}^{d_x} \to \mathbb{R}$ is μ_x-strongly convex, i.e.,*

$$G(x') - G(x'') - \langle \nabla G(x''), x' - x'' \rangle \geq \frac{\mu_x}{2} \|x' - x''\|^2, \qquad \forall x', x'' \in \mathbb{R}^{d_x}. \tag{3}$$

Assumption 2. *The function $G : \mathbb{R}^{d_x} \to \mathbb{R}$ is L_x-smooth, i.e.*

$$\|\nabla G(x') - \nabla G(x'')\| \leq L_x \|x' - x''\|, \qquad \forall x', x'' \in \mathbb{R}^{d_x}. \tag{4}$$

Assumption 3. *Function $F : \mathbb{R}^{d_y} \to \mathbb{R} \cup \{+\infty\}$ is proper, closed and convex.*

Assumption 4 (See Kovalev et al. [2021a])**.** *There exists a constant $\mu_{xy} > 0$ such that*

$$\mu_{xy}^2 \leq \begin{cases} \lambda_{\min}^+(KK^\top), & \text{if } \partial F^\star(y) \in \operatorname{range} K \text{ for all } y \in \mathbb{R}^{d_y}, \\ \lambda_{\min}(KK^\top), & \text{otherwise.} \end{cases}$$

1.3 ProxSkip

The most general version of the ProxSkip method[5] of Mishchenko et al. [2022] was designed to solve problems of the form (1). In each iteration, ProxSkip evaluates the gradient of G and then flips a biased coin: with probability p, it additionally evaluates the proximity operator of $F^\star$, and performs a matrix-vector multiplication involving K. The method becomes relevant to the standard optimization formulation of federated learning (FL), i.e., to the finite-sum optimization problem

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^n f_i(x), \tag{5}$$

[5]This variant is called SplitSkip in their paper.

Table 1: Summary of the key complexity results obtained by our methods APDA (Algorithm 1; Theorem 1) and APDA with Inexact Prox (Algorithm 2; Theorem 2) for solving the saddle-point problem (7).

Algorithm	No Prox G	No Prox $F^\star$	Works with $L_y = \infty$	Linear rate with $\mu_y = 0$	# Outer Iterations[(1)]	# Inner Iterations[(1)]
CP[(a)]	✗	✗	✓	✗	$\frac{L_{xy}}{\sqrt{\mu_x \mu_y}}$[(2)]	uses prox of G
AltGDA[(b)]	✓	✓	✗	✓	$\max\left\{\frac{L}{\mu_x}, \frac{L^2}{\mu_{xy}^2}\right\}$	—[(3)]
APDG[(c)]	✓	✓	✗	✓	$\max\left\{\frac{\sqrt{L_x L_y}}{\mu_{xy}}, A_{xy}\right\}$	—[(3)]
Alg 1	✗	✗	✓	✓	A_{xy}[(2)]	uses prox of G
Alg 2	✓	✗	✓	✓	A_{xy}[(2)]	$\max\left\{\kappa_x \kappa_{xy}, \kappa_x^{1/2}\kappa_{xy}^2\right\}$[(4)]
Alg 2	✓	✗	✓	✓	A_{xy}[(2)]	$\max\left\{\kappa_x^{5/6}\kappa_{xy}, \kappa_x^{1/3}\kappa_{xy}^2\right\}$[(5)]
Alg 2	✓	✗	✓	✓	A_{xy}[(2)]	$\max\left\{\kappa_x^{3/4}\kappa_{xy}, \kappa_x^{1/4}\kappa_{xy}^2\right\}$[(6)]

[(a)] Chambolle and Pock [2011] assume that G and $F^\star$ are μ_x and μ_y-strongly-convex, respectively. We do not assume $F^\star$ to be strongly convex.
[(b)] Zhang et al. [2022] assume that the functions G and $F^\star$ are L-smooth (i.e., $L = \max\{L_x, L_y, L_{xy}\}$), and that G is μ_x-strongly-convex.
[(c)] Kovalev et al. [2021a] assume that the functions G and $F^\star$ are L_x and L_y-smooth, respectively, and that G is μ_x-strongly-convex.
[(1)] For brevity, we let $\kappa_{xy} \overset{\text{def}}{=} \frac{L_{xy}}{\mu_{xy}}$, $\kappa_x \overset{\text{def}}{=} \frac{L_x}{\mu_x}$, and $A_{xy} \overset{\text{def}}{=} \max\left\{\kappa_x^{1/2}\kappa_{xy}, \kappa_{xy}^2\right\}$. We omit constant factors and a $\log \frac{1}{\varepsilon}$ factor in all expressions, for brevity. So, for example, the expression A_{xy} in the case of the method of our methods should be interpreted as $\mathcal{O}\left(A_{xy} \log \frac{1}{\varepsilon}\right)$.
[(2)] # outer iterations = # evaluations of the prox of $F^\star$.
[(3)] There is no prox operator and hence no inner iterations.
[(4)] The iterative method $\mathcal{M}$ evaluating the prox of G inexactly in this case is: $\mathcal{M}$ = GD (see Lemma 1).
[(5)] The iterative method $\mathcal{M}$ evaluating the prox of G inexactly in this case is: $\mathcal{M}$ = FGD (Fast Gradient Descent) + GD. See Lemma 1.
[(6)] The iterative method $\mathcal{M}$ evaluating the prox of G inexactly in this case is: $\mathcal{M}$ = FSFOM + FGD (Fast Gradient Descent). See Lemma 1.

through its application to its consensus reformulation

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \left\{ \sum_{i=1}^n f_i(x_i) + \psi(x_1,\ldots,x_n) \right\} = \min_{x \in \mathbb{R}^{d_x}} G(x) + F(x), \tag{6}$$

where $d_x \overset{\text{def}}{=} nd$, $x \overset{\text{def}}{=} (x_1,\ldots,x_n) \in \mathbb{R}^{nd}$, $G(x) \overset{\text{def}}{=} \sum_{i=1}^n f_i(x_i)$, and $F \overset{\text{def}}{=} \psi$ is the indicator function of the constraint $x_1 = \cdots = x_n$, i.e.,

$$\psi(x_1,\ldots,x_n) \overset{\text{def}}{=} \begin{cases} 0 & \text{if} \quad x_1 = \cdots = x_n \\ +\infty & \text{otherwise} \end{cases}.$$

The evaluation of the proximity operator of F is equivalent to averaging of the vectors $x_1,\ldots,x_n$, which necessitates communication. Therefore, if p is small, ProxSkip communicates very rarely. Since G is block separable, the gradient steps involving G, taken in between two communications, correspond to gradient steps with respect to the local loss functions $\{f_i\}$ taken by the clients. See [Mishchenko et al., 2022] for the details; we will elaborate on this as well in Section 6.

ProxSkip inexactly solves problem (1) in $\mathcal{O}\left(\kappa \log {}^1\!/_\varepsilon\right)$ iterations, out of which only $\mathcal{O}\left(\sqrt{\kappa\chi} \log \frac{1}{\varepsilon}\right)$ involve communication, where χ is a condition number measuring the connectivity of the graph (the standard setup in FL corresponds to a fully connected graph, in which case $\chi = 1$; see (26) for definition).

2 Summary of Contributions

Inspired by the results of Mishchenko et al. [2022], we propose an alternative and substantially different algorithm which obtains the same guarantees for the number of prox evaluations (wrt F) as ProxSkip, but has better guarantees for the number gradient steps (wrt G) in between the prox evaluations. Below we summarize the main contributions:

Saddle-point formulation. Unlike Mishchenko et al. [2022], we consider the saddle-point reformulation of (1)

$$\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} \left\{G(x) + \langle y, Kx\rangle - F^\star(y)\right\}, \tag{7}$$

where $F^\star(y) \stackrel{\text{def}}{=} \sup_{y'\in\mathbb{R}^{d_y}} \{\langle y, y'\rangle - F(y')\}$ is the convex conjugate of F. Since F is proper, closed and convex, so is $F^\star$. We assume throughout that (7) is solvable, i.e., there exists at least one solution $(x^\star, y^\star)$. Such a solution then satisfies the first-oder optimality conditions[6]

$$0 \in \partial G(x^\star) + K^\top y^\star, \qquad 0 \in \partial F^\star(y^\star) - Kx^\star, \tag{8}$$

where ∂ denotes the subdifferential. By working with this reformulation, we can tap into the rich and powerful philosophical and technical toolbox offered by proximal-point theory, fixed point theory, and primal-dual methods, which facilitates the algorithm development and analysis. This ultimately enables us to shed new light on the nature of local gradient-type steps as inexact computations of the prox operator of G in a new appropriately designed Accelerated Primal-Dual Algorithm (APDA; see Algorithm 1).

Modifications of Chambolle-Pock Our Algorithms 1 and 2 are inspired by the celebrated Chambolle-Pock method [Chambolle and Pock, 2011], but with several important and nontrivial modifications. While Chambolle-Pock achieves linear convergence when both G and $F^\star$ are strongly convex, $F^\star$ is merely convex in our setting. Our modifications are:

i) Inspired by the ideas of Kovalev et al. [2021a], we perform a careful modification of the dual update step (update of y) in order to retain linear convergence despite lack of strong convexity in $F^\star$. On the other hand, in contrast to the method of Kovalev et al. [2021a], we do not assume $F^\star$ to be smooth. This modification leads to a new method, which we call APDA (Algorithm 1). APDA relies on the evaluation of the prox operators of both G and $F^\star$.
ii) Next, we remove the reliance on the prox operator of G, and instead allow for its inexact evaluation via a suitable user-defined gradient-based method, which we call $\mathcal{M}$ (see (2) and Lemma 1). We call the resulting method "APDA with Inexact Prox" (Algorithm 2). The choice of method $\mathcal{M}$ will have a strong impact on the number of inexact/local steps, and this is one of the places in which we improve upon the results of Mishchenko et al. [2022].

General theory Our general complexity results for Algorithms 1 and 2, covered in Theorems 7 and 2, respectively, contrasted with the key baselines, are summarized in Table 1). While the method of [Chambolle and Pock, 2011] needs $F^\star$ to be strongly convex to obtain a linear rate, we only need convexity. While the AltGDA [Zhang et al., 2022] and APDG [Kovalev et al., 2021a] methods enjoy linear rates without strong convexity of $F^\star$, both require $F^\star$ to be L_y-smooth. In contrast, we do not need this assumption (i.e., we allow $L_y = \infty$). Our methods are the first to obtain linear convergence rates in the regime when G is L_x-smooth and μ_x-strongly-convex, and $F^\star$ is merely proper, closed and convex, without requiring it to be L_y-smooth, nor μ_y-strongly-convex. This is important in some applications. Our two methods offer two alternative ways of dealing with this regime: while APDA (Algorithm 1) relies on the evaluation of the prox of G, APDA with Inexact Prox (Algorithm 2) does not. As we shall see, the latter method has an important application in federated learning.

Federated learning and a third method When applied to the distributed/federated problem (5) (see Section 6), APDA with Inexact Prox (Algorithm 2) turns out to be a theoretically better alternative to ProxSkip [Mishchenko et al., 2022]. In the centralized case, our method requires the same optimal number of communication rounds ($\widetilde{\mathcal{O}}(\sqrt{\kappa})$, where $\kappa = {}^{L_x}/_{\mu_x}$) as ProxSkip, but requires fewer local gradient-type steps ($\mathcal{O}(\kappa^{1/3})$ or $\mathcal{O}(\kappa^{1/4})$, compared to $\mathcal{O}(\kappa^{1/2})$ of ProxSkip, depending on the choice of the inner method $\mathcal{M}$). Like ProxSkip, our method can be applied to optimization over a connected network, and we obtain theoretical improvements in this decentralized scenario as well. However, in the decentralized regime, neither ProxSkip nor Algorithm 2 obtain the optimal bound for the number of communication rounds. For this reason, we propose a third method (Algorithm 5) which employs an accelerated gossip routine to remedy this situation. It is also notable that while ProxSkip uses a random number of local steps, all our methods perform a deterministic number of local steps. Our complexity results are summarized in Table 2.

[6]Whenever we invoke Assumption 2 (L_x-smoothness of G), we have $\partial G(x) = \{\nabla G(x)\}$, and hence the first condition can be replaced by $0 = \nabla G(x^\star) + K^\top y^\star$.

Table 2: Summary of our general convergence results provided by Theorem 2 for Algorithm 2 (APDA with Inexact Prox) and Theorem 3 for Algorithm 5 (APDA with Inexact Prox and Accelerated Gossip) for solving the saddle-point reformulation (25) of the federated learning problem (5).

Algorithm	Method $\mathcal{M}$[(2)] for Inexact Prox	Deter-ministic # comm. rounds	Centralized case		Decentralized case	
			Optimal #comm. rounds?	#Local steps per round	Optimal #comm. rounds?	#Local steps per round
ProxSkip [Mishchenko et al., 2022]	GD	✗	✓	$\mathcal{O}\left(\sqrt{\kappa}\right)$	✓[(1)]	$\widetilde{\mathcal{O}}\left(\sqrt{\kappa}\right)$[(1)]
Alg 2; Thm 2	GD	✓	✓	$\widetilde{\mathcal{O}}\left(\sqrt{\kappa}\right)$	✗	$\widetilde{\mathcal{O}}\left(\sqrt{\kappa}\right)$
	FGD+GD	✓	✓	$\widetilde{\mathcal{O}}\left(\sqrt[3]{\kappa}\right)$	✗	$\widetilde{\mathcal{O}}\left(\sqrt[3]{\kappa}\right)$
	FGD+FSFOM	✓	✓	$\widetilde{\mathcal{O}}\left(\sqrt[4]{\kappa}\right)$	✗	$\widetilde{\mathcal{O}}\left(\sqrt[4]{\kappa}\right)$
Alg 5; Thm 3	GD	✓	✓	$\widetilde{\mathcal{O}}\left(\sqrt{\kappa}\right)$	✓	$\widetilde{\mathcal{O}}\left(\sqrt{\kappa}\right)$
	FGD+GD	✓	✓	$\widetilde{\mathcal{O}}\left(\sqrt[3]{\kappa}\right)$	✓	$\widetilde{\mathcal{O}}\left(\sqrt[3]{\kappa}\right)$
	FGD+FSFOM	✓	✓	$\widetilde{\mathcal{O}}\left(\sqrt[4]{\kappa}\right)$	✓	$\widetilde{\mathcal{O}}\left(\sqrt[4]{\kappa}\right)$

[(1)] This is true only when $\kappa \leq \chi$.
[(1)] GD = Gradient Descent; FGD = Fast Gradient Descent (i.e., Nesterov's accelerated GD); FSFOM = a fixed-step first-order method from [Kim and Fessler, 2021].

3 From Proximal Point Method to Chambolle-Pock

In this section, we briefly motivate the development of the celebrated Chambolle-Pock method which acts as a starting point of our algorithm design.

3.1 Proximal-Point Method for finding zeros of monotone operators

Our starting point is the general problem of finding a zero of an (set-valued) operator $A : \mathcal{H} \to 2^{\mathcal{H}}$, where $\mathcal{H}$ is a Hilbert space, i.e., find $z \in \mathcal{H}$ such that

$$0 \in A(z). \tag{9}$$

If A is maximally monotone, its resolvent $(\mathrm{Id} + \eta A)^{-1}$ is single valued, nonexpansive, and has full domain. Moreover, $0 \in A(z)$ iff $z = (\mathrm{Id} + \eta A)^{-1}(z)$. The corresponding fixed point iteration, i.e., $z^{k+1} = (\mathrm{Id} + \eta A)^{-1}(z^k)$, is called the proximal point method (PPM) [Rockafellar, 1976]. This can be equivalently written as $z^k \in (\mathrm{Id} + \eta A)(z^{k+1})$, and subsequently as

$$z^{k+1} \in z^k - \eta A(z^{k+1}).$$

From now on, for simplicity only, we will ignore the fact that in general, $A(z^{k+1})$ is a set, and will write $z^{k+1} = z^k - \eta A(z^{k+1})$ instead to mean the same thing, i.e., that there exists $u \in A(z^{k+1})$ such that $z^{k+1} = z^k - \eta u$.

3.2 PPM applied to the saddle-point problem

The optimality conditions (8) of the saddle point problem (7) can be written in the form (9) with $z = (x; y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ as follows[7]:

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in A \begin{pmatrix} x \\ y \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} \partial G(x) + K^\top y \\ \partial F^\star(y) - Kx \end{pmatrix}. \tag{10}$$

Allowing for different stepsizes η_x, η_y for each block of the vector $z = (x; y)$, PPM applied to (10) takes the form

$$\begin{aligned} x^{k+1} &\in x^k - \eta_x \left(\partial G(x^{k+1}) + K^\top y^{k+1}\right) \\ y^{k+1} &\in y^k - \eta_y \left(\partial F^\star(y^{k+1}) - K x^{k+1}\right). \end{aligned}$$

The main advantage of this method is its unboundedly fast convergence rate under weak assumptions. According to Theorem 4, the proof of which we provide in the appendix for completeness, if G and

[7] We replaced ∇G by ∂G here as the beginning of our story does not require G to be smooth.

Algorithm 1 APDA

Input: Initial point $(x^0, y^0) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, $\bar{y}^0 = y^0$; Step sizes $\eta_x, \eta_y, \beta_y > 0$, $\theta \in [0,1]$
for $k = 0, 1, \ldots$ **do**
  $x^{k+1} = x^k - \eta_x \left(\nabla G(x^{k+1}) + K^\top \bar{y}^k\right)$
  $y^{k+1} \in y^k - \eta_y \left(\partial F^\star(y^{k+1}) - K x^{k+1}\right) - \eta_y \beta_y K \left(K^\top y^k + \nabla G(x^{k+1})\right)$
  $\bar{y}^{k+1} = y^{k+1} + \theta \left(y^{k+1} - y^k\right)$
end for

$F^\star$ are proper and closed, G is μ_x strongly convex and $F^\star$ is μ_y strongly convex, then any choice of stepsizes $\eta_x > 0$ and $\eta_y > 0$ (yes, without an upper bound!), PPM find an ε-accurate solution in

$$\mathcal{O}\left(\left(1 + \frac{1}{\min\{\eta_x \mu_x, \eta_y \mu_y\}}\right) \log \frac{1}{\varepsilon}\right) \tag{11}$$

iterations. Unfortunately, PPM is not implementable since in order to compute x^{k+1}, we need to know y^{k+1}, and vice versa.

3.3 Chambolle-Pock: Making PPM implementable, and fast

In order to overcome the above problem, Chambolle and Pock [2011] proposed to replace y^{k+1} with y^k (see Algorithm 1 in [Chambolle and Pock, 2011]), which leads to

$$\begin{aligned} x^{k+1} &\in x^k - \eta_x \left(\partial G(x^{k+1}) + K^\top y^k\right) \\ y^{k+1} &\in y^k - \eta_y \left(\partial F^\star(y^{k+1}) - K x^{k+1}\right). \end{aligned}$$

Although this method is implementable, it has its own disadvantages, one of which is its weak iteration complexity bound

$$\mathcal{O}\left(\frac{L_{xy}^2}{\mu_x \mu_y} \log \frac{1}{\varepsilon}\right). \tag{12}$$

Chambolle and Pock [2011] proposed to fix this problem via an extrapolation step of the dual variable (see Algorithm 3 in [Chambolle and Pock, 2011]):

$$\begin{aligned} x^{k+1} &\in x^k - \eta_x \left(\partial G(x^{k+1}) + K^\top \bar{y}^k\right) \\ y^{k+1} &\in y^k - \eta_y \left(\partial F^\star(y^{k+1}) - K x^{k+1}\right) \\ \bar{y}^{k+1} &= y^{k+1} + \theta(y^{k+1} - y^k). \end{aligned}$$

This new method enjoys the much better iteration complexity bound

$$\mathcal{O}\left(\frac{L_{xy}}{\sqrt{\mu_x \mu_y}} \log \frac{1}{\varepsilon}\right). \tag{13}$$

4 Accelerated Primal-Dual Algorithm (Algorithm 1)

Recall that the Chambolle-Pock method requires $F^\star$ to be strongly-convex to obtain a linear convergence rate. However, in our setting, $F^\star$ is not strongly convex[8] (see Assumption 3), and Chambolle-Pock method does not converge linearly in this scenario.

4.1 Modifying Chambolle-Pock to preserve linear rate without strong convexity of $F^\star$

To obtain a linear rate, we modify the dual update step of the algorithm using a trick proposed by Kovalev et al. [2021a] that was shown to work in the regime when $F^\star$ is smooth; the innovation here is that we do not need this assumption (see Table 1). From this point onwards, we will also need to

[8] We would need to assume F to be smooth to ensure that $F^\star$ is strongly convex. However, we do not want to do this as this is not satisfied in many scenarios, in particular, in our key application to federated learning.

assume G to be L_x-smooth (see Assumption 2). In particular, we propose to modify the update step for y^{k+1} in the Chambolle-Pock method as follows:

$$\begin{aligned} x^{k+1} &= x^k - \eta_x \left(\nabla G(x^{k+1}) + K^\top \bar{y}^k\right) \\ y^{k+1} &\in y^k - \eta_y \left(\partial F^\star(y^{k+1}) - K x^{k+1}\right) - \eta_y \beta_y K \left(K^\top y^k + \nabla G(x^{k+1})\right) \\ \bar{y}^{k+1} &= y^{k+1} + \theta(y^{k+1} - y^k). \end{aligned}$$

This is a new method, which we call APDA (formalized as Algorithm 1).

4.2 APDA converges linearly

Our first result shows that APDA indeed converges linearly, without the need for $F^\star$ to be strongly convex.

Theorem 1 (Convergence of APDA; informal)**.** *Let Assumptions 1, 2,3 and 4 hold. Then, with a suitable selection of stepsizes, APDA (Algorithm 1) solves problem* (7) *in*

$$\mathcal{O}\left(\max\left\{\sqrt{\frac{L_x}{\mu_x}}\frac{L_{xy}}{\mu_{xy}}, \frac{L_{xy}^2}{\mu_{xy}^2}\right\}\log\frac{1}{\varepsilon}\right) \tag{14}$$

iterations.

The formal statement and proof can be found in the appendix (see Theorem 7).

5 Accelerated Primal-Dual Algorithm with Inexact Prox (Algorithm 2)

The key disadvantage of APDA is that it requires the evaluations of the proximity operator of G, which can be very expensive in some applications. To remedy the situation, we first notice that Step 3 of APDA can be equivalently written in the form

$$x^{k+1} = \arg\min_{x\in\mathbb{R}^{d_x}} \left\{\Psi^k(x) \overset{\text{def}}{=} G(x) + \frac{1}{2\eta_x}\left\|x - \left(x^k - \eta_x K^\top \bar{y}^k\right)\right\|^2\right\}; \tag{15}$$

that is, this step involves the evaluation of the prox of G. The key idea of this section is to replace this by an inexact prox computation via a suitably selected iterative method $\mathcal{M}$ (this is the method performing the inner iterations in Table 1). This leads to our next method: APDA with Inexact Prox (Algorithm 2).

Algorithm 2 APDA with Inexact Prox

1: **Input**: Initial point $(x^0, y^0) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, $\bar{y}^0 = y^0$; Step sizes $\eta_x, \eta_y, \beta_y > 0$, $\theta \in [0,1]$; # inner iterations T
2: **for** $k = 0, 1, \ldots$ **do**
3: Find $\hat{x}^k$ as a final point of T iteration of some method $\mathcal{M}$ for following problem:

$$\hat{x}^k \approx \arg\min_{x\in\mathbb{R}^{d_x}} \left\{\Psi^k(x) \overset{\text{def}}{=} G(x) + \frac{1}{2\eta_x}\left\|x - \left(x^k - \eta_x K^\top \bar{y}^k\right)\right\|^2\right\}$$

4: $x^{k+1} = x^k - \eta_x \left(\nabla G(\hat{x}^k) + K^\top \bar{y}^k\right)$
5: $y^{k+1} \in y^k - \eta_y \left(\partial F^\star(y^{k+1}) - K\hat{x}^k\right) - \eta_y \beta_y K \left(K^\top y^k + \nabla G(\hat{x}^k)\right)$
6: $\bar{y}^{k+1} = y^{k+1} + \theta\left(y^{k+1} - y^k\right)$
7: **end for**

5.1 Gradient methods for finding a stationary point of convex functions

A key feature of Algorithm 2 is its reliance on a subroutine $\mathcal{M}$ for an inexact evaluation of the prox of G via solving the auxiliary problem (15). Our theory requires the method $\mathcal{M}$ to be able to output, after T iterations, a point $\hat{x}^k$ such that

$$\|\nabla\Psi^k(\hat{x}^k)\|^2 \le \mathcal{O}\left(\frac{1}{T^\alpha}\right), \tag{16}$$

where $\alpha \geq 2$. In other words, we require a reduction of the squared norm of the gradient with a fast sublinear rate. In the next lemma, we present three examples of such methods.

Lemma 1. *Let $\Psi : \mathbb{R}^{d_x} \to \mathbb{R}$ be an L-smooth convex function, and let $w^\star$ be a minimizer of Ψ. Then there exists a gradient-based method $\mathcal{M}$ which after T iterations outputs a point w^T satisfying*

$$\|\nabla \Psi(w^T)\|^2 \leq \frac{AL^2\|w^0 - w^\star\|^2}{T^\alpha}, \tag{17}$$

for all starting points $x^0 \in \mathbb{R}^{d_x}$ and some universal constant $A > 0$. In particular,
(i) if $\mathcal{M}$ is GD, then $\|\nabla \Psi(w^T)\|^2 \leq \frac{4L^2\|w^0-w^\star\|^2}{T^2}$,
(ii) if $\mathcal{M}$ is a combination[9] of Fast GD [Nesterov, 2004] and GD, then $\|\nabla \Psi(w^T)\|^2 \leq \frac{64L^2\|w^0-w^\star\|^2}{T^3}$,
(iii) if $\mathcal{M}$ is a combination[10] of Fast GD [Nesterov, 2004] and FSFOM [Kim and Fessler, 2021], then $\|\nabla \Psi(w^T)\|^2 \leq \frac{256L^2\|w^0-w^\star\|^2}{T^4}$.

Let $w^{\star k} = \arg\min_{w \in \mathbb{R}^{d_x}} \Psi^k(w)$. Since Ψ^k is $\left(L_x + \eta_x^{-1}\right)$-smooth, Lemma 1 implies that T iterations of a method $\mathcal{M}$ satisfying (17) applied to the auxiliary problem (15) with starting point $w^0 = x^k$ yield point $w^T = \hat{x}^k$ for which

$$\|\nabla \Psi^k(\hat{x}^k)\|^2 \leq \frac{A\left(\eta_x^{-1} + L_x\right)^2 \|x^k - w^{\star k}\|^2}{T^\alpha} = \frac{A\left(1 + \eta_x L_x\right)^2 \|x^k - w^{\star k}\|^2}{\eta_x^2 T^\alpha}. \tag{18}$$

5.2 APDA with Inexact Prox converges linearly

Now we can provide the main theorem with the total complexity of gradient computation ∇G and proximity operator computation $\partial F^\star$.

Theorem 2. *Let Assumptions 1, 2, 3, 4 hold. Then there exist parameters of Algorithm 2 such that the total # of evaluations of prox $F^\star$ and the total number evaluations of the gradient of ∇G to find an ε solution of problem* (7) *are*

$$\mathcal{O}\left(\max\left\{\sqrt{\frac{L_x}{\mu_x}}\frac{L_{xy}}{\mu_{xy}}, \frac{L_{xy}^2}{\mu_{xy}^2}\right\}\log\frac{1}{\varepsilon}\right), \ \mathcal{O}\left(\max\left\{\left(\frac{L_x}{\mu_x}\right)^{\frac{2+\alpha}{2\alpha}}\frac{L_{xy}}{\mu_{xy}}, \sqrt[\alpha]{\frac{L_x}{\mu_x}}\frac{L_{xy}^2}{\mu_{xy}^2}\right\}\log\frac{1}{\varepsilon}\right), \tag{19}$$

respectively. In particular,
(i) if the inner method $\mathcal{M}$ is GD, then the total number of ∇G computations is equal to

$$\mathcal{O}\left(\max\left\{\frac{L_x}{\mu_x}\frac{L_{xy}}{\mu_{xy}}, \sqrt{\frac{L_x}{\mu_x}}\frac{L_{xy}^2}{\mu_{xy}^2}\right\}\log\frac{1}{\varepsilon}\right), \tag{20}$$

(ii) if the inner method $\mathcal{M}$ is combination of Fast GD and GD, then the total number of ∇G computations is equal to

$$\mathcal{O}\left(\max\left\{\left(\frac{L_x}{\mu_x}\right)^{\frac{5}{6}}\frac{L_{xy}}{\mu_{xy}}, \sqrt[3]{\frac{L_x}{\mu_x}}\frac{L_{xy}^2}{\mu_{xy}^2}\right\}\log\frac{1}{\varepsilon}\right), \tag{21}$$

(iii)if the inner method $\mathcal{M}$ is combination of Fast GD and FSFOM, then the total number of ∇G computations is equal to

$$\mathcal{O}\left(\max\left\{\left(\frac{L_x}{\mu_x}\right)^{\frac{3}{4}}\frac{L_{xy}}{\mu_{xy}}, \sqrt[4]{\frac{L_x}{\mu_x}}\frac{L_{xy}^2}{\mu_{xy}^2}\right\}\log\frac{1}{\varepsilon}\right). \tag{22}$$

The proof relies on several lemmas; their statements and the proof of the theorem can be found in the appendix.

[9] The first half of the iterations is solved via the Fast Gradient Descent (FGD) method of Nesterov [2004], and the second half via Gradient Descent (GD).

[10] The first half of the iterations is solved via the Fast Gradient Descent (FGD) method of Nesterov [2004], and the second half via the fixed-step first-order method (FSFOM) of Kim and Fessler [2021].

Note that our way of performing inexact computation of prox of G allows us to keep the same complexity as APDA (Algorithm 1) in terms of the number of evaluations of the prox of $F^\star$. When $\mathcal{M}$ is chosen to be Gradient Decent ($\alpha = 2$), the number of computations of the gradient of ∇G, given by (20), is larger than the number of computations of the prox of G for APDA. Fortunately, we can reduce this if GD is replaced with a faster method. For example, if we choose $\mathcal{M}$ to be a simple combination of Fast GD (FGD) and GD, in which case $\alpha = 3$, Theorem 2 says that the number of computations of the gradient of ∇G is can be reduced to (21) (this choice is mentioned in the second-to-last row of Table 2). A further reduction is possible if we instead employ a combination of FSFOM and Fast GD; see (22) and the last row of Table 2.

6 Accelerated Primal Dual Algorithm with Inexact Prox and Accelerated Gossip (Algorithm 5)

In this section, we consider the most significant applications of Algorithm 2: decentralized optimization over a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and federated learning. In particular, we consider the finite-sum optimization problem

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^{n} f_i(x)$$

(see (5)) interpreted as follows: $n = |\mathcal{V}|$ is the number of clients/agents in the network. Communication can only happen between clients connected by an edge. The prevalent paradigm in federated learning, where a single server orchestrates communication in rounds, arises as a special case of this with $\mathcal{G}$ being the fully connected network.

If $\hat{W}$ is the Laplacian[11] of graph $\mathcal{G}$, and let $W = \hat{W} \otimes I_{dn}$. Then problem (5) can be rewritten in following equivalent way:

$$\min_{\sqrt{W}\mathbf{x}=0} P(\mathbf{x}) = \min_{\mathbf{x} \in \mathbb{R}^{dn}} P(\mathbf{x}) + \psi(\sqrt{W}\mathbf{x}), \tag{23}$$

where $\mathbf{x}^\top = (x_1^\top, x_2^\top, \ldots, x_n^\top)$, $P(\mathbf{x}) = \sum_{i=1}^n f_i(x_i)$ and $\psi(\mathbf{x}) = 0$ iff $\mathbf{x} = 0$, and $\psi(\mathbf{x}) = +\infty$ otherwise. Problem (23) is a special case of (1). By dualizing the nonsmooth (but proper, closed and convex) penalty, we get the equivalent saddle point formulation

$$\min_{\sqrt{W}\mathbf{x}=0} P(\mathbf{x}) = \min_{\mathbf{x} \in \mathbb{R}^{dn}} \max_{\mathbf{y} \in \mathbb{R}^{dn}} \left\{ P(\mathbf{x}) + \langle \mathbf{y}, \sqrt{W}\mathbf{x} \rangle - \psi^\star(\mathbf{y}) \right\}, \tag{24}$$

As we can see, this problem (24) is the particular case of the problem (7). It means that we can solve it by Algorithm 2, for example. Moreover, we do not have to compute the prox of $\psi^\star$ due to the fact that $\psi^\star(\cdot) \equiv 0$, because $\psi(\cdot)$ is the indicator function of $\{0\}$. We thus arrive at the final formulation

$$\min_{\mathbf{x} \in \mathbb{R}^{dn}} \max_{\mathbf{y} \in \mathbb{R}^{dn}} \left\{ P(\mathbf{x}) + \langle \mathbf{y}, \sqrt{W}\mathbf{x} \rangle \right\}. \tag{25}$$

6.1 Application of Algorithm 2 to (25)

Before providing the complexity results related to the application of Algorithm 2 to problem (25), note that $L_{xy} = \sqrt{\lambda_{\max}(W)}$ and $\mu_{xy} = \sqrt{\lambda_{\min}^+(W)}$ and define

$$\chi \overset{\text{def}}{=} \frac{\lambda_{\max}(W)}{\lambda_{\min}^+(W)}. \tag{26}$$

According to Theorem 2, the total number of evaluations of the prox of $\psi^\star$, i.e., the communication complexity, and the total number of evaluations of ∇P, i.e., computation complexity, are

$$\sharp\text{comm} = \widetilde{\mathcal{O}}\left(\max\left\{\sqrt{\kappa\chi}, \chi\right\}\right), \quad \sharp\text{comp} = \widetilde{\mathcal{O}}\left(\max\left\{\kappa^{\frac{2+\alpha}{2\alpha}}\sqrt{\chi},\ \sqrt[\alpha]{\kappa}\chi\right\}\right), \tag{27}$$

respectively. For example, in centralized case, when $\mathcal{G}$ is the complete graph ($\chi = 1$) and $\mathcal{M}$ is chosen to be GD ($\alpha = 2$), we obtain the same complexities as ProxSkip [Mishchenko et al., 2022]:

$$\sharp\text{comm} = \widetilde{\mathcal{O}}\left(\sqrt{\kappa}\right), \qquad \sharp\text{comp} = \widetilde{\mathcal{O}}\left(\kappa\right). \tag{28}$$

[11] In fact, it is enough for $\hat{W}$ to satisfy the less restrictive Assumption 6.

However, this can be improved by using a more elaborate subroutine $\mathcal{M}$. If instead of GD we use either FGD + GD ($\alpha = 3$) or FGD + FSFOM ($\alpha = 4$) in place of $\mathcal{M}$, the number of communication rounds will be the same as in the case of ProxSkip (or Algorithm 2 used with $\mathcal{M}$ = GD). However, the total number of gradient computations gets improved to $\sharp\text{comp} = \widetilde{\mathcal{O}}\left(\kappa^{\frac{5}{6}}\right)$ in the first case, and to $\sharp\text{comp} = \widetilde{\mathcal{O}}\left(\kappa^{\frac{3}{4}}\right)$ in the second case.

6.2 Improvement on Algorithm 2 via accelerated gossip

Compared to the ProxSkip computation complexity $\widetilde{\mathcal{O}}\left(\sqrt{\kappa\chi}\right)$, in the general decentralized case (i.e., $\chi > 1$), complexity (27) of our Algorithm 2 is worse if $\kappa \leq \chi$. To tackle this problem, we propose to enhance Algorithm 2 using the accelerated gossip technique [Scaman et al., 2017]. Based on this approach, we propose one more (and final) method: Algorithm 5. For this method, we prove the following result.

Theorem 3. *Let Assumptions 1 and 2 hold for function P. Then there exist parameters of Algorithm 5 such that in order to find an ε-solution of problem* (7)*, the total number of communications and the total number of gradient computations can be bounded by*

$$\sharp comm = \widetilde{\mathcal{O}}\left(\sqrt{\kappa\chi}\right), \qquad \sharp comp = \widetilde{\mathcal{O}}\left(\kappa^{\frac{2+\alpha}{2\alpha}}\right), \tag{29}$$

respectively. In particular, if the inner method $\mathcal{M}$ is
(i) GD, then the number of local steps is equal to $\widetilde{\mathcal{O}}\left(\kappa^{1/2}\right)$ and the total number of gradient computations is equal to $\widetilde{\mathcal{O}}\left(\kappa\right)$;
(ii) FGD + GD, then the number of local steps is equal to $\widetilde{\mathcal{O}}\left(\kappa^{1/3}\right)$ and the total number of gradient computations is equal to $\widetilde{\mathcal{O}}\left(\kappa^{5/6}\right)$;
(iii)FGD + FSFOM, then the number of local steps is equal to $\widetilde{\mathcal{O}}\left(\kappa^{1/4}\right)$ and the total number of gradient computations is equal to $\widetilde{\mathcal{O}}\left(\kappa^{3/4}\right)$.

The communication complexities obtained this way are substantially better than those of decentralized ProxSkip; see Table 3 and the commentary in the next subsection.

Recall that in Table 2, we already compared ProxSkip and our Algorithm 2 in the centralized case. In Table 3 we add to this a comparison in the decentralized case, and include our Algorithm 5 as well. In particular, in Table 3 we compare the complexity results of our methods (Algorithms 2 and 5) for solving the decentralized optimization problem (5) to two selected benchmarks: D-SGD [Koloskova et al., 2020] and ProxSkip [Mishchenko et al., 2022].

First, observe that ProxSkip has vastly superior communication complexity to D-SGD, both in the centralized case (i.e., for fully-connected network; $\chi = 1$), where the improvement is from $\mathcal{O}(\kappa)$ to $\widetilde{\mathcal{O}}(\sqrt{\kappa})$, and the decentralized case ($\chi > 1$), where the improvement is from $\widetilde{\mathcal{O}}(\kappa\chi)$ to $\widetilde{\mathcal{O}}(\sqrt{\kappa\chi})$.

In the decentralized case, both our methods match the $\widetilde{\mathcal{O}}(\sqrt{\kappa\chi})$ communication complexity of ProxSkip. However, Algorithm 2 does so only when $\sqrt{\kappa\chi} \geq \chi$ (i.e., when $\kappa > \chi$). On the other hand, both our methods have an improved bound on the number of local gradient computations, depending on what subroutine $\mathcal{M}$ they employ. The improvement is from $\widetilde{\mathcal{O}}(\sqrt{\kappa})$ to $\widetilde{\mathcal{O}}\left(\sqrt[3]{\kappa}\right)$ (when $\mathcal{M} =$ FGD + GD) to $\widetilde{\mathcal{O}}\left(\sqrt[4]{\kappa}\right)$ (when $\mathcal{M} =$ FGD + FSFOM).

Acknowledgments and Disclosure of Funding

The work of all authors was supported by the KAUST Baseline Research Grant awarded to P. Richtárik. The work of A. Sadiev was supported by the Visiting Student Research Program (VSRP) at KAUST. The work of A. Sadiev was also partially 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.

References

D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

D. Alistarh, T. Hoefler, M. Johansson, S. Khirirat, N. Konstantinov, and C. Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems*, 2018.

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

L. Condat, K. Yi, and P. Richtárik. EF-BV: A unified theory of error feedback and variance reduction mechanisms for biased and unbiased compression in distributed optimization. *arXiv preprint arXiv:2205.04180*, 2022.

I. Fatkhullin, I. Sokolov, E. Gorbunov, Z. Li, and P. Richtárik. Ef21 with bells & whistles: practical algorithmic extensions of modern error feedback. *arXiv preprint arXiv:2110.03294*, 2021.

E. Gorbunov, F. Hanzely, and P. Richtárik. Local SGD: unified theory and new efficient methods. In *The 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020a.

E. Gorbunov, F. Hanzely, and P. Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. In *The 23rd International Conference on Artificial Intelligence and Statistics*, 2020b.

E. Gorbunov, D. Kovalev, D. Makarenko, and P. Richtárik. Linearly converging error compensated SGD. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020c.

E. Gorbunov, K. Burlachenko, Z. Li, and P. Richtárik. MARINA: Faster non-convex distributed learning with compression. In *38th International Conference on Machine Learning*, 2021.

F. Haddadpour and M. Mahdavi. On the convergence of local descent methods infederated learning. *arXiv preprint arXiv:1910.14425*, 2019.

F. Hanzely, K. Mishchenko, and P. Richtárik. SEGA: variance reduction via gradient sketching. In *Advances in Neural Information Processing Systems 31*, pages 2082–2093, 2018.

S. Horváth and P. Richtárik. A better alternative to error feedback for communication-efficient distributed learning. In *9th International Conference on Learning Representations (ICLR)*, 2021.

S. Horváth, C.-Y. Ho, Ľudovít Horváth, A. N. Sahu, M. Canini, and P. Richtárik. Natural compression for distributed deep learning. *arXiv preprint arXiv:1905.10988*, 2019a.

S. Horváth, D. Kovalev, K. Mishchenko, S. Stich, and P. Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*, 2019b.

S. Horváth, M. Sanjabi, L. Xiao, P. Richtárik, and M. Rabbat. FedShuffle: Recipes for better use of local work in federated learning. *arXiv preprint arXiv:2204.13169*, 2022.

R. Islamov, X. Qian, and P. Richtárik. Distributed second order methods with fast rates and compressed communication. In *International Conference on Machine Learning*, 2021.

R. Islamov, X. Qian, S. Hanzely, M. Safaryan, and P. Richtárik. Distributed Newton-type methods with communication compression and bernoulli aggregation. *arXiv preprint arXiv:2206.03588*, 2022.

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

S. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. Suresh. SCAFFOLD: Stochastic controlled averaging for on-device federated learning. In *The 27th International Conference on Machine Learning (ICML)*, 2020.

A. Khaled, K. Mishchenko, and P. Richtárik. First analysis of local GD on heterogeneous data. In *NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality*, pages 1–11, 2019a.

A. Khaled, K. Mishchenko, and P. Richtárik. Better communication complexity for local SGD. In *NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality*, pages 1–11, 2019b.

A. Khaled, K. Mishchenko, and P. Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, 2020.

D. Kim and J. A. Fessler. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *Journal of Optimization Theory and Applications (JOTA)*, 188(1): 192–219, 2021.

A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning (ICML)*, pages 5381–5393. PMLR, 2020.

J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik. Federated optimization: distributed machine learning for on-device intelligence. *arXiv:1610.02527*, 2016a.

J. Konečný, H. B. McMahan, F. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: strategies for improving communication efficiency. In *NIPS Private Multi-Party Machine Learning Workshop*, 2016b.

D. Kovalev, A. Salim, and P. Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. *Advances in Neural Information Processing Systems*, 33: 18342–18352, 2020.

D. Kovalev, A. Gasnikov, and P. Richtárik. Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling. *arXiv preprint arXiv:2112.15199*, 2021a.

D. Kovalev, A. Koloskova, M. Jaggi, P. Richtárik, and S. Stich. A linearly convergent algorithm for decentralized optimization: Sending less bits for free! In *The 24th International Conference on Artificial Intelligence and Statistics (AISTATS 2021)*, 2021b.

T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization for heterogeneous networks. In *Proceedings of the 1st Adaptive & Multitask Learning Workshop*, 2019a.

T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated learning: challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a. doi: 10.1109/MSP.2020.2975749.

X. Li, W. Yang, S. Wang, and Z. Zhang. Communication-efficient local decentralized SGD methods. *arXiv preprint arXiv:1910.09126*, 2019b.

X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the conver-gence of FedAvg on non-IID data. In *International Conference on Learning Representations (ICLR)*, 2020b.

Z. Li, D. Kovalev, X. Qian, and P. Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. In *International Conference on Machine Learning*, 2020c.

C. Ma, J. Konečný, M. Jaggi, V. Smith, M. I. Jordan, P. Richtárik, and M. Takáč. Distributed optimization with arbitrary local solvers. *Optimization Methods and Software*, 32(4):813–848, 2017.

G. Malinovsky, D. Kovalev, E. Gasanov, L. Condat, and P. Richtárik. From local SGD to local fixed point methods for federated learning. In *International Conference on Machine Learning (ICML)*, 2020.

G. Malinovsky, K. Mishchenko, and P. Richtárik. Server-side stepsizes and sampling without replacement provably help in federated optimization. *arXiv preprint arXiv:2201.11066*, 2021.

O. L. Mangasarian. Parallel gradient distribution in unconstrained optimization. *SIAM Journal on Control and Optimization*, 33(6):1916–1925, 1995.

B. McMahan, E. Moore, D. Ramage, and B. Agüera y Arcas. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*, 2016.

H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

K. Mishchenko, E. Gorbunov, M. Takáč, and P. Richtárik. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019.

K. Mishchenko, F. Hanzely, and P. Richtárik. 99% of worker-master communication in distributed optimization is not needed. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124, pages 979–988, 2020.

K. Mishchenko, A. Khaled, and P. Richtárik. Proximal and federated random reshuffling. *arXiv preprint arXiv:2102.06704*, 2021a.

K. Mishchenko, D. Kovalev, B. Wang, and P. Richtárik. Intsgd: Floatless compression of stochastic gradients. *preprint*, 2021b.

K. Mishchenko, G. Malinovsky, S. Stich, and P. Richtárik. ProxSkip: Yes! Local gradient steps provably lead to communication acceleration! Finally! In *39th International Conference on Machine Learning (ICML 2022)*, 2022.

P. Moritz, R. Nishihara, I. Stoica, and M. I. Jordan. SparkNet: Training deep networks in Spark. In *International Conference on Learning Representations (ICLR)*, 2016.

Y. Nesterov. *Introductory lectures on convex optimization: a basic course (Applied Optimization)*. Kluwer Academic Publishers, 2004.

R. Pathak and M. J. Wainwright. FedSplit: An algorithmic framework for fast federated optimization. *arXiv preprint arXiv:2005.05238*, 2020.

C. Philippenko and A. Dieuleveut. Bidirectional compression in heterogeneous settings for distributed or federated learning with partial participation: tight convergence guarantees. *arXiv preprint arXiv:2006.14591*, 2020.

D. Povey, X. Zhang, and S. Khudanpur. Parallel training of DNNs with natural gradient and parameter averaging. In *ICLR 2015 Workshop (arXiv preprint arXiv:1410.7455)*, 2014.

X. Qian, P. Richtárik, and T. Zhang. Error compensated distributed SGD can be accelerated. *arXiv preprint arXiv:2010.00091*, 2020.

X. Qian, R. Islamov, M. Safaryan, and P. Richtárik. Basis matters: better communication-efficient second order methods for federated learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

P. Richtárik, I. Sokolov, and I. Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. In *Advances in Neural Information Processing Systems 34*, 2021.

P. Richtárik, I. Sokolov, I. Fatkhullin, E. Gasanov, Z. Li, and E. Gorbunov. 3pc: Three point compressors for communication-efficient distributed training and a better theory for lazy aggregation. *arXiv preprint arXiv:2202.00998*, 2022.

R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

M. Safaryan, F. Hanzely, and P. Richtárik. Smoothness matrices beat smoothness constants: better communication compression techniques for distributed optimization. In *ICLR Workshop: Distributed and Private Machine Learning*, 2021.

M. Safaryan, R. Islamov, X. Qian, and P. Richtárik. FedNL: Making Newton-type methods applicable to federated learning. In *Internatioanl Conference on Machine Learning*, 2022.

K. Scaman, F. Bach, S. Bubeck, Y. T. Lee, and L. Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *International Conference on Machine Learning (ICML)*, pages 3027–3036. PMLR, 2017.

E. Shulgin and P. Richtárik. Shifted compression framework: Generalizations and improvements. In *OPT2021: 13th Annual Workshop on Optimization for Machine Learning*, 2021.

S. U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations (ICLR)*, 2020.

S. U. Stich and S. P. Karimireddy. The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.

S. U. Stich, J.-B. Cordonnier, and M. Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems*, 2018.

R. Szlendak, A. Tyurin, and P. Richtárik. Permutation compressors for provably faster distributed nonconvex optimization. In *10th International Conference on Learning Representations*, 2022.

H. Tang, C. Yu, X. Lian, T. Zhang, and J. Liu. Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6155–6165, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/tang19d.html`.

A. Tyurin and P. Richtárik. Distributed nonconvex optimization with communication compression, optimal oracle complexity, and no client synchronization. *arXiv preprint arXiv:2202.01268*, 2022.

T. Vogels, S. P. Karimireddy, and M. Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. In *Neural Information Processing Systems*, 2019.

B. Wang, M. Safaryan, and P. Richtárik. Smoothness-aware quantization techniques. *arXiv preprint arXiv:2106.03524*, 2021a.

H. Wang, S. Sievert, Z. Charles, D. Papailiopoulos, and S. Wright. Atomo: Communication-efficient learning via atomic sparsification. *arXiv preprint arXiv:1806.04090*, 2018.

J. Wang and G. Joshi. Cooperative SGD: A unified framework for the design andanalysis of local-update SGD algorithms. *Journal of Machine Learning Research (JMLR)*, 21:1–50, 2021.

J. Wang, Z. Charles, Z. Xu, G. Joshi, H. B. McMahan, B. A. y Arcas, M. Al-Shedivat, G. Andrew, S. Avestimehr, K. Daly, D. Data, S. Diggavi, H. Eichner, A. Gadhikar, Z. Garrett, A. M. Girgis, F. Hanzely, A. Hard, C. He, S. Horvath, Z. Huo, A. Ingerman, M. Jaggi, T. Javidi, P. Kairouz, S. Kale, S. P. Karimireddy, J. Konečný, S. Koyejo, T. Li, L. Liu, M. Mohri, H. Qi, S. J. Reddi, P. Richtárik, K. Singhal, V. Smith, M. Soltanolkotabi, W. Song, A. T. Suresh, S. U. Stich, A. Talwalkar, H. Wang, B. Woodworth, S. Wu, F. X. Yu, H. Yuan, M. Zaheer, M. Zhang, T. Zhang, C. Zheng, C. Zhu, and W. Zhu. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021b.

J. Wangni, J. Wang, J. Liu, and T. Zhang. Gradient sparsification for communication-efficient distributed optimization. *arXiv preprint arXiv:1710.09854*, 2017.

B. Woodworth, K. K. Patel, S. U. Stich, Z. Dai, B. Bullins, H. B. McMahan, O. Shamir, and N. Srebro. Is local SGD better than minibatch SGD? *arXiv preprint arXiv:2002.07839*, 2020.

H. Xu, C.-Y. Ho, A. M. Abdelmoniem, A. Dutta, E. H. Bergou, K. Karatsenidis, M. Canini, and P. Kalnis. Compressed communication for distributed deep learning: Survey and quantitative evaluation. Technical report, KAUST, 2020.

G. Zhang, Y. Wang, L. Lessard, and R. B. Grosse. Near-optimal local convergence of alternating gradient descent-ascent for minimax optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 7659–7679. PMLR, 2022.

C. Zhu, S. Han, H. Mao, and W. J. Dally. Trained ternary quantization. In *ICLR*, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We crearly state all theorem
   (b) Did you describe the limitations of your work? [Yes] Assumptions are our limitations.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work is mainly theoretical, with no foreseeable negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 1.2
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Additional Related Work

An alternative approach to achieving improved communication efficiency is through the use of communication compression via (unbiased) quantization [Zhu et al., 2017, Alistarh et al., 2017, Horváth et al., 2019a, Mishchenko et al., 2021b], sketching [Hanzely et al., 2018, Safaryan et al., 2021] or sparsification [Wangni et al., 2017, Wang et al., 2018, Mishchenko et al., 2020]. Modern variants offering with variance reduction for the variance caused by compression Mishchenko et al. [2019], Horváth et al. [2019b], Gorbunov et al. [2020b], Safaryan et al. [2021], Mishchenko et al. [2021b], Wang et al. [2021a], Shulgin and Richtárik [2021], adaptivity [Mishchenko et al., 2021b], bidirectional compression [Horváth et al., 2019a, Tang et al., 2019, Philippenko and Dieuleveut, 2020] or acceleration Li et al. [2020c] enjoy better theoretical rates and practical performance. Variance-reduction for communication compression has been extended to work over arbitrary connected networks [Kovalev et al., 2021b], and to second-order methods [Islamov et al., 2021, Safaryan et al., 2022, Qian et al., 2022, Islamov et al., 2022]. The current state-of-the-art communication complexity in the smooth nonconvex regime is offered by the MARINA [Gorbunov et al., 2021, Szlendak et al., 2022] and DASHA [Tyurin and Richtárik, 2022] methods.

Greedy (biased) compressors, such as Top-K sparsification [Alistarh et al., 2018] or Rank-K approximation [Vogels et al., 2019], require a different approach via an error-feedback/compensation mechanism [Stich et al., 2018, Stich and Karimireddy, 2019, Tang et al., 2019, Gorbunov et al., 2020c]. For a more modern treatment of error feedback offering current state-of-the-art rates, we refer the reader to [Richtárik et al., 2021, Fatkhullin et al., 2021, Richtárik et al., 2022, Qian et al., 2020]. An alternative approach based on the transformation of a biased compressor into a related induced unbiased compressor was proposed in [Horváth and Richtárik, 2021], and a unified treatment of variance reduction and error-feedback was proposed in [Condat et al., 2022].

For a systems-oriented survey, we recommend the reader the work of Xu et al. [2020].

In this work we do not consider the communication compression approach to communication efficiency since this area is much more understood, and many methods already improve on the theoretical communication complexity of vanilla GD and SGD, often by significant data and dimension-dependent margins. Instead, we focus on the practice of delayed parameter synchronization via local training, and contribute to the theoretical foundations of this immensely popular yet poorly understood approach to achieving communication efficiency.

A.1 Summary of complexity results for Algorithms 2 and 5 applied to decentralized optimization

B Analysis of the Proximal-Point Method

In this section we justify the claims we made in Section 3.2 about the Proximal-Point Method (Algorithm 3). In particular, we prove the complexity result (11). The result is not new of course, but we could not find a source for the proof we include here.

B.1 The Proximal-Point Method

We have described the Proximal-Point Method informally in Section 3.2. Here we state it formally as Algorithm 3.

Algorithm 3 Proximal-Point Method

Input: Initial point $(x^0, y^0) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$; Stepsizes $\eta_x, \eta_y > 0$
for $k = 0, 1, \ldots$ **do**
  $x^{k+1} = x^k - \eta_x \left(\partial G(x^{k+1}) + K^\top y^{k+1}\right)$
  $y^{k+1} = y^k - \eta_y \left(\partial F^\star(y^{k+1}) - K x^{k+1}\right)$
end for

Table 3: Complexity results for Algorithms 2 and 5 applied to solving the decentralized optimization problem (5) formulated as the saddle-point problem (25). Our results improve upon those of ProxSkip, both in communication complexity (for Algorithm 5), and the number of local gradient steps (for both our methods, given a proper choice of $\mathcal{M}$).

Alg.	Inner Method $\mathcal{M}$	Deter-ministic #local steps	Centralized case		Decentralized case		
			#comm. rounds	#local steps per round	GA[1]	#comm. rounds	#local steps per round
D-SGD[3]	GD	✓	$\widetilde{\mathcal{O}}(\kappa)$	τ	✗	$\widetilde{\mathcal{O}}(\kappa\chi)$	τ
ProxSkip[4]	GD	✗	$\widetilde{\mathcal{O}}(\sqrt{\kappa})$	$\mathcal{O}(\sqrt{\kappa})$	✗	$\widetilde{\mathcal{O}}(\sqrt{\kappa\chi})$[1]	$\widetilde{\mathcal{O}}(\sqrt{\kappa})$[2]
Alg. 2	GD	✓	$\widetilde{\mathcal{O}}(\sqrt{\kappa})$	$\widetilde{\mathcal{O}}(\sqrt{\kappa})$	✗	$\widetilde{\mathcal{O}}(\sqrt{\kappa\chi}\vee\chi)$[5]	$\widetilde{\mathcal{O}}(\sqrt{\kappa})$
	FGD + GD	✓		$\widetilde{\mathcal{O}}(\sqrt[3]{\kappa})$	✗		$\widetilde{\mathcal{O}}(\sqrt[3]{\kappa})$
	FGD + FSFOM	✓		$\widetilde{\mathcal{O}}(\sqrt[4]{\kappa})$	✗		$\widetilde{\mathcal{O}}(\sqrt[4]{\kappa})$
Alg. 5	GD	✓	$\widetilde{\mathcal{O}}(\sqrt{\kappa})$	$\widetilde{\mathcal{O}}(\sqrt{\kappa})$	✓	$\widetilde{\mathcal{O}}(\sqrt{\kappa\chi})$	$\widetilde{\mathcal{O}}(\sqrt{\kappa})$
	FGD + GD	✓		$\widetilde{\mathcal{O}}(\sqrt[3]{\kappa})$	✓		$\widetilde{\mathcal{O}}(\sqrt[3]{\kappa})$
	FGD + FSFOM	✓		$\widetilde{\mathcal{O}}(\sqrt[4]{\kappa})$	✓		$\widetilde{\mathcal{O}}(\sqrt[4]{\kappa})$

(1) Does not use Accelerated Gossip technique (AG).
(2) Valid only when $\kappa \leq \chi$.
(3) This method was analyzed by Koloskova et al. [2020].
(4) This method was proposed and analyzed by Mishchenko et al. [2022].
(5) For $a, b \in \mathbb{R}$, we denote $a \vee b = \max\{a, b\}$

B.2 Bonding the distance of the primal iterates to the primal solution

In our first lemma, we will provide a bound on $\|x^{k+1} - x^\star\|^2$.

Lemma 2. *Let Assumption 1 hold and choose any $\eta_x, \eta_y > 0$. Then the iterates of the Proximal-Point Method (Algorithm 3) for all $k \geq 0$ satisfy*

$$(1 + 2\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 \leq \frac{1}{\eta_x}\|x^k-x^\star\|^2-\frac{1}{\eta_x}\|x^{k+1}-x^k\|^2-2\langle K^\top y^{k+1}-K^\top y^\star, x^{k+1}-x^\star\rangle.$$

Proof. By writing x^{k+1} as $x^k + (x^{k+1} - x^k)$, and using line 3 of Algorithm 3, which reads $x^{k+1} = x^k - \eta_x\left(\partial G(x^{k+1}) + K^\top y^{k+1}\right)$, we get

$$\begin{aligned}
\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 &= \frac{1}{\eta_x}\|x^k - x^\star\|^2 + \frac{2}{\eta_x}\langle x^{k+1} - x^k, x^k - x^\star\rangle + \frac{1}{\eta_x}\|x^{k+1} - x^k\|^2 \\
&= \frac{1}{\eta_x}\|x^k - x^\star\|^2 + \frac{2}{\eta_x}\langle x^{k+1} - x^k, x^{k+1} - x^\star\rangle - \frac{1}{\eta_x}\|x^{k+1} - x^k\|^2 \\
&= \frac{1}{\eta_x}\|x^k - x^\star\|^2 - 2\langle \partial G(x^{k+1}) + K^\top y^{k+1}, x^{k+1} - x^\star\rangle - \frac{1}{\eta_x}\|x^{k+1} - x^k\|^2.
\end{aligned}$$

We now split the inner product into two parts by applying the optimality condition (see (8))

$$0 \in \partial G(x^\star) + K^\top y^\star,$$

obtaining[12]

$$\begin{aligned}
\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 &= \frac{1}{\eta_x}\|x^k - x^\star\|^2 - 2\langle \partial G(x^{k+1}) - \partial G(x^\star), x^{k+1} - x^\star\rangle - \frac{1}{\eta_x}\|x^{k+1} - x^k\|^2 \\
&\quad -2\langle K^\top y^{k+1} - K^\top y^\star, x^{k+1} - x^\star\rangle.
\end{aligned}$$

[12]Abusing notation, here by $\partial G(x^\star)$ we refer to the subgradient $g_G(x^\star) \in \partial G(x^\star)$ for which $0 = g_G(x^\star) + K^\top y^\star$, i.e., $g_G(x^\star) = -K^\top y^\star$.

Finally, this allows us to replace the first inner product using strong convexity of G as follows:

$$\begin{aligned}\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 &\leq \frac{1}{\eta_x}\|x^k-x^\star\|^2 - 2\mu_x\|x^{k+1}-x^\star\|^2 - \frac{1}{\eta_x}\|x^{k+1}-x^k\|^2 \\ &\quad -2\langle K^\top y^{k+1} - K^\top y^\star, x^{k+1}-x^\star\rangle.\end{aligned}$$

□

B.3 Bonding the distance of the dual iterates to the dual solution

In our second lemma, we will provide a bound on $\|y^{k+1}-y^\star\|^2$. For this, we will rely on an additional assumption (strong convexity of $F^\star$):

Assumption 5. *The function $F^\star:\mathbb{R}^{d_y}\to\mathbb{R}$ is μ_y-strongly convex (but can be non-differentiable), i.e.,*

$$F^\star(y') - F^\star(y'') - \langle g_{F^\star}(y''), y'-y''\rangle \geq \frac{\mu_y}{2}\|y'-y''\|^2, \qquad \forall y', y''\in\mathbb{R}^{d_y}, \tag{30}$$

where $g_{F^\star}(y'')\in\partial F^\star(y'')$ is any subgradient of $F^\star$ at $y''\in\mathbb{R}^{d_y}$.

Lemma 3. *Let Assumptions 3 and 5 hold and choose any $\eta_x,\eta_y>0$. Then the iterates of the Proximal-Point Method (Algorithm 3) for all $k\geq 0$ satisfy*

$$(1+2\mu_y\eta_y)\frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 \leq \frac{1}{\eta_y}\|y^k-y^\star\|^2 - \frac{1}{\eta_y}\|y^{k+1}-y^k\|^2 + 2\langle K^\top y^{k+1}-K^\top y^\star, x^{k+1}-x^\star\rangle.$$

Proof. By writing y^{k+1} as $y^k+(y^{k+1}-y^k)$, and using line 4 of Algorithm 3, which reads $y^{k+1} = y^k - \eta_y\left(\partial F^\star(y^{k+1}) - Kx^{k+1}\right)$, we get

$$\begin{aligned}\frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 &= \frac{1}{\eta_y}\|y^k-y^\star\|^2 + \frac{2}{\eta_y}\langle y^{k+1}-y^k, y^k-y^\star\rangle + \frac{1}{\eta_y}\|y^{k+1}-y^k\|^2 \\ &= \frac{1}{\eta_y}\|y^k-y^\star\|^2 + \frac{2}{\eta_y}\langle y^{k+1}-y^k, y^{k+1}-y^\star\rangle - \frac{1}{\eta_y}\|y^{k+1}-y^k\|^2 \\ &= \frac{1}{\eta_y}\|y^k-y^\star\|^2 - 2\langle \partial F^\star(y^{k+1}) - Kx^{k+1}, y^{k+1}-y^\star\rangle - \frac{1}{\eta_y}\|y^{k+1}-y^k\|^2.\end{aligned}$$

We now split the inner product into two parts by applying the optimality condition (see (8))

$$0\in\partial F^\star(y^\star) - Kx^\star,$$

obtaining[13]

$$\begin{aligned}\frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 &= \frac{1}{\eta_y}\|y^k-y^\star\|^2 - 2\langle \partial F^\star(y^{k+1}) - \partial F^\star(y^\star), y^{k+1}-y^\star\rangle - \frac{1}{\eta_y}\|y^{k+1}-y^k\|^2 \\ &\quad +2\langle Kx^{k+1} - Kx^\star, y^{k+1}-y^\star\rangle.\end{aligned}$$

Finally, this allows us to replace the first inner product using strong convexity of $F^\star$ as follows:

$$\begin{aligned}\frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 &\leq \frac{1}{\eta_y}\|y^k-y^\star\|^2 - 2\mu_y\|y^{k+1}-y^\star\|^2 - \frac{1}{\eta_y}\|y^{k+1}-y^k\|^2 \\ &\quad +2\langle K^\top y^{k+1} - K^\top y^\star, x^{k+1}-x^\star\rangle.\end{aligned}$$

□

B.4 Complexity of the Proximal-Point Method

We now formulate the main result describing the iteration complexity of the Proximal-Point Method. As we shall see, it follows by combining the above two lemmas. Note that the theorem postulates *arbitrarily fast linear convergence*, with the speed controlled via the primal and dual stepsizes η_x and η_y, respectively. The larger the stepsizes, the faster the rate becomes. In the limit, as $\eta_x\to+\infty$ and $\eta_y\to+\infty$, the result obtained predicts convergence in a single iteration.

[13] Abusing notation, here by $\partial F^\star(y^\star)$ we refer to the subgradient $g_{F^\star}(y^\star)\in\partial F^\star(y^\star)$ for which $0 = g_{F^\star}(y^\star) - Kx^\star$, i.e., $g_{F^\star}(y^\star) = Kx^\star$.

Theorem 4. *Let Assumptions 1 and 5 hold, and choose any $\eta_x, \eta_y > 0$. Then the iterates of the Proximal-Point Method (Algorithm 3) for all $k \geq 0$ satisfy*

$$\Delta^{k+1} \leq \frac{\Delta^k}{\min\{1 + 2\mu_x\eta_x, 1 + 2\mu_y\eta_y\}},$$

where the Lyapunov function is defined by

$$\Delta^k \overset{def}{=} \frac{1}{\eta_x}\|x^k - x^\star\|^2 + \frac{1}{\eta_y}\|y^k - y^\star\|^2.$$

This implies the following statement:

$$k \geq \left(1 + \frac{1}{\min\{2\mu_x\eta_x, 2\mu_y\eta_y\}}\right)\log\frac{1}{\varepsilon} \qquad \Rightarrow \qquad \Delta^k \leq \varepsilon\Delta^0.$$

Proof. By adding the inequalities from Lemma 2 and Lemma 3, we obtain

$$\begin{aligned}
(1 + 2\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 \quad &+ \quad (1 + 2\mu_y\eta_y)\frac{1}{\eta_y}\|y^{k+1} - y^\star\|^2 \\
&\leq \quad \frac{1}{\eta_x}\|x^k - x^\star\|^2 - \frac{1}{\eta_x}\|x^{k+1} - x^k\|^2 + \frac{1}{\eta_y}\|y^k - y^\star\|^2 - \frac{1}{\eta_y}\|y^{k+1} - y^k\|^2 \\
&\quad -2\langle K^\top y^{k+1} - K^\top y^\star, x^{k+1} - x^\star\rangle + 2\langle K^\top y^{k+1} - K^\top y^\star, x^{k+1} - x^\star\rangle \\
&\leq \quad \frac{1}{\eta_x}\|x^k - x^\star\|^2 + \frac{1}{\eta_y}\|y^k - y^\star\|^2.
\end{aligned}$$

Thus, if we denote $\Delta^k \overset{\text{def}}{=} \frac{1}{\eta_x}\|x^k - x^\star\|^2 + \frac{1}{\eta_y}\|y^k - y^\star\|^2$ and $m \overset{\text{def}}{=} \min\{1 + 2\mu_x\eta_x, 1 + 2\mu_y\eta_y\}$, the above inequality can be written in the following way:

$$\Delta^{k+1} \leq \frac{1}{m}\Delta^k = \left(1 - \frac{m-1}{m}\right)\Delta^k.$$

Using standard arguments, the above implies that

$$k \geq \frac{m}{m-1}\log\frac{1}{\varepsilon} \quad \Rightarrow \quad \Delta^k \leq \varepsilon\Delta^0.$$

□

C Analysis of the Chambolle-Pock Method

In this section we justify the claims we made in Section 3.3 about the Chambolle-Pock Method (Algorithm 4). In particular, we provide formal statements and proofs of the complexity results (12) and (13) mentioned in Sections C.4 and C.5, respectively.

C.1 The Chambolle-Pock Method

We have described the Chambolle-Pock Method informally in Section 3.2. Here we state it formally as Algorithm 4.

Algorithm 4 Chambolle-Pock Method [Chambolle and Pock, 2011]

1: **Input**: Initial point $(x^0, y^0) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, $\bar{y}^0 = y^0$; Stepsizes $\eta_x, \eta_y > 0$, Extrapolation parameter $\theta \in [0, 1]$
2: **for** $k = 0, 1, \ldots$ **do**
3: $\quad x^{k+1} = x^k - \eta_x \left(\partial G(x^{k+1}) + K^\top \bar{y}^k\right)$
4: $\quad y^{k+1} = y^k - \eta_y \left(\partial F^\star(y^{k+1}) - K x^{k+1}\right)$
5: $\quad \bar{y}^{k+1} = y^{k+1} + \theta \left(y^{k+1} - y^k\right)$
6: **end for**

C.2 Bonding the distance of the primal iterates to the primal solution

In our first lemma, we will provide a bound on $\|x^{k+1} - x^\star\|^2$. The result is identical to Lemma 2 with a single exception: instead of the fresh dual point y^{k+1}, we now have $\bar{y}^k$ on the right-hand side. This lemma applies both to the $\theta = 0$ and $\theta > 0$ cases.

Lemma 4. *Let Assumption 1 hold and choose any $\eta_x > 0$. Then the iterates of the Chambolle-Pock Method (Algorithm 4) for all $k \geq 0$ satisfy*

$$(1 + 2\mu_x\eta_x)\,\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 \leq \frac{1}{\eta_x}\|x^k - x^\star\|^2 - \frac{1}{\eta_x}\|x^{k+1} - x^k\|^2 - 2\langle K^\top \bar{y}^k - K^\top y^\star, x^{k+1} - x^\star\rangle.$$

Proof. Identical to the proof of Lemma 2; one just needs to replace y^{k+1} by $\bar{y}^k$ everywhere. □

C.3 Bonding the distance of the dual iterates to dual solution

We do not need a new bound on $\|y^{k+1} - y^\star\|^2$ for Algorithm 4 since Lemma 3 we proved for the Proximal-Point Method applies here as well.

C.4 Formal statement and proof of (12) (Chambolle-Pock in the $\theta = 0$ case)

We are now ready to state and prove the main complexity result for the Chambolle-Pock Method in the $\theta = 0$ case. This is the formal version of the informal complexity result (12).

Theorem 5 (Complexity of the Chambolle-Pock Method in the $\theta = 0$ case)**.** *Let Assumptions 1, 5 hold. Consider the Chambolle-Pock Method (Algorithm 4) with the extrapolation parameter set as*

$$\theta = 0,$$

and the primal and dual stepsizes set as

$$\eta_x = \frac{\mu_y}{L_{xy}^2}, \qquad \eta_y = \frac{\mu_x}{L_{xy}^2}.$$

Then for the Lyapunov function

$$\Delta^k \overset{def}{=} \frac{1}{\eta_x}\|x^k - x^\star\|^2 + \frac{1}{\eta_y}\|y^k - y^\star\|^2$$

and all $k \geq 0$ we have

$$\Delta^{k+1} \leq \frac{\Delta^k}{\min\{1 + \mu_x\eta_x, 1 + 2\mu_y\eta_y\}} = \frac{\Delta^k}{1 + \frac{\mu_x\mu_y}{L_{xy}^2}}.$$

This implies the following statement:

$$k \geq \left(1 + \frac{1}{\min\{\mu_x\eta_x, 2\mu_y\eta_y\}}\right)\log\frac{1}{\varepsilon} = \left(1 + \frac{L_{xy}^2}{\mu_x\mu_y}\right)\log\frac{1}{\varepsilon} \qquad \Rightarrow \qquad \Delta^k \leq \varepsilon\Delta^0.$$

Proof. By adding the inequalities from Lemma 4 (and noting that $\bar{y}^k = y^k$ in the $\theta = 0$ case) and Lemma 3, we obtain

$$\begin{aligned}
(1+2\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 \quad &+ \quad (1+2\mu_y\eta_y)\frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 \\
&\leq \quad \frac{1}{\eta_x}\|x^k-x^\star\|^2 - \frac{1}{\eta_x}\|x^{k+1}-x^k\|^2 + \frac{1}{\eta_y}\|y^k-y^\star\|^2 - \frac{1}{\eta_y}\|y^{k+1}-y^k\|^2 \\
&- \quad 2\langle K^\top y^k - K^\top y^\star, x^{k+1}-x^\star\rangle + 2\langle K^\top y^{k+1} - K^\top y^\star, x^{k+1}-x^\star\rangle \qquad (31) \\
&\leq \quad \frac{1}{\eta_x}\|x^k-x^\star\|^2 + \frac{1}{\eta_y}\|y^k-y^\star\|^2 \\
&- \quad \frac{1}{\eta_y}\|y^{k+1}-y^k\|^2 + 2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1}-x^\star\rangle, \qquad (32)
\end{aligned}$$

where in the last step we have used the bound $-\frac{1}{\eta_x}\|x^{k+1}-x^k\|^2 \leq 0$.

Using the Cauchy-Schwarz inequality, the definition of L_{xy} as the norm of K (see (2)), and applying Young's inequality, we can bound the inner product by

$$\begin{aligned}
2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1}-x^\star\rangle \quad &\leq \quad 2\|K^\top y^{k+1} - K^\top y^k\|\|x^{k+1}-x^\star\| \\
&\overset{(2)}{\leq} \quad 2L_{xy}\|y^{k+1}-y^k\|\|x^{k+1}-x^\star\| \\
&\leq \quad L_{xy}\left(C\|y^{k+1}-y^k\|^2 + \frac{1}{C}\|x^{k+1}-x^\star\|^2\right),
\end{aligned}$$

for any $C > 0$. Plugging this into (32), and rearranging the inequality so that all terms involving $\|x^{k+1}-x^\star\|^2$ appear on the left-hand side, we get

$$\begin{aligned}
\left(1+2\mu_x\eta_x - \frac{L_{xy}\eta_x}{C}\right)\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 \quad &+ \quad (1+2\mu_y\eta_y)\frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 \\
&\leq \quad \frac{1}{\eta_x}\|x^k-x^\star\|^2 + \frac{1}{\eta_y}\|y^k-y^\star\|^2 \\
&\quad - (1 - CL_{xy}\eta_y)\frac{1}{\eta_y}\|y^{k+1}-y^k\|^2.
\end{aligned}$$

Taking $C = \frac{L_{xy}}{\mu_x}$ and $\eta_y = \frac{\mu_x}{L_{xy}^2}$, we obtain the simplified bound

$$\begin{aligned}
(1+\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 \quad &+ \quad (1+2\mu_y\eta_y)\frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 \\
&\leq \quad \frac{1}{\eta_x}\|x^k-x^\star\|^2 + \frac{1}{\eta_y}\|y^k-y^\star\|^2 - \left(1 - \frac{L_{xy}^2\eta_y}{\mu_x}\right)\frac{1}{\eta_y}\|y^{k+1}-y^k\|^2 \\
&\leq \quad \frac{1}{\eta_x}\|x^k-x^\star\|^2 + \frac{1}{\eta_y}\|y^k-y^\star\|^2. \qquad (33)
\end{aligned}$$

If we let

$$m \overset{\text{def}}{=} \min\{1+\mu_x\eta_x, 1+2\mu_y\eta_y\} = \min\left\{1 + \frac{\mu_x\mu_y}{L_{xy}^2}, 1 + 2\frac{\mu_x\mu_y}{L_{xy}^2}\right\} = 1 + \frac{\mu_x\mu_y}{L_{xy}^2},$$

then inequality (33) implies

$$\Delta^{k+1} \leq \frac{1}{m}\Delta^k = \left(1 - \frac{m-1}{m}\right)\Delta^k \qquad \forall k \geq 0.$$

Using standard arguments, the above implies that

$$k \geq \frac{m}{m-1}\log\frac{1}{\varepsilon} = \left(1+\frac{1}{m-1}\right)\log\frac{1}{\varepsilon} = \left(1+\frac{L_{xy}^2}{\mu_x\mu_y}\right)\log\frac{1}{\varepsilon} \qquad \Rightarrow \qquad \Delta^k \leq \varepsilon\Delta^0.$$

□

C.5 Formal statement and proof of (13) (Chambolle-Pock in the $\theta > 0$ case)

We now study the case when the extrapolation parameter θ is set to a positive value, and show that this helps to get better rates.

Lemma 5. *Under Assumption 1, the iterates of the Chambolle-Pock Method (Algorithm 4) with $\theta > 0$ for any $C > 0$ and all $k \geq 1$ satisfy*

$$\begin{aligned} 2\langle K^\top y^{k+1} - K^\top \bar{y}^k, x^{k+1} - x^\star\rangle \quad \leq \quad & 2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1} - x^\star\rangle \\ & -2\theta\langle K^\top y^k - K^\top y^{k-1}, x^k - x^\star\rangle \\ & +\theta L_{xy} C\|x^{k+1} - x^k\|^2 + \frac{\theta L_{xy}}{C}\|y^k - y^{k-1}\|^2. \end{aligned} \tag{34}$$

Proof. Using line 5 from Algorithm 4, which reads $\bar{y}^k = y^k + \theta\left(y^k - y^{k-1}\right)$, we obtain

$$\begin{aligned} 2\langle K^\top y^{k+1} - K^\top \bar{y}^k, x^{k+1} - x^\star\rangle \quad = \quad & 2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1} - x^\star\rangle \\ & -2\theta\langle K^\top y^k - K^\top y^{k-1}, x^{k+1} - x^\star\rangle \\ = \quad & 2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1} - x^\star\rangle \\ & -2\theta\langle K^\top y^k - K^\top y^{k-1}, x^k - x^\star\rangle \\ & -2\theta\langle K^\top y^k - K^\top y^{k-1}, x^{k+1} - x^k\rangle. \end{aligned} \tag{35}$$

Using the Cauchy-Schwarz inequality, the definition of L_{xy} as the norm of K (see (2)), and applying Young's inequality, we can bound the last inner product by

$$\begin{aligned} -2\theta\langle K^\top y^k - K^\top y^{k-1}, x^{k+1} - x^k\rangle \quad \leq \quad & 2\theta\|K^\top y^k - K^\top y^{k-1}\|\|x^{k+1} - x^k\| \\ \leq \quad & 2\theta L_{xy}\|y^k - y^{k-1}\|\|x^{k+1} - x^k\| \\ \leq \quad & \theta L_{xy}\left(C\|y^k - y^{k-1}\|^2 + \frac{1}{C}\|x^{k+1} - x^k\|^2\right). \end{aligned}$$

It only remains to plug this inequality into (35).

□

We are now ready to state and prove the general theorem.

Theorem 6 (Complexity of the Chambolle-Pock Method in the $\theta > 0$ case)**.** *Let Assumptions 1, 5 hold. Consider the Chambolle-Pock Method (Algorithm 4) with the extrapolation parameter set as*

$$\theta = \max\left\{\frac{1}{1+2\mu_x\eta_x}, \frac{1}{1+2\mu_y\eta_y}\right\}, \tag{36}$$

and primal and dual stepsizes set as

$$\eta_x = \frac{1}{L_{xy}}\sqrt{\frac{\mu_y}{\mu_x}}, \qquad \eta_y = \frac{1}{L_{xy}}\sqrt{\frac{\mu_x}{\mu_y}}. \tag{37}$$

Then for the Lyapunov function defined for $k \geq 0$ via

$$\begin{aligned} \Delta^{k+1} \quad \stackrel{def}{=} \quad & (1+2\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 + (1+2\mu_y\eta_y)\frac{1}{\eta_y}\|y^{k+1} - y^\star\|^2 \\ & +\frac{1}{\eta_y}\|y^{k+1} - y^k\|^2 - 2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1} - x^\star\rangle \end{aligned} \tag{38}$$

and for $k = 0$ via

$$\Delta^0 \stackrel{def}{=} (1 + 2\mu_x\eta_x)\frac{1}{\eta_x}\|x^0 - x^\star\|^2 + (1 + 2\mu_y\eta_y)\frac{1}{\eta_y}\|y^0 - y^\star\|^2, \tag{39}$$

we have

$$0 \le \Delta^k \le \theta^k \Delta^0 \qquad \forall k \ge 0. \tag{40}$$

Proof. We will proceed in several steps.

Showing that $\Delta^k \ge 0$ for all $k \ge 0$. First, let us show that $\Delta^k \ge 0$ for every k. This is clear for $k = 0$ from (39). Let us show that $\Delta^{k+1} \ge 0$ for $k \ge 0$. Using Young's inequality and (2), we obtain the inequality

$$\begin{aligned}
\Delta^{k+1} &\stackrel{(38)}{=} (1 + 2\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 + (1 + 2\mu_y\eta_y)\frac{1}{\eta_y}\|y^{k+1} - y^\star\|^2 \\
&\quad + \frac{1}{\eta_y}\|y^{k+1} - y^k\|^2 - 2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1} - x^\star\rangle \\
&\ge (1 + 2\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 + (1 + 2\mu_y\eta_y)\frac{1}{\eta_y}\|y^{k+1} - y^\star\|^2 \\
&\quad + \frac{1}{\eta_y}\|y^{k+1} - y^k\|^2 - L_{xy}B\|y^{k+1} - y^k\|^2 - \frac{L_{xy}}{B}\|x^{k+1} - x^\star\|^2 \\
&= \left(1 + 2\mu_x\eta_x - \frac{L_{xy}\eta_x}{B}\right)\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 + (1 + 2\mu_y\eta_y)\frac{1}{\eta_y}\|y^{k+1} - y^\star\|^2 \\
&\quad + (1 - L_{xy}B\eta_y)\frac{1}{\eta_y}\|y^{k+1} - y^k\|^2,
\end{aligned} \tag{41}$$

which holds for all $B > 0$. Selecting $B = \frac{1}{L_{xy}\eta_y}$, the blue term is zeroed out, and using the primal and dual stepsizes (37), we get

$$\begin{aligned}
\Delta^{k+1} &\stackrel{(41)}{\ge} \left(1 + 2\mu_x\eta_x - L_{xy}^2\eta_x\eta_y\right)\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 + (1 + 2\mu_y\eta_y)\frac{1}{\eta_y}\|y^{k+1} - y^\star\|^2 \\
&\stackrel{(37)}{=} \frac{2\sqrt{\mu_x\mu_y}}{L_{xy}}\|x^{k+1} - x^\star\|^2 + (1 + 2\mu_y\eta_y)\frac{1}{\eta_y}\|y^{k+1} - y^\star\|^2 \\
&\ge 0.
\end{aligned}$$

Establishing technical bounds. Denote

$$\mathcal{W}^{k+1} \stackrel{\text{def}}{=} (1 + 2\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 + (1 + 2\mu_y\eta_y)\frac{1}{\eta_y}\|y^{k+1} - y^\star\|^2. \tag{42}$$

Combining Lemma 4, which provides a bound on $\|x^{k+1} - x^\star\|^2$, and Lemma 3, which provides a bound on $\|y^{k+1} - y^\star\|^2$, we obtain

$$\begin{aligned}
\mathcal{W}^{k+1} &\le \frac{1}{\eta_x}\|x^k - x^\star\|^2 - \frac{1}{\eta_x}\|x^{k+1} - x^k\|^2 - 2\langle K^\top \bar{y}^k - K^\top y^\star, x^{k+1} - x^\star\rangle \\
&\quad + \frac{1}{\eta_y}\|y^k - y^\star\|^2 - \frac{1}{\eta_y}\|y^{k+1} - y^k\|^2 + 2\langle K^\top y^{k+1} - K^\top y^\star, x^{k+1} - x^\star\rangle \\
&\le \frac{1}{\eta_x}\|x^k - x^\star\|^2 - \frac{1}{\eta_x}\|x^{k+1} - x^k\|^2 + \frac{1}{\eta_y}\|y^k - y^\star\|^2 - \frac{1}{\eta_y}\|y^{k+1} - y^k\|^2 \\
&\quad + 2\langle K^\top y^{k+1} - K^\top \bar{y}^k, x^{k+1} - x^\star\rangle.
\end{aligned} \tag{43}$$

This is the same inequality as (31) with the exception that y^k was replaced by $\bar{y}^k$. Using Lemma 5 to bound the inner product in (43), we get

$$\begin{aligned}
\mathcal{W}^{k+1} &\overset{(43)+(34)}{\leq} \frac{1}{\eta_x}\|x^k - x^\star\|^2 - \frac{1}{\eta_x}\|x^{k+1} - x^k\|^2 + \frac{1}{\eta_y}\|y^k - y^\star\|^2 - \frac{1}{\eta_y}\|y^{k+1} - y^k\|^2 \\
&\quad +2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1} - x^\star\rangle - 2\theta\langle K^\top y^k - K^\top y^{k-1}, x^k - x^\star\rangle \\
&\quad +\theta L_{xy} C\|x^{k+1} - x^k\|^2 + \frac{\theta L_{xy}}{C}\|y^k - y^{k-1}\|^2 \\
&\leq \frac{1}{\eta_x}\|x^k - x^\star\|^2 - \left(\frac{1}{\eta_x} - \theta L_{xy} C\right)\|x^{k+1} - x^k\|^2 + \frac{1}{\eta_y}\|y^k - y^\star\|^2 - \frac{1}{\eta_y}\|y^{k+1} - y^k\|^2 \\
&\quad +2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1} - x^\star\rangle - 2\theta\langle K^\top y^k - K^\top y^{k-1}, x^k - x^\star\rangle \\
&\quad +\frac{\theta L_{xy}}{C}\|y^k - y^{k-1}\|^2 \\
&= \frac{1}{\eta_x}\|x^k - x^\star\|^2 + \frac{1}{\eta_y}\|y^k - y^\star\|^2 - \frac{1}{\eta_y}\|y^{k+1} - y^k\|^2 \\
&\quad +2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1} - x^\star\rangle - 2\theta\langle K^\top y^k - K^\top y^{k-1}, x^k - x^\star\rangle \\
&\quad +\theta^2 L_{xy}^2 \eta_x \eta_y \frac{1}{\eta_y}\|y^k - y^{k-1}\|^2, \qquad (44)
\end{aligned}$$

where in the last step we have made the choice $C \overset{\text{def}}{=} (\theta \eta_x L_{xy})^{-1}$.

Showing that $\Delta^{k+1} \leq \theta\Delta^k$ for $k \geq 1$. By combining (38) and (42), and applying inequality (44), for $k \geq 1$ we get

$$\begin{aligned}
\Delta^{k+1} &\overset{(38)+(42)}{=} \mathcal{W}^{k+1} + \frac{1}{\eta_y}\|y^{k+1} - y^k\|^2 - 2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1} - x^\star\rangle \\
&\overset{(44)}{\leq} \frac{1}{\eta_x}\|x^k - x^\star\|^2 + \frac{1}{\eta_y}\|y^k - y^\star\|^2 - \frac{1}{\eta_y}\|y^{k+1} - y^k\|^2 \\
&\quad +2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1} - x^\star\rangle - 2\theta\langle K^\top y^k - K^\top y^{k-1}, x^k - x^\star\rangle \\
&\quad +\theta^2 L_{xy}^2 \eta_x \eta_y \frac{1}{\eta_y}\|y^k - y^{k-1}\|^2 \\
&\quad +\frac{1}{\eta_y}\|y^{k+1} - y^k\|^2 - 2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1} - x^\star\rangle \\
&\overset{(A)}{\leq} \frac{1}{\eta_x}\|x^k - x^\star\|^2 + \frac{1}{\eta_y}\|y^k - y^\star\|^2 - 2\theta\langle K^\top y^k - K^\top y^{k-1}, x^k - x^\star\rangle + \theta^2 L_{xy}^2 \eta_x \eta_y \frac{1}{\eta_y}\|y^k - y^{k-1}\|^2 \\
&\overset{(B)}{\leq} \frac{1}{\eta_x}\|x^k - x^\star\|^2 + \frac{1}{\eta_y}\|y^k - y^\star\|^2 - 2\theta\langle K^\top y^k - K^\top y^{k-1}, x^k - x^\star\rangle + \theta^2 \frac{1}{\eta_y}\|y^k - y^{k-1}\|^2 \\
&\overset{(C)}{\leq} \theta\left((1 + 2\mu_x\eta_x)\frac{1}{\eta_x}\|x^k - x^\star\|^2 + (1 + 2\mu_y\eta_y)\frac{1}{\eta_y}\|y^k - y^\star\|^2\right) \\
&\quad +\theta\left(\frac{1}{\eta_y}\|y^k - y^{k-1}\|^2 - 2\langle K^\top y^k - K^\top y^{k-1}, x^k - x^\star\rangle\right) \\
&\overset{(38)}{=} \theta\Delta^k, \qquad (45)
\end{aligned}$$

where in step (A) we annihilated the red and blue terms as their sum is zero, in step (B) we used the fact that $L_{xy}^2 \eta_x \eta_y \leq 1$, which follows from the stepsize choice, and in (C) we used the inequalities $1 \leq \theta(1 + 2\mu_x\eta_x)$, $1 \leq \theta(1 + 2\mu_y\eta_y)$, and $\theta \leq 1$, which follow from (36).

Showing that $\Delta^{k+1} \leq \theta\Delta^k$ for $k = 0$. We start with inequality (43) for $k = 0$:

$$\begin{aligned}
\mathcal{W}^1 &\overset{(43)}{\leq} \frac{1}{\eta_x}\|x^0 - x^\star\|^2 - \frac{1}{\eta_x}\|x^1 - x^0\|^2 + \frac{1}{\eta_y}\|y^0 - y^\star\|^2 - \frac{1}{\eta_y}\|y^1 - y^0\|^2 \\
&\quad +2\langle K^\top y^1 - K^\top y^0, x^1 - x^\star\rangle. \qquad (46)
\end{aligned}$$

According to (38) and (36), we have

$$
\begin{aligned}
\Delta^1 &\overset{(38)+(42)}{=} \mathcal{W}^1 + \frac{1}{\eta_y}\|y^1 - y^0\|^2 - 2\langle K^\top y^1 - K^\top y^0, x^1 - x^\star\rangle \\
&\overset{(46)}{\leq} \frac{1}{\eta_x}\|x^0 - x^\star\|^2 - \frac{1}{\eta_x}\|x^1 - x^0\|^2 + \frac{1}{\eta_y}\|y^0 - y^\star\|^2 \\
&\leq \frac{1}{\eta_x}\|x^0 - x^\star\|^2 + \frac{1}{\eta_y}\|y^0 - y^\star\|^2 \\
&\leq \theta\left((1 + 2\mu_x\eta_x)\frac{1}{\eta_x}\|x^0 - x^\star\|^2 + (1 + 2\mu_y\eta_y)\frac{1}{\eta_y}\|y^0 - y^\star\|^2\right) \\
&= \theta\Delta^0
\end{aligned}
$$

where in the last inequality we used the inequalities $1 \leq \theta(1 + 2\mu_x\eta_x)$ and $1 \leq \theta(1 + 2\mu_y\eta_y)$, which follow from (36).

□

The informal result (13) mentioned in Section C.5 is a simple corollary of the above theorem. Indeed, using the definition of θ, we can obtain:

$$
k \geq \frac{1}{1-\theta}\log\frac{1}{\varepsilon} \qquad \Rightarrow \qquad \Delta^k \leq \varepsilon\Delta^0.
$$

It remains to remark that

$$
\begin{aligned}
\frac{1}{1-\theta}\log\frac{1}{\varepsilon} &= \mathcal{O}\left(\left(1 + \max\left\{\frac{1}{2\mu_x\eta_x}, \frac{1}{2\mu_y\eta_y}\right\}\right)\log\frac{1}{\varepsilon}\right) \\
&= \mathcal{O}\left(\left(1 + \frac{L_{xy}}{\sqrt{\mu_x\mu_y}}\right)\log\frac{1}{\varepsilon}\right),
\end{aligned}
$$

which is the result from (13).

D Analysis of the Accelerated Primal-Dual Algorithm (APDA; Algorithm 1)

In this section we perform convergence analysis for our new method APDA (Algorithm 1). We start by establishing three lemmas, followed by the proof of the main theorem.

D.1 Three lemmas

The first result is a variant of Lemma 4, offering a bound on the distance of the primal iterates from the primal optimal solution. Compared to Lemma 4, this result is strengthened by the additional assumption of L_x-smoothness of G (Assumption 2).

Lemma 6. *Let Assumptions 1 and 2 hold and choose any $\eta_x, \eta_y > 0$. Then the iterates of APDA (Algorithm 1) for all $k \geq 0$ satisfy*

$$(1+\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 \leq \frac{1}{\eta_x}\|x^k-x^\star\|^2 - \frac{1}{\eta_x}\|x^{k+1}-x^k\|^2 - 2\langle K^\top\bar{y}^k - K^\top y^\star, x^{k+1}-x^\star\rangle - \frac{1}{L_x}\|\nabla G(x^{k+1})-\nabla G(x^\star)\|^2.$$

Proof. By writing x^{k+1} as $x^k + (x^{k+1}-x^k)$, and using line 3 of Algorithm 1, which reads $x^{k+1} = x^k - \eta_x\left(\nabla G(x^{k+1}) + K^\top\bar{y}^k\right)$, we get

$$\begin{aligned}\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 &= \frac{1}{\eta_x}\|x^k-x^\star\|^2 + \frac{2}{\eta_x}\langle x^{k+1}-x^k, x^{k+1}-x^\star\rangle - \frac{1}{\eta_x}\|x^{k+1}-x^k\|^2 \\ &= \frac{1}{\eta_x}\|x^k-x^\star\|^2 - 2\langle \nabla G(x^{k+1}) + K^\top\bar{y}^k, x^{k+1}-x^\star\rangle - \frac{1}{\eta_x}\|x^{k+1}-x^k\|^2.\end{aligned}$$

We now split the inner product into two parts by applying the optimality condition (see (8))

$$\nabla G(x^\star) + K^\top y^\star = 0,$$

obtaining

$$\begin{aligned}\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 &= \frac{1}{\eta_x}\|x^k-x^\star\|^2 - 2\langle \nabla G(x^{k+1}) - \nabla G(x^\star), x^{k+1}-x^\star\rangle - \frac{1}{\eta_x}\|x^{k+1}-x^k\|^2 \\ &\quad -2\langle K^\top\bar{y}^k - K^\top y^\star, x^{k+1}-x^\star\rangle. \end{aligned} \tag{47}$$

Since G is μ_x-strongly-convex and L_x-smooth, we can lower-bound [Nesterov, 2004] the inner product appearing in (47) as follows:

$$2\langle \nabla G(x^{k+1}) - \nabla G(x^\star), x^{k+1}-x^\star\rangle \geq \mu_x\|x^{k+1}-x^\star\|^2 + \frac{1}{L_x}\|\nabla G(x^{k+1}) - \nabla G(x^\star)\|^2. \tag{48}$$

Finally, by plugging (48) into (47), we get

$$\begin{aligned}\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 &\leq \frac{1}{\eta_x}\|x^k-x^\star\|^2 - \mu_x\|x^{k+1}-x^\star\|^2 - \frac{1}{\eta_x}\|x^{k+1}-x^k\|^2 \\ &\quad -2\langle K^\top\bar{y}^k - K^\top y^\star, x^{k+1}-x^\star\rangle - \frac{1}{L_x}\|\nabla G(x^{k+1}) - \nabla G(x^\star)\|^2.\end{aligned}$$

□

The second result is a technical lemma borrowed from [Kovalev et al., 2021a] for lower-bounding the term $\|K^\top y - K^\top y^\star\|$.

Lemma 7 (See Kovalev et al. [2021a])**.** *Let Assumption 4 hold and let $(x^\star, y^\star)$ be a solution of* (7)*. Then*

$$\|K^\top y - K^\top y^\star\| \geq \mu_{xy}\|y - y^\star\|, \qquad \forall y \in \mathbb{R}^{d_y}.$$

The third result offers a bound on the distance between the dual iterates and the dual optimal solution. When analyzing the Proximal Point method and the Chambolle-Pock method, in this step we relied on Lemma 3, in the proof of which we required $F^\star$ to be μ_x-strongly convex. However, for APDA we explicitly wish to avoid using this assumption. As we shall see, in order to obtain linear convergence, it is enough will invoke Assumption 4. The next lemma is an analogue of Lemma 3 without the need to assume μ_x-strong-convexity of $F^\star$.

Lemma 8. *Let Assumptions 2, 3, and 4 be satisfied, and choose*

$$\beta_y \leq \frac{1}{2L_{xy}^2\eta_y}. \tag{49}$$

Then the iterates of APDA (Algorithm 1) for all $k \geq 0$ satisfy

$$\frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 \leq (1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^k-y^\star\|^2 - \frac{1}{2\eta_y}\|y^{k+1}-y^k\|^2 + 2\langle K^\top y^{k+1}-K^\top y^\star, x^{k+1}-x^\star\rangle + \beta_y\|\nabla G(x^{k+1})-\nabla G(x^\star)\|^2.$$

Proof. By writing y^{k+1} as $y^k + (y^{k+1}-y^k)$, and using line 4 from Algorithm 1, which reads

$$y^{k+1} = y^k - \eta_y\left(\partial F^\star(y^{k+1}) - Kx^{k+1}\right) - \eta_y\beta_y K\left(K^\top y^k + \nabla G(x^{k+1})\right),$$

we get

$$\begin{aligned}\frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 &= \frac{1}{\eta_y}\|y^k-y^\star\|^2 + \frac{2}{\eta_y}\langle y^{k+1}-y^k, y^{k+1}-y^\star\rangle - \frac{1}{\eta_y}\|y^{k+1}-y^k\|^2 \\ &= \underbrace{\frac{1}{\eta_y}\|y^k-y^\star\|^2 - 2\langle\partial F^\star(y^{k+1})-Kx^{k+1}, y^{k+1}-y^\star\rangle - \frac{1}{\eta_y}\|y^{k+1}-y^k\|^2}_{A^k} \\ &\quad \underbrace{-2\beta_y\langle K^\top y^k + \nabla G(x^{k+1}), K^\top y^{k+1} - K^\top y^\star\rangle}_{B^k}.\end{aligned} \tag{50}$$

We split the inner product appearing in A^k into two parts by applying the optimality condition (see (8))

$$0 \in \partial F^\star(y^\star) - Kx^\star,$$

obtaining

$$\begin{aligned}A^k &= \frac{1}{\eta_y}\|y^k-y^\star\|^2 - 2\langle\partial F^\star(y^{k+1}) - \partial F^\star(y^\star), y^{k+1}-y^\star\rangle - \frac{1}{\eta_y}\|y^{k+1}-y^k\|^2 \\ &\quad +2\langle Kx^{k+1} - Kx^\star, y^{k+1}-y^\star\rangle.\end{aligned} \tag{51}$$

Applying the parallelogram identity[14] to B^k, and using the optimality condition (see (8)),

$$\nabla G(x^\star) + K^\top y^\star = 0, \tag{52}$$

we can write

$$\begin{aligned}B^k &= \beta_y\|\nabla G(x^{k+1}) + K^\top y^\star\|^2 - \beta_y\|K^\top y^k - K^\top y^\star\|^2 \\ &\quad +\beta_y\|K^\top y^k - K^\top y^{k+1}\|^2 - \beta_y\|\nabla G(x^{k+1}) + K^\top y^{k+1}\|^2 \\ &\overset{(52)}{=} \beta_y\|\nabla G(x^{k+1}) - \nabla G(x^\star)\|^2 - \beta_y\|K^\top y^k - K^\top y^\star\|^2 \\ &\quad +\beta_y\|K^\top y^k - K^\top y^{k+1}\|^2 - \beta_y\|\nabla G(x^{k+1}) + K^\top y^{k+1}\|^2.\end{aligned} \tag{53}$$

[14] Here we refer to the identity $2\langle a+b, c-d\rangle = -\|b+d\|^2 + \|a-d\|^2 - \|a-c\|^2 + \|b+c\|^2$ which holds for all $a, b, c, d \in \mathbb{R}^{d_x}$.

Plugging (51) and (53) back into (50), and using convexity of $F^\star$, we get

$$\begin{aligned}
\frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 \quad &\overset{(50)}{=} \quad A^k + B^k \\
&\overset{(51)+(53)}{=} \quad \frac{1}{\eta_y}\|y^k-y^\star\|^2 \underbrace{-2\langle \partial F^\star(y^{k+1}) - \partial F^\star(y^\star), y^{k+1}-y^\star\rangle}_{\leq 0} - \frac{1}{\eta_y}\|y^{k+1}-y^k\|^2 \\
&\qquad +2\langle Kx^{k+1}-Kx^\star, y^{k+1}-y^\star\rangle \\
&\qquad +\beta_y\|\nabla G(x^{k+1}) - \nabla G(x^\star)\|^2 - \beta_y\|K^\top y^k - K^\top y^\star\|^2 \\
&\qquad +\beta_y\|K^\top y^k - K^\top y^{k+1}\|^2 \underbrace{-\beta_y\|\nabla G(x^{k+1}) + K^\top y^{k+1}\|^2}_{\leq 0} \\
&\leq \quad \frac{1}{\eta_y}\|y^k-y^\star\|^2 - \frac{1}{\eta_y}\|y^{k+1}-y^k\|^2 \\
&\qquad +2\langle Kx^{k+1}-Kx^\star, y^{k+1}-y^\star\rangle \\
&\qquad +\beta_y\|\nabla G(x^{k+1}) - \nabla G(x^\star)\|^2 - \beta_y\|K^\top y^k - K^\top y^\star\|^2 \\
&\qquad +\beta_y\|K^\top y^k - K^\top y^{k+1}\|^2. \qquad (54)
\end{aligned}$$

It now only remains to plug the bounds $\|K^\top y^k - K^\top y^\star\|^2 \geq \mu_{xy}^2\|y^k - y^\star\|^2$ (see Lemma 7) and $\|K^\top y^k - K^\top y^{k+1}\|^2 \leq L_{xy}^2\|y^k - y^{k+1}\|^2$ into (54), and apply the restriction (49) on β_y.

□

D.2 Main result

We are now ready to state the formal version of Theorem 1.

Theorem 7 (Covergence of APDA; formal)**.** *Let Assumptions 1, 2, 3, and 4 hold and choose the various parameters of the method as follows:*

$$\beta_y = \min\left\{\frac{1}{L_x}, \frac{1}{2L_{xy}^2\eta_y}\right\}, \qquad (55)$$

$$\eta_x = \frac{1}{2\sqrt{L_x\mu_x}}\frac{\mu_{xy}}{L_{xy}}, \qquad (56)$$

$$\eta_y = \frac{\sqrt{L_x\mu_x}}{L_{xy}\mu_{xy}}, \qquad (57)$$

$$\theta = \max\left\{\frac{1}{1+\mu_x\eta_x}, 1-\mu_{xy}^2\beta_y\eta_y\right\}. \qquad (58)$$

Then for the Lyapunov function defined for $k \geq 0$ via

$$\begin{aligned}
\Xi^{k+1} \quad &\overset{def}{=} \quad (1+\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 + \frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 \\
&\qquad + \frac{1}{2\eta_y}\|y^{k+1}-y^k\|^2 - 2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1}-x^\star\rangle \qquad (59)
\end{aligned}$$

and for $k = 0$ via

$$\Xi^0 \overset{def}{=} (1+\mu_x\eta_x)\frac{1}{\eta_x}\|x^0-x^\star\|^2 + \frac{1}{\eta_y}\|y^0-y^\star\|^2,$$

we have

$$0 \leq \mu_x\|x^{k+1}-x^\star\|^2 + \frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 \leq \Xi^k \leq \theta^k\Delta^0 \qquad \forall k \geq 0. \qquad (60)$$

Proof. We will proceed in several steps.

Showing that $\Xi^k \geq 0$ for all $k \geq 0$. First, we will show that $\Xi^k \geq 0$ for every k. This is clear for $k = 0$. Let us show that $\Xi^{k+1} \geq 0$ for every $k \geq 0$. Using Young's inequality with any parameter $B > 0$, and the definition of L_{xy} from (2), we obtain

$$\begin{aligned}
\Xi^{k+1} &\overset{(59)}{=} (1+\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 + \frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 + \frac{1}{2\eta_y}\|y^{k+1}-y^k\|^2 \\
&\quad -2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1}-x^\star\rangle \\
&\geq (1+\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 + \frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 + \frac{1}{2\eta_y}\|y^{k+1}-y^k\|^2 \\
&\quad -L_{xy}B\|y^{k+1}-y^k\|^2 - \frac{L_{xy}}{B}\|x^{k+1}-x^\star\|^2 \\
&= (1+\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 + \frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 + \left(\frac{1}{2} - L_{xy}B\eta_y\right)\frac{1}{\eta_y}\|y^{k+1}-y^k\|^2 \\
&\quad -\frac{L_{xy}}{B}\|x^{k+1}-x^\star\|^2. \qquad (61)
\end{aligned}$$

Using (61) with $B = \frac{1}{2L_{xy}\eta_y}$, and noticing that $2L_{xy}^2\eta_x\eta_y = 1$ (this follows from (56) and (57)), we get

$$\begin{aligned}
\Xi^{k+1} &\geq (1+\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 + \frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 - \frac{2L_{xy}^2\eta_x\eta_y}{\eta_x}\|x^{k+1}-x^\star\|^2 \\
&= \mu_x\|x^{k+1}-x^\star\|^2 + \frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 \\
&\geq 0.
\end{aligned}$$

Establishing technical bounds. Denote

$$\mathcal{W}^{k+1} \overset{\text{def}}{=} (1+\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 + \frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2. \qquad (62)$$

By adding the inequalities from Lemmas 6 and 8, we obtain

$$\begin{aligned}
\mathcal{W}^{k+1} &\leq \frac{1}{\eta_x}\|x^k-x^\star\|^2 - \frac{1}{\eta_x}\|x^{k+1}-x^k\|^2 - \frac{1}{L_x}\|\nabla G(x^{k+1}) - \nabla G(x^\star)\|^2 \\
&\quad +(1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^k-y^\star\|^2 - \frac{1}{2\eta_y}\|y^{k+1}-y^k\|^2 + \beta_y\|\nabla G(x^{k+1}) - \nabla G(x^\star)\|^2 \\
&\quad +2\langle K^\top y^{k+1} - K^\top y^\star, x^{k+1}-x^\star\rangle - 2\langle K^\top \bar{y}^k - K^\top y^\star, x^{k+1}-x^\star\rangle \\
&\leq \frac{1}{\eta_x}\|x^k-x^\star\|^2 - \frac{1}{\eta_x}\|x^{k+1}-x^k\|^2 + (1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^k-y^\star\|^2 - \frac{1}{2\eta_y}\|y^{k+1}-y^k\|^2 \\
&\quad +2\langle K^\top y^{k+1} - K^\top \bar{y}^k, x^{k+1}-x^\star\rangle, \qquad (63)
\end{aligned}$$

where in the last step we used the bound

$$\beta_y\|\nabla G(x^{k+1}) - \nabla G(x^\star)\|^2 - \frac{1}{L_x}\|\nabla G(x^{k+1}) - \nabla G(x^\star)\|^2 \leq 0,$$

which follows from the restriction $\beta_y \leq \frac{1}{L_x}$; see (55). Using line 5 from Algorithm 1, which says

$$\bar{y}^{k+1} = y^{k+1} + \theta(y^{k+1} - y^k), \qquad (64)$$

applying Cauchy-Schwarz inequality, and subsequently using Young's inequality with constant $C > 0$, the inner product from (63) can for $k \geq 1$ be further bounded as follows

$$
\begin{aligned}
2\langle K^\top y^{k+1} - K^\top \bar{y}^k, x^{k+1} - x^\star\rangle &\overset{(64)}{=} 2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1} - x^\star\rangle \\
&\quad -2\theta\langle K^\top y^k - K^\top y^{k-1}, x^k - x^\star\rangle \\
&\quad +2\theta\langle K^\top y^k - K^\top y^{k-1}, x^{k+1} - x^k\rangle \\
&\leq 2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1} - x^\star\rangle \\
&\quad -2\theta\langle K^\top y^k - K^\top y^{k-1}, x^k - x^\star\rangle \\
&\quad +2\theta L_{xy}\|y^k - y^{k-1}\|\|x^{k+1} - x^k\| \\
&\leq +2\langle K^\top y^{k+1} - K^\top y^k, x^{k+1} - x^\star\rangle \\
&\quad -2\theta\langle K^\top y^k - K^\top y^{k-1}, x^k - x^\star\rangle \\
&\quad +\theta L_{xy} C\|y^k - y^{k-1}\|^2 + \frac{\theta L_{xy}}{C}\|x^{k+1} - x^k\|^2. \qquad (65)
\end{aligned}
$$

Showing that $\Xi^{k+1} \leq \theta\Xi^k$ for all $k \geq 1$. Plugging the bounds (62) and (65) into the Lyapunov function (59), for any $k \geq 1$ we get

$$
\begin{aligned}
\Xi^{k+1} &\overset{(59)+(62)+(65)}{\leq} \frac{1}{\eta_x}\|x^k - x^\star\|^2 - \left(\frac{1}{\eta_x} - \frac{\theta L_{xy}}{C}\right)\|x^{k+1} - x^k\|^2 + (1 - \mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^k - y^\star\|^2 \\
&\quad -2\theta\langle K^\top y^k - K^\top y^{k-1}, x^k - x^\star\rangle + \theta L_{xy} C\|y^k - y^{k-1}\|^2 \\
&\overset{C=\theta\eta_x L_{xy}}{\leq} \frac{1}{\eta_x}\|x^k - x^\star\|^2 + (1 - \mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^k - y^\star\|^2 \\
&\quad -2\theta\langle K^\top y^k - K^\top y^{k-1}, x^k - x^\star\rangle + \theta^2 L_{xy}^2\eta_x\eta_y\frac{1}{\eta_y}\|y^k - y^{k-1}\|^2 \\
&\overset{(56)+(57)\ \&\ \theta\leq 1}{\leq} \frac{1}{\eta_x}\|x^k - x^\star\|^2 + (1 - \mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^k - y^\star\|^2 \\
&\quad -2\theta\langle K^\top y^k - K^\top y^{k-1}, x^k - x^\star\rangle + \theta\frac{1}{2\eta_y}\|y^k - y^{k-1}\|^2 \\
&\leq \max\left\{\frac{1}{1+\mu_x\eta_x}, 1 - \mu_{xy}^2\beta_y\eta_y\right\}\left((1+\mu_x\eta_x)\frac{1}{\eta_x}\|x^k - x^\star\|^2 + \frac{1}{\eta_y}\|y^k - y^\star\|^2\right) \\
&\quad +\theta\left(\frac{1}{2\eta_y}\|y^k - y^{k-1}\|^2 - 2\langle K^\top y^k - K^\top y^{k-1}, x^k - x^\star\rangle\right) \\
&\overset{(58)+(59)}{=} \theta\Xi^k.
\end{aligned}
$$

Showing that $\Xi^{k+1} \leq \theta\Xi^k$ for $k = 0$. This can be done using similar arguments as those used in the proof of Theorem 6.

□

D.3 Proof of Theorem 1 (informal)

We now provide the iteration complexity of Algorithm 1 as a corollary of Theorem 7. Note that in view of (60), we get

$$
k \geq \mathcal{O}\left(\frac{1}{1-\theta}\log\frac{1}{\varepsilon}\right) \qquad \Rightarrow \qquad \Xi^k \leq \varepsilon\Xi^0.
$$

Using the definition of θ given in (58), we have

$$\begin{aligned}
\mathcal{O}\left(\frac{1}{1-\theta}\log\frac{1}{\varepsilon}\right) &= \mathcal{O}\left(\max\left\{1+\frac{1}{\mu_x\eta_x}, \frac{1}{\mu_{xy}^2\beta_y\eta_y}\right\}\log\frac{1}{\varepsilon}\right) \\
&= \mathcal{O}\left(\max\left\{1+\frac{1}{\mu_x\eta_x}, \frac{L_x}{\mu_{xy}^2\eta_y}, \frac{L_{xy}^2}{\mu_{xy}^2}\right\}\log\frac{1}{\varepsilon}\right) \\
&= \mathcal{O}\left(\max\left\{1+\sqrt{\frac{L_x}{\mu_x}}\frac{L_{xy}}{\mu_{xy}}, \frac{L_{xy}^2}{\mu_{xy}^2}\right\}\log\frac{1}{\varepsilon}\right).
\end{aligned}$$

This proves the statement of Theorem 1 mentioned in Section 4 in the main body of the paper.

E Analysis of the Accelerated Primal-Dual Algorithm with Inexact Prox (Algorithm 2)

In this section we provide convergence analysis for our second new method, APDA with Inexact Prox (Algorithm 2). We start with statements of three lemmas, followed by the proof of the main theorem.

E.1 Three Lemmas

The first result is essentially a modification of Lemma 6, providing a bound on the distance of the primal iterates from the primal optimal solution. Compare to Lemma 6, these changes consist in using the definition of function Ψ^k (see the problem (15)) to prove this key fact.

Lemma 9. *Let $w^{\star k}$ be an exact solution to the problem* (15)*. Then under Assumptions 1, 2, we have*

$$\begin{aligned}\left(1+\frac{\mu_x\eta_x}{2}\right)\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 \quad \leq \quad & \frac{1}{\eta_x}\|x^k-x^\star\|^2+(2\eta_x+\mu_x\eta_x^2)\|\nabla\Psi^k(\hat{x}^k)\|^2-\frac{1}{2\eta_x}\|x^k-w^{\star k}\|^2\\ & -2\langle K^\top\bar{y}^k-K^\top y^\star,\hat{x}^k-x^\star\rangle-\frac{1}{L_x}\|\nabla G(\hat{x}^k)-\nabla G(x^\star)\|^2.\end{aligned}$$

Proof. In view of line 4 of Algorithm 2, which reads

$$x^{k+1}=x^k-\eta_x\left(\nabla G(\hat{x}^k)+K^\top\bar{y}^k\right), \tag{66}$$

and writing x^{k+1} as $x^k+(x^{k+1}-x^k)$, we get

$$\begin{aligned}\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 \quad &= \quad \frac{1}{\eta_x}\|x^k-x^\star\|^2+\frac{2}{\eta_x}\langle x^{k+1}-x^k,x^{k+1}-x^\star\rangle-\frac{1}{\eta_x}\|x^{k+1}-x^k\|^2\\ &\overset{(66)}{=} \quad \frac{1}{\eta_x}\|x^k-x^\star\|^2-\frac{2}{\eta_x}\langle\eta_x\left(\nabla G(\hat{x}^k)+K^\top\bar{y}^k\right),x^{k+1}-x^\star\rangle-\frac{1}{\eta_x}\left\|\eta_x\left(\nabla G(\hat{x}^k)+K^\top\bar{y}^k\right)\right\|^2\\ &= \quad \frac{1}{\eta_x}\|x^k-x^\star\|^2-2\langle\nabla G(\hat{x}^k)+K^\top\bar{y}^k,\hat{x}^k-x^\star\rangle-\eta_x\left\|\nabla G(\hat{x}^k)+K^\top\bar{y}^k\right\|^2\\ &\quad -\frac{2}{\eta_x}\left\langle\eta_x\left(\nabla G(\hat{x}^k)+K^\top\bar{y}^k\right),x^{k+1}-\hat{x}^k\right\rangle.\end{aligned}$$

Using the identity $-2\langle a,b\rangle=\|a\|^2+\|b\|^2-\|a+b\|^2$ to rewrite the second inner product from the previous equation, we get

$$\begin{aligned}\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 \quad &= \quad \frac{1}{\eta_x}\|x^k-x^\star\|^2-2\langle\nabla G(\hat{x}^k)+K^\top\bar{y}^k,\hat{x}^k-x^\star\rangle-\eta_x\|\nabla G(\hat{x}^k)+K^\top\bar{y}^k\|^2\\ &\quad +\frac{1}{\eta_x}\left(\left\|\eta_x\left(\nabla G(\hat{x}^k)+K^\top\bar{y}^k\right)\right\|^2+\|x^{k+1}-\hat{x}^k\|^2\right)\\ &\quad -\frac{1}{\eta_x}\left\|\eta_x\left(\nabla G(\hat{x}^k)+K^\top\bar{y}^k\right)+x^{k+1}-\hat{x}^k\right\|^2\\ &\overset{(66)}{=} \quad \frac{1}{\eta_x}\|x^k-x^\star\|^2-2\langle\nabla G(\hat{x}^k)-\nabla G(x^\star),\hat{x}^k-x^\star\rangle\\ &\quad -2\langle K^\top\bar{y}^k-K^\top y^\star,\hat{x}^k-x^\star\rangle+\frac{1}{\eta_x}\|x^{k+1}-\hat{x}^k\|^2-\frac{1}{\eta_x}\|x^k-\hat{x}^k\|^2,\end{aligned} \tag{67}$$

where in the last equation we applied the optimality condition (see (8))

$$\nabla G(x^\star)+K^\top y^\star=0.$$

Due to the fact that G is μ_x-strongly-convex and L_x-smooth, we can estimate the inner product $2\langle\nabla G(\hat{x}^k)-\nabla G(x^\star),\hat{x}^k-x^\star\rangle$ from below similarly as in the proof of Lemma 6:

$$2\langle\nabla G(\hat{x}^k)-\nabla G(x^\star),\hat{x}^k-x^\star\rangle\geq\mu_x\|\hat{x}^k-x^\star\|^2+\frac{1}{L_x}\|\nabla G(\hat{x}^k)-\nabla G(x^\star)\|^2. \tag{68}$$

Now, by plugging (68) into (67), we obtain

$$\begin{aligned}\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 &\leq \frac{1}{\eta_x}\|x^k-x^\star\|^2-\mu_x\|\hat{x}^k-x^\star\|^2-\frac{1}{L_x}\|\nabla G(\hat{x}^k)+\nabla G(x^\star)\|^2\\ &\quad+\frac{1}{\eta_x}\|x^{k+1}-\hat{x}^k\|^2-\frac{1}{\eta_x}\|x^k-\hat{x}^k\|^2-2\langle K^\top\bar{y}^k-K^\top y^\star,\hat{x}^k-x^\star\rangle.\end{aligned}$$

Applying the inequality $\|a-b\|^2\geq\frac{1}{2}\|a-c\|^2-\|b-c\|^2$ to $\|\hat{x}^k-x^\star\|^2$ and taking $c=x^{k+1}$, we obtain

$$\begin{aligned}\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 &\leq \frac{1}{\eta_x}\|x^k-x^\star\|^2-\frac{\mu_x}{2}\|x^{k+1}-x^\star\|^2-\frac{1}{L_x}\|\nabla G(\hat{x}^k)-\nabla G(x^\star)\|^2\\ &\quad+(1+\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1}-\hat{x}^k\|^2-\frac{1}{\eta_x}\|x^k-\hat{x}^k\|^2\\ &\quad-2\langle K^\top\bar{y}^k-K^\top y^\star,\hat{x}^k-x^\star\rangle\\ &\leq \frac{1}{\eta_x}\|x^k-x^\star\|^2-\frac{\mu_x}{2}\|x^{k+1}-x^\star\|^2-\frac{1}{L_x}\|\nabla G(\hat{x}^k)-\nabla G(x^\star)\|^2\\ &\quad+(1+\mu_x\eta_x)\frac{1}{\eta_x}\|x^{k+1}-\hat{x}^k\|^2-\frac{1}{2\eta_x}\|x^k-w^{\star k}\|^2+\frac{1}{\eta_x}\|\hat{x}^k-w^{\star k}\|^2\\ &\quad-2\langle K^\top\bar{y}^k-K^\top y^\star,\hat{x}^k-x^\star\rangle,\end{aligned}\tag{69}$$

where in the last inequality we used the bound $\|a-b\|^2\geq\frac{1}{2}\|a-c\|^2-\|b-c\|^2$ to estimate $\frac{1}{\eta_x}\|x^k-\hat{x}^k\|^2$. From line 3 of Algorithm 2, according to the definition of function $\Psi^k(x)$, we get

$$\begin{aligned}\nabla\Psi^k(\hat{x}^k) &= \nabla G(\hat{x}^k)+K^\top\bar{y}^k+\frac{1}{\eta_x}\left(\hat{x}^k-x^k\right)\\ &= \nabla G(\hat{x}^k)+K^\top\bar{y}^k+\frac{1}{\eta_x}\left(\hat{x}^k-x^{k+1}\right)+\frac{1}{\eta_x}\left(x^{k+1}-x^k\right)\\ &\overset{(66)}{=} \frac{1}{\eta_x}\left(\hat{x}^k-x^{k+1}\right).\end{aligned}\tag{70}$$

Finally, substituting (70) into (69), we gain

$$\begin{aligned}\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2 &= \frac{1}{\eta_x}\|x^k-x^\star\|^2-\frac{\mu_x}{2}\|x^{k+1}-x^\star\|^2-\frac{1}{L_x}\|\nabla G(\hat{x}^k)-\nabla G(x^\star)\|^2\\ &\quad+\left(2\eta_x+\mu_x\eta_x^2\right)\|\nabla\Psi^k(\hat{x}^k)\|^2-\frac{1}{2\eta_x}\|x^k-w^{\star k}\|^2\\ &\quad-2\langle K^\top\bar{y}^k-K^\top y^\star,\hat{x}^k-x^\star\rangle,\end{aligned}$$

where we also use $\frac{1}{\eta_x}$-strong convexity of function $\Psi^k(x)$ (see problem (15)). □

The second result offers a bound on the distance between the dual itarates and the dual optimal solution as Lemma 8. Since the line 4 of Algorithm 1 coincides with the line 5 of Algorithm 2, the statement and proof of the following lemma coincides with Lemma 8 and its proof with the only change of replacing x^{k+1} with $\hat{x}^k$.

Lemma 10. *Then under Assumptions 2, 3, and 4 be satisfied, and choose*

$$\beta_y\leq\frac{1}{2L_{xy}^2\eta_y}.\tag{71}$$

Then the iterates of APDA with Inexact Prox (Algorithm 2) for all $k\geq 0$ satisfy

$$\begin{aligned}\frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 &\leq (1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^k-y^\star\|^2-\frac{1}{2\eta_y}\|y^{k+1}-y^k\|^2+2\langle K^\top y^{k+1}-K^\top y^\star,\hat{x}^k-x^\star\rangle\\ &\quad+\beta_y\|\nabla G(\hat{x}^k)-\nabla G(x^\star)\|^2.\end{aligned}$$

Proof. The proof is identical to the proof of Lemma 8, with the sole distinction that x^{k+1} should be replaced everywhere by $\hat{x}^k$. □

The third result is a technical lemma, which offers a bound on an inner product .

Lemma 11. *Let $w^{\star k}$ be an exact solution to the problem* (15)*. Then the following inequality holds*

$$\begin{aligned}-2\theta\langle K^\top y^k - K^\top y^{k-1}, \hat{x}^k - \hat{x}^{k-1}\rangle &\leq 16L_{xy}^2\theta\eta_x\|y^k - y^{k-1}\|^2 + \frac{\theta}{4\eta_x}\|\hat{x}^k - w^{\star k}\|^2 \\ &\quad + \frac{\theta}{4\eta_x}\|x^k - w^{\star k}\|^2 + \frac{\theta\eta_x}{8}\|\nabla\Psi^{k-1}(\hat{x}^{k-1})\|^2.\end{aligned}$$

Proof. Using the Cauchy-Schwarz inequality, the definition of L_{xy} as the norm of K (see (2)), and applying Young's inequality, we can bound the inner product by

$$\begin{aligned}-2\theta\langle K^\top y^k - K^\top y^{k-1}, \hat{x}^k - \hat{x}^{k-1}\rangle &\leq 2\theta\|K^\top y^k - K^\top y^{k-1}\|\|\hat{x}^k - \hat{x}^{k-1}\| \\ &\leq L_{xy}\theta C\|y^k - y^{k-1}\|^2 + \frac{L_{xy}\theta}{C}\|\hat{x}^k - \hat{x}^{k-1}\|^2 \\ &\leq L_{xy}\theta C\|y^k - y^{k-1}\|^2 + \frac{2L_{xy}\theta}{C}\|\hat{x}^k - x^k\|^2 \\ &\quad + \frac{2L_{xy}\theta}{C}\|x^k - \hat{x}^{k-1}\|^2\end{aligned}$$

for any $C > 0$. Taking $C = 16L_{xy}\eta_x$, we obtain

$$\begin{aligned}-2\theta\langle K^\top y^k - K^\top y^{k-1}, \hat{x}^k - \hat{x}^{k-1}\rangle &\leq L_{xy}\theta C\|y^k - y^{k-1}\|^2 + \frac{4L_{xy}\theta}{C}\|\hat{x}^k - w^{\star k}\|^2 \\ &\quad \frac{4L_{xy}\theta}{C}\|\hat{x}^k - w^{\star k}\|^2 + \frac{2L_{xy}\theta}{C}\|x^k - \hat{x}^{k-1}\|^2 \\ &= 16L_{xy}^2\theta\eta_x\|y^k - y^{k-1}\|^2 + \frac{\theta}{4\eta_x}\|\hat{x}^k - w^{\star k}\|^2 \\ &\quad + \frac{\theta}{4\eta_x}\|x^k - w^{\star k}\|^2 + \frac{\theta\eta_x}{8}\|\nabla\Psi^{k-1}(\hat{x}^{k-1})\|^2,\end{aligned}$$

where in the last equation we use (70) for $k-1$, which reads $\nabla\Psi^k(\hat{x}^{k-1}) = \frac{1}{\eta_x}\left(\hat{x}^{k-1} - x^k\right)$. □

E.2 Detailed theorem

Theorem 8 (Convergence of APDA with Inexact Prox; formal)**.** *Let Assumptions 1, 2, 3 and 4 hold and select the various parameters of the method as follows:*

$$\beta_y = \min\left\{\frac{1}{L_x}, \frac{1}{2L_{xy}^2\eta_y}\right\}; \tag{72}$$

$$T = \sqrt[\alpha]{20A}\left(1 + \sqrt{\frac{L_x}{\mu_x}}\right)^{2/\alpha}; \tag{73}$$

$$\eta_x = \frac{1}{4\sqrt{L_x\mu_x}}\frac{\mu_{xy}}{L_{xy}}; \tag{74}$$

$$\eta_y = \frac{\sqrt{L_x\mu_x}}{8L_{xy}\mu_{xy}}; \tag{75}$$

$$\theta = \max\left\{\frac{2}{2 + \mu_x\eta_x}, 1 - \mu_{xy}^2\beta_y\eta_y\right\}, \tag{76}$$

Then for the Lyapunov function defined for $k \geq 0$ via

$$\begin{aligned}\Delta^{k+1} &\stackrel{def}{=} \left(1 + \frac{\mu_x\eta_x}{2}\right)\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 + \frac{1}{\eta_y}\|y^{k+1} - y^\star\|^2 + \frac{1}{2\eta_y}\|y^{k+1} - y^k\|^2 \\ &\quad + \frac{1}{8\eta_x}\|x^k - w^{\star k}\|^2 - 2\langle K^\top y^{k+1} - K^\top y^k, \hat{x}^k - x^\star\rangle,\end{aligned} \tag{77}$$

and for $k=0$ via

$$\Delta^0 \stackrel{def}{=} \left(1+\frac{\mu_x\eta_x}{2}\right)\frac{1}{\eta_x}\|x^0-x^\star\|^2+\frac{1}{\eta_y}\|y^0-y^\star\|^2, \tag{78}$$

we have

$$\frac{1}{2\eta_x}\|x^{k+1}-x^\star\|^2+\frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 \le \Delta^k \le \theta^k\Delta^0, \quad \forall\, k\ge 0. \tag{79}$$

Proof. We denote $\mathcal{V}^{k+1}$ as follows:

$$\mathcal{V}^{k+1} = \left(1+\frac{\mu_x\eta_x}{2}\right)\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2+\frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2+\frac{1}{2\eta_y}\|y^{k+1}-y^k\|^2. \tag{80}$$

Combining Lemmas 9, 10 and (72), we obtain

$$\begin{aligned}
\mathcal{V}^{k+1} \le\ & \frac{1}{\eta_x}\|x^k-x^\star\|^2+(1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^k-y^\star\|^2+(2\eta_x+\mu_x\eta_x^2)\|\nabla\Psi^k(\hat{x}^k)\|^2\\
& -\frac{1}{2\eta_x}\|x^k-w^{\star k}\|^2+\left(\beta_y-\frac{1}{L_x}\right)\|\nabla G(\hat{x}^k)-\nabla G(x^\star)\|^2\\
& -2\langle K^\top\bar{y}^k-K^\top y^\star,\hat{x}^k-x^\star\rangle+2\langle K^\top y^{k+1}-K^\top y^\star,\hat{x}^k-x^\star\rangle \qquad (81)\\
\le\ & \frac{1}{\eta_x}\|x^k-x^\star\|^2+(1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^k-y^\star\|^2+(2\eta_x+\mu_x\eta_x^2)\|\nabla\Psi^k(\hat{x}^k)\|^2\\
& -\frac{1}{2\eta_x}\|x^k-w^{\star k}\|^2+2\langle K^\top y^{k+1}-K^\top y^k,\hat{x}^k-x^\star\rangle\\
& -2\theta\langle K^\top y^k-K^\top y^{k-1},\hat{x}^{k-1}-x^\star\rangle-2\theta\langle K^\top y^k-K^\top y^{k-1},\hat{x}^k-\hat{x}^{k-1}\rangle,
\end{aligned}$$

where in last inequality the line 6 from Algorithm 2 is used. Next, according to Lemma 11, we gain

$$\begin{aligned}
\mathcal{V}^{k+1} \le\ & \frac{1}{\eta_x}\|x^k-x^\star\|^2+(1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^k-y^\star\|^2+(2\eta_x+\mu_x\eta_x^2)\|\nabla\Psi^k(\hat{x}^k)\|^2-\frac{1}{2\eta_x}\|x^k-w^{\star k}\|^2\\
& +2\langle K^\top y^{k+1}-K^\top y^k,\hat{x}^k-x^\star\rangle-2\theta\langle K^\top y^k-K^\top y^{k-1},\hat{x}^{k-1}-x^\star\rangle\\
& +16\theta L_{xy}^2\eta_x\|y^k-y^{k-1}\|^2+\frac{\theta}{4\eta_x}\|\hat{x}^k-w^{\star k}\|^2+\frac{\theta}{4\eta_x}\|x^k-w^{\star k}\|^2+\frac{\theta\eta_x}{8}\|\nabla\Psi_{\eta_x}^{k-1}(\hat{x}^{k-1})\|^2\\
\le\ & \frac{1}{\eta_x}\|x^k-x^\star\|^2+(1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^k-y^\star\|^2+16\theta L_{xy}^2\eta_x\eta_y\frac{1}{\eta_y}\|y^k-y^{k-1}\|^2\\
& +(2\eta_x+\mu_x\eta_x^2)\|\nabla\Psi^k(\hat{x}^k)\|^2-\frac{1}{4\eta_x}\|x^k-w^{\star k}\|^2+\frac{\theta}{4\eta_x}\|\hat{x}^k-w^{\star k}\|^2+\frac{\theta\eta_x}{8}\|\nabla\Psi_{\eta_x}^{k-1}(\hat{x}^{k-1})\|^2\\
& +2\langle K^\top y^{k+1}-K^\top y^k,\hat{x}^k-x^\star\rangle-2\theta\langle K^\top y^k-K^\top y^{k-1},\hat{x}^{k-1}-x^\star\rangle.
\end{aligned}$$

Using definition of function Ψ^k and its $\frac{1}{\eta_x}$-strong convexity in the following form

$$\|\nabla\Psi^k(\hat{x}^k)\|^2 \ge \frac{1}{\eta_x}\|\hat{x}^k-w^{\star k}\|^2,$$

and assuming that $32L_{xy}^2\eta_x\eta_y\le 1$ and $0<\theta\le 1$, we get

$$\begin{aligned}
\mathcal{V}^{k+1} \le\ & \frac{1}{\eta_x}\|x^k-x^\star\|^2+(1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^k-y^\star\|^2+\frac{\theta}{2\eta_y}\|y^k-y^{k-1}\|^2\\
& +(2\eta_x+\mu_x\eta_x^2)\|\nabla\Psi^k(\hat{x}^k)\|^2-\frac{1}{4\eta_x}\|x^k-w^{\star k}\|^2+\frac{\theta\eta_x}{4}\|\nabla\Psi^k(\hat{x}^k)\|^2+\frac{\theta\eta_x}{8}\|\nabla\Psi_{\eta_x}^{k-1}(\hat{x}^{k-1})\|^2\\
& +2\langle K^\top y^{k+1}-K^\top y^k,\hat{x}^k-x^\star\rangle-2\theta\langle K^\top y^k-K^\top y^{k-1},\hat{x}^{k-1}-x^\star\rangle\\
\stackrel{0<\theta\le 1}{\le}\ & \frac{1}{\eta_x}\|x^k-x^\star\|^2+(1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^k-y^\star\|^2+\frac{\theta}{2\eta_y}\|y^k-y^{k-1}\|^2\\
& +\left(\frac{9}{4}\eta_x+\mu_x\eta_x^2\right)\|\nabla\Psi^k(\hat{x}^k)\|^2-\frac{1}{4\eta_x}\|x^k-w^{\star k}\|^2+\frac{\theta\eta_x}{8}\|\nabla\Psi_{\eta_x}^{k-1}(\hat{x}^{k-1})\|^2\\
& +2\langle K^\top y^{k+1}-K^\top y^k,\hat{x}^k-x^\star\rangle-2\theta\langle K^\top y^k-K^\top y^{k-1},\hat{x}^{k-1}-x^\star\rangle.
\end{aligned}$$

To estimate $\|\nabla\Psi^k(\hat{x}^k)\|^2$ we apply Lemma 17 for the problem (15) as follows (see (18)):

$$\|\nabla\Psi^k(\hat{x}^k)\|^2 \leq \frac{A\left(1+\eta_x L_x\right)^2\|x^k - w^{\star k}\|^2}{\eta_x^2 T^\alpha}.$$

According to (18), (73) and (74), we obtain

$$\begin{aligned}
\mathcal{V}^{k+1} \overset{(18)}{\leq} & \frac{1}{\eta_x}\|x^k - x^\star\|^2 + (1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^k - y^\star\|^2 + \frac{\theta}{2\eta_y}\|y^k - y^{k-1}\|^2 \\
& + \left(\frac{9}{4}\eta_x + \mu_x\eta_x^2\right)\frac{A\left(1+\eta_x L_x\right)^2}{\eta_x^2 T^\alpha}\|x^k - w^{\star k}\|^2 - \frac{1}{4\eta_x}\|x^k - w^{\star k}\|^2 \\
& + \frac{\theta\eta_x}{8}\frac{A\left(1+\eta_x L_x\right)^2}{\eta_x^2 T^\alpha}\|x^{k-1} - w^{\star k-1}\|^2 \\
& + 2\langle K^\top y^{k+1} - K^\top y^k, \hat{x}^k - x^\star\rangle - 2\theta\langle K^\top y^k - K^\top y^{k-1}, \hat{x}^{k-1} - x^\star\rangle \\
\overset{(73)+(74)}{\leq} & \frac{1}{\eta_x}\|x^k - x^\star\|^2 + (1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^k - y^\star\|^2 + \frac{\theta}{2\eta_y}\|y^k - y^{k-1}\|^2 \\
& + \frac{\theta}{8\eta_x}\|x^{k-1} - w^{\star k-1}\|^2 - \frac{1}{8\eta_x}\|x^k - w^{\star k}\|^2 \\
& + 2\langle K^\top y^{k+1} - K^\top y^k, \hat{x}^k - x^\star\rangle - 2\theta\langle K^\top y^k - K^\top y^{k-1}, \hat{x}^{k-1} - x^\star\rangle.
\end{aligned}$$

Showing that $\Delta^{k+1} \leq \theta\Delta^k$ for $k \geq 1$. Using the definition of Δ^{k+1} (see (77)), we have

$$\begin{aligned}
\Delta^{k+1} \leq & \frac{1}{\eta_x}\|x^k - x^\star\|^2 + (1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^k - y^\star\|^2 + \frac{\theta}{2\eta_y}\|y^k - y^{k-1}\|^2 \\
& + \frac{\theta}{8\eta_x}\|x^{k-1} - w^{\star k-1}\|^2 - 2\theta\langle K^\top y^k - K^\top y^{k-1}, \hat{x}^{k-1} - x^\star\rangle \\
\leq & \max\left\{\frac{2}{2+\mu_x\eta_x}, 1-\mu_{xy}^2\beta_y\eta_y\right\}\left(\left(1+\frac{\mu_x\eta_x}{2}\right)\frac{1}{\eta_x}\|x^k - x^\star\|^2 + \frac{1}{\eta_y}\|y^k - y^\star\|^2\right) \\
& + \theta\left(\frac{1}{2\eta_y}\|y^k - y^{k-1}\|^2 + \frac{1}{8\eta_x}\|x^{k-1} - w^{\star k-1}\|^2 - 2\langle K^\top y^k - K^\top y^{k-1}, \hat{x}^{k-1} - x^\star\rangle\right) \\
\overset{(76)}{\leq} & \theta\Delta^k,
\end{aligned}$$

where in the last inequality we take $\theta = \max\left\{\frac{2}{2+\mu_x\eta_x}, 1-\mu_{xy}^2\beta_y\eta_y\right\}$ (see (76)).

Showing that $\Delta^{k+1} \leq \theta\Delta^k$ for $k = 0$. We start with inequality (81) for $k = 0$:

$$\begin{aligned}
\mathcal{V}^1 \overset{(81)}{\leq} & \frac{1}{\eta_x}\|x^0 - x^\star\|^2 + (1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^0 - y^\star\|^2 + (2\eta_x + \mu_x\eta_x^2)\|\nabla\Psi^k(\hat{x}^0)\|^2 \\
& - \frac{1}{2\eta_x}\|x^0 - w^{\star 0}\|^2 + \left(\beta_y - \frac{1}{L_x}\right)\|\nabla G(\hat{x}^0) - \nabla G(x^\star)\|^2 \\
& - 2\langle K^\top\bar{y}^0 - K^\top y^\star, \hat{x}^0 - x^\star\rangle + 2\langle K^\top y^1 - K^\top y^\star, \hat{x}^0 - x^\star\rangle \\
\leq & \frac{1}{\eta_x}\|x^0 - x^\star\|^2 + (1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^0 - y^\star\|^2 - \frac{1}{2\eta_x}\|x^0 - w^{\star 0}\|^2 \\
& + (2\eta_x + \mu_x\eta_x^2)\|\nabla\Psi^k(\hat{x}^0)\|^2 + 2\langle K^\top y^1 - K^\top y^0, \hat{x}^0 - x^\star\rangle,
\end{aligned} \tag{82}$$

where in the last inequality we take $\beta_y \leq \frac{1}{L_x}$ (see (72)). Now, we apply Lemma 1 to the problem (15) for $k = 0$ (see (18)):

$$\|\nabla\Psi^k(\hat{x}^0)\|^2 \leq \frac{A(1+\eta_x L_x)^2\|x^0 - w^{\star 0}\|^2}{\eta_x T^\alpha}. \tag{83}$$

Substituting (83) into (82), we can get

$$\begin{aligned}
\mathcal{V}^1 &\leq \frac{1}{\eta_x}\|x^0 - x^\star\|^2 + (1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^0 - y^\star\|^2 - \frac{1}{2\eta_x}\|x^0 - w^{\star 0}\|^2 \\
&\quad +(2\eta_x + \mu_x\eta_x^2)\frac{A(1+\eta_x L_x)^2}{\eta_x T^\alpha}\|x^0 - w^{\star 0}\|^2 + 2\langle K^\top y^1 - K^\top y^0, \hat{x}^0 - x^\star\rangle \\
&\overset{(73)}{\leq} \frac{1}{\eta_x}\|x^0 - x^\star\|^2 + (1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^0 - y^\star\|^2 - \frac{1}{8\eta_x}\|x^0 - w^{\star 0}\|^2 \\
&\quad +2\langle K^\top y^1 - K^\top y^0, \hat{x}^0 - x^\star\rangle, \qquad (84)
\end{aligned}$$

where in the last inequality we take T according to (73). According to (77) and (78), we have

$$\begin{aligned}
\Delta^1 &\overset{(77)}{=} \mathcal{V}^1 + \frac{1}{8\eta_x}\|x^0 - w^{\star 0}\|^2 - 2\langle K^\top y^1 - K^\top y^0, \hat{x}^0 - x^\star\rangle \\
&\overset{(84)}{\leq} \frac{1}{\eta_x}\|x^0 - x^\star\|^2 + (1-\mu_{xy}^2\beta_y\eta_y)\frac{1}{\eta_y}\|y^0 - y^\star\|^2 \\
&\leq \theta\left(\left(1+\frac{\mu_x\eta_x}{2}\right)\frac{1}{\eta_x}\|x^0 - x^\star\|^2 + \frac{1}{\eta_y}\|y^0 - y^\star\|^2\right) \\
&= \theta\Delta^0,
\end{aligned}$$

where in the last inequality we used the inequalities $1 \leq \theta\left(1+\frac{\mu_x\eta_x}{2}\right)$ and $1-\mu_{xy}^2\beta_y\eta_y \leq \theta$, which follow from (76).

Showing that $\Delta^{k+1} \geq 0$ for $k \geq 0$. Finally, we need to show that $\Delta^k \geq 0$ for every k. This is obvious for $k = 0$ from . Using the Cauchy-Schwarz inequality, the definition of L_{xy} as the norm of K (see (2)), and applying Young's inequality, we get

$$\begin{aligned}
\Delta^{k+1} &\overset{(77)}{=} \left(1+\frac{\mu_x\eta_x}{2}\right)\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 + \frac{1}{\eta_y}\|y^{k+1} - y^\star\|^2 + \frac{1}{2\eta_y}\|y^{k+1} - y^k\|^2 \\
&\quad +\frac{1}{8\eta_x}\|x^k - w^{\star k}\|^2 - 2\langle K^\top y^{k+1} - K^\top y^k, \hat{x}^k - x^\star\rangle \\
&\overset{(2)}{\geq} \left(1+\frac{\mu_x\eta_x}{2}\right)\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 + \frac{1}{\eta_y}\|y^{k+1} - y^\star\|^2 + \frac{1}{2\eta_y}\|y^{k+1} - y^k\|^2 \\
&\quad +\frac{1}{8\eta_x}\|x^k - w^{\star k}\|^2 - 2L_{xy}\|y^{k+1} - y^k\|\|\hat{x}^k - x^\star\| \\
&\geq \left(1+\frac{\mu_x\eta_x}{2}\right)\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 + \frac{1}{\eta_y}\|y^{k+1} - y^\star\|^2 + \left(\frac{1}{2} - L_{xy}B\eta_y\right)\frac{1}{\eta_y}\|y^{k+1} - y^k\|^2 \\
&\quad +\frac{1}{8\eta_x}\|x^k - w^{\star k}\|^2 - \frac{L_{xy}}{B}\|\hat{x}^k - x^\star\|^2 \\
&\geq \left(1+\frac{\mu_x\eta_x}{2}\right)\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 + \frac{1}{\eta_y}\|y^{k+1} - y^\star\|^2 + \left(\frac{1}{2} - L_{xy}B\eta_y\right)\frac{1}{\eta_y}\|y^{k+1} - y^k\|^2 \\
&\quad +\frac{1}{8\eta_x}\|x^k - w^{\star k}\|^2 - \frac{2L_xy}{B}\|\hat{x}^k - x^{k+1}\|^2 - \frac{2L_xy}{B}\|x^{k+1} - x^\star\|^2.
\end{aligned}$$

for any $B > 0$. According to the definition of function Ψ^k (see 15), selecting $B = \frac{1}{2L_{xy}\eta_y}$, we get

$$\begin{aligned}
\Delta^{k+1} &\geq \left(1+\frac{\mu_x\eta_x}{2}\right)\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 + \frac{1}{\eta_y}\|y^{k+1} - y^\star\|^2 + \frac{1}{8\eta_x}\|x^k - w^{\star k}\|^2 \\
&\quad -4L_{xy}^2\eta_y\eta_x\frac{1}{\eta_x}\|\eta_x\nabla\Psi_{\eta_x}^k(\hat{x}^k)\|^2 - 4L_{xy}^2\eta_y\eta_x\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 \\
&= \left(1 - 4L_{xy}^2\eta_y\eta_x + \frac{\mu_x\eta_x}{2}\right)\frac{1}{\eta_x}\|x^{k+1} - x^\star\|^2 \\
&\quad +\frac{1}{\eta_y}\|y^{k+1} - y^\star\|^2 + \frac{1}{8\eta_x}\|x^k - w^{\star k}\|^2 - 4L_{xy}^2\eta_y\eta_x\frac{1}{\eta_x}\|\eta_x\nabla\Psi_{\eta_x}^k(\hat{x}^k)\|^2. \qquad (85)
\end{aligned}$$

Choosing stepsizes η_x, η_y according to (74) and (75), we can derive the following inequality:

$$32L_{xy}^2\eta_x\eta_y \leq 1. \tag{86}$$

Now, plugging (86) and (18) into (85), we obtain

$$\begin{aligned}
\Delta^{k+1} &\overset{(86)}{\geq} \left(\frac{7}{8}+\frac{\mu_x\eta_x}{2}\right)\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2+\frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 \\
&\quad +\frac{1}{8\eta_x}\|x^k-w^{\star k}\|^2-\frac{1}{8\eta_x}\|\eta_x\nabla\Psi_{\eta_x}^k(\hat{x}^k)\|^2 \\
&\overset{(18)}{\geq} \left(\frac{7}{8}+\frac{\mu_x\eta_x}{2}\right)\frac{1}{\eta_x}\|x^{k+1}-x^\star\|^2+\frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2 \\
&\quad +\frac{1}{8\eta_x}\|x^k-w^{\star k}\|^2-\frac{1}{8\eta_x}\frac{C\left(1+\eta_xL_x\right)^2}{T^\alpha}\|x^k-w^{\star k}\|^2 \\
&\overset{(73)}{\geq} \frac{1}{2\eta_x}\|x^{k+1}-x^\star\|^2+\frac{1}{\eta_y}\|y^{k+1}-y^\star\|^2,
\end{aligned}$$

where in last inequality we use (73) to evaluate two last terms from below by zero. □

E.3 Proof of Theorem 2

To prove the informal result (19), we simply apply the above theorem. Using the definition of θ, the theorem implies that

$$k \geq \frac{1}{1-\theta}\log\frac{1}{\varepsilon} \quad \Rightarrow \quad \Delta^k \leq \varepsilon\Delta^0.$$

It remains to remark that the number of outer iterations to find ε-solution is equal to

$$\begin{aligned}
\mathcal{O}\left(\frac{1}{1-\theta}\log\frac{1}{\varepsilon}\right) &= \mathcal{O}\left(\max\left\{1+\frac{2}{\mu_x\eta_x},\frac{1}{\mu_{xy}^2\beta_y\eta_y}\right\}\log\frac{1}{\varepsilon}\right) \\
&= \mathcal{O}\left(\max\left\{1+\frac{1}{\mu_x\eta_x},\frac{L_x}{\mu_{xy}^2\eta_y},\frac{L_{xy}^2}{\mu_{xy}^2}\right\}\log\frac{1}{\varepsilon}\right) \\
&= \mathcal{O}\left(\max\left\{1+\sqrt{\frac{L_x}{\mu_x}}\frac{L_{xy}}{\mu_{xy}},\frac{L_{xy}^2}{\mu_{xy}^2}\right\}\log\frac{1}{\varepsilon}\right).
\end{aligned}$$

In other words, the number of computations of prox is equal to

$$\sharp\text{prox} = \mathcal{O}\left(\max\left\{1+\sqrt{\frac{L_x}{\mu_x}}\frac{L_{xy}}{\mu_{xy}},\frac{L_{xy}^2}{\mu_{xy}^2}\right\}\log\frac{1}{\varepsilon}\right),$$

and the number of gradient evaluations is

$$\begin{aligned}
\sharp\nabla G &= k\cdot T \\
&= \mathcal{O}\left(\max\left\{1+\sqrt{\frac{L_x}{\mu_x}}\frac{L_{xy}}{\mu_{xy}},\frac{L_{xy}^2}{\mu_{xy}^2}\right\}\log\frac{1}{\varepsilon}\right)\cdot\mathcal{O}\left(\left(\frac{L_x}{\mu_x}\right)^{1/\alpha}\right) \\
&= \mathcal{O}\left(\max\left\{\left(\frac{L_x}{\mu_x}\right)^{1/\alpha}+\left(\frac{L_x}{\mu_x}\right)^{\frac{2+\alpha}{2\alpha}}\frac{L_{xy}}{\mu_{xy}},\left(\frac{L_x}{\mu_x}\right)^{1/\alpha}\frac{L_{xy}^2}{\mu_{xy}^2}\right\}\log\frac{1}{\varepsilon}\right).
\end{aligned}$$

Algorithm 5 APDA with Inexact Prox and Accelerated Gossip

Input: Initial point $(\mathbf{x}^0, \mathbf{y}^0) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, $\bar{\mathbf{y}}^0 = \mathbf{y}^0$; Step sizes $\eta_x, \eta_y, \beta_y > 0$, $\theta \in [0,1]$; Number of inner iterations T; Number of iterations of Accelerated Gossip N;
for $k = 0, 1, 2, \ldots$ **do**
  Find $\hat{\mathbf{x}}^k$ as a final point of T iteration of method $\mathcal{M}$ for following problem:

$$\min_{\mathbf{x} \in \mathbb{R}^{d_x}} \left\{ P(\mathbf{x}) + \frac{1}{2\eta_x} \left\| \mathbf{x} - \left(\mathbf{x}^k - \eta_x \mathrm{AG}(W, \bar{\mathbf{y}}^k, N) \right) \right\|^2 \right\} \tag{87}$$

  $\mathbf{x}^{k+1} = \mathbf{x}^k - \eta_x \left(\nabla P(\hat{\mathbf{x}}^k) + \mathrm{AG}(W, \bar{\mathbf{y}}^k, N) \right)$
  $\mathbf{y}^{k+1} = \mathbf{y}^k + \eta_y \left(\mathrm{AG}(W, \hat{\mathbf{x}}^k, N) - \beta_y \mathrm{AG}(W, \mathrm{AG}(W, \mathbf{y}^k, N) + \nabla P(\hat{\mathbf{x}}^k), N) \right)$
  $\bar{\mathbf{y}}^{k+1} = \mathbf{y}^{k+1} + \theta \left(\mathbf{y}^{k+1} - \mathbf{y}^k \right)$
end for
procedure $\mathrm{AG}(W, \mathbf{x}, N)$ = Accelerated Gossip
  Set $a_0 = 1$, $a_1 = c_2$, $\mathbf{x}_0 = \mathbf{x}$, $\mathbf{x}_1 = c_2 \left(I - c_3 W \right) \mathbf{x}$
  for $i = 1, \ldots, N-1$ **do**
    $a_{i+1} = 2c_2 a_i - a_{i-1}$, $\mathbf{x}_{i+1} = 2c_2 \left(I - c_2 W \right) \mathbf{x}_i - \mathbf{x}_{i-1}$
  end for
  return $\mathbf{x} - \frac{\mathbf{x}_N}{a_N}$
end procedure

F Analysis of the Accelerated Primal-Dual Algorithm with Inexact Prox and Accelerated Gossip (Algorithm 5)

F.1 Gossip matrices

In Algorithm 5 we work with a more general notion of a gossip matrix (beyond graph Laplacians), defined next.

Assumption 6 (see [Scaman et al., 2017]). *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a connected communication network. The communication process is represented via multiplication by a matrix $\hat{W} \in \mathbb{R}^{n \times n}$ which satisfies the following conditions:*

- *$\hat{W}$ is symmetric,*
- *$\hat{W}$ is positive semi-definite,*
- *the kernel of $\hat{W}$ satisfies $\mathrm{Ker}\, \hat{W} = \mathrm{span}\{(1, \ldots, 1)^\top \in \mathbb{R}^n\}$, and*
- *$\hat{W}_{i,j} \neq 0$ if and only if $i = j$ or $(i,j) \in \mathcal{E}$.*

It is easy to see that $\sigma(\hat{W}) \subset \sigma(W)$, where $\sigma(\cdot)$ denotes the spectrum, and $W = \hat{W} \otimes I_{dn}$.

F.2 Theorem 3 follows from Theorem 2

Our main result for Algorithm 5, i.e., Theorem 3, follows from Theorem 2. Below we state it more formally.

Theorem 9 (Formal version of Theorem 3; convergence for Algorithm 5). *Let us invoke the same assumptions as those made in Theorem 2. Additionally, let Assumption 6 hold. Assume that the auxiliary problem* (87) *is solved by one of the methods from Lemma 1. Set the parameters N, c_1, c_2, c_3 to*

$$N = \lfloor \sqrt{\chi} \rfloor, \quad c_1 = \frac{\sqrt{\chi} - 1}{\sqrt{\chi} + 1}, \quad c_2 = \frac{\sqrt{\chi} + 1}{\sqrt{\chi} - 1}, \quad c_3 = \frac{2\chi}{(1 + \chi)\, \lambda_{\max}(W)}$$

and let

$$\lambda_1 = 1 + \frac{2c_1^N}{1 + c_1^{2N}}, \quad \lambda_2 = 1 - \frac{2c_1^N}{1 + c_1^{2N}},$$

$$\beta_y = \min\left\{\frac{1}{L_x}, \frac{1}{2\lambda_1^2\eta_y}\right\}, \quad \eta_x = \frac{1}{2\sqrt{L_x\mu_x}}\frac{\lambda_2}{\lambda_1}, \quad \eta_y = \frac{\sqrt{L_x\mu_x}}{\sqrt{2}\lambda_1\lambda_2},$$

$$\theta = \max\left\{\frac{2}{2+\mu_x\eta_x}, 1-\lambda_2^2\beta_y\eta_y\right\},$$

$$T = \sqrt[\alpha]{20A}\left(1+\sqrt{\frac{L_x}{\mu_x}}\right)^{2/\alpha}.$$

Then for the Lyapunov function Δ^k from Theorem 8, we have

$$0 \leq \Delta^k \leq \theta^k \Delta^0 \qquad \forall k \geq 0,$$

Moreover, for every $\varepsilon > 0$, Algorithm 5 finds $(\mathbf{x}^k, \mathbf{y}^k)$ for which $\Delta^k \leq \varepsilon\Delta^0$ in at most $\mathcal{O}\left(\kappa^{\frac{2+\alpha}{2\alpha}} \log\left(1/\varepsilon\right)\right)$ gradient computations and at most $\mathcal{O}\left(\sqrt{\kappa\chi}\log\left(1/\varepsilon\right)\right)$ communication rounds.

Proof. The main idea of the proof is this methods is to show two following things: i) Theorem 2 holds true with some replacements, i.e. $L_{xy} \to \lambda_1$, $\mu_{xy} \to \lambda_2$, where λ_1, λ_2 is upper bound and lower bound of the spectrum of a matrix, which is defined below; ii) Estimate λ_1 and λ_2.

The proof of Theorem 9 is similar to the proof of Corollary 1 from [Kovalev et al., 2020]. According to the proof from [Kovalev et al., 2020], we need to replace

$$L_{xy} = \lambda_{\max}(\sqrt{W}) \to \lambda_1 = 1 + \frac{2c_1^N}{1+c_1^{2N}}, \quad \mu_{xy} = \lambda_{\min}^+(\sqrt{W}) \to \lambda_2 = 1 - \frac{2c_1^N}{1+c_1^{2N}}.$$

□

G Proof of Lemma 1

Proof. We consider the three options separately. We will use f instead of Ψ and x instead of w in the proof. Let $T = 2K$.

(i) **K iterations of GD followed by K iterations GD.** We consider the GD method with stepsize $\gamma = \frac{1}{L}$, which performs the iterations

$$x^{k+1} = x^k - \gamma \nabla f(x^k) = x^k - \frac{1}{L}\nabla f(x^k),$$

where L is the smoothness constant of f. We assume that f is convex.

• **Gradient decreases monotonically.** According to the chain of inequalities

$$\begin{aligned}
\|\nabla f(x^{k+1})\|^2 &= \|\nabla f(x^{k+1}) - \nabla f(x^k)\|^2 + 2\langle \nabla f(x^{k+1}) - \nabla f(x^k), \nabla f(x^k)\rangle + \|\nabla f(x^k)\|^2 \\
&= \|\nabla f(x^{k+1}) - \nabla f(x^k)\|^2 - \frac{2}{\gamma}\langle \nabla f(x^{k+1}) - \nabla f(x^k), x^{k+1} - x^k\rangle + \|\nabla f(x^k)\|^2 \\
&\leq \left(1 - \frac{2}{\gamma L}\right)\|\nabla f(x^{k+1}) - \nabla f(x^k)\|^2 + \|\nabla f(x^k)\|^2 \\
&\leq \|\nabla f(x^k)\|^2, \qquad (88)
\end{aligned}$$

where the first inequality follows from convexity and L-smoothness, the gradient norm decreases monotonically.

• **Bound on best gradient norm.** Further, using L-smoothness of f, we obtain

$$\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&= f(x^k) - \gamma\|\nabla f(x^k)\|^2 + \frac{\gamma^2 L}{2}\|\nabla f(x^k)\|^2 \\
&= f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2.
\end{aligned}$$

Summing up from $k = K$ to $k = 2K$, we get

$$\sum_{k=K}^{2K} \|\nabla f(x^k)\|^2 \leq 2L\left(f(x^K) - f(x^{2K})\right) \leq 2L\left(f(x^K) - f^\star\right),$$

which implies

$$\min_{k\in[K,2K]} \|\nabla f(x^k)\|^2 \leq \frac{2L\left(f(x^K) - f^\star\right)}{K+1}. \qquad (89)$$

• **Bound on function suboptimality.** It is known that for convex L-smooth functions, Gradient Descent (GD) with constant stepsize $\gamma = \frac{1}{L}$, i.e., the method

$$x^{k+1} = x^k - \gamma \nabla f(x^k) = x^k - \frac{1}{L}\nabla f(x^k),$$

satisfies [Nesterov, 2004]

$$f(x^K) - f^\star \leq \frac{L\|x^0 - x^\star\|^2}{2K}. \qquad (90)$$

• **Bound on last gradient norm.** By combining (88), (90) and (89), we get

$$\begin{aligned}
\|\nabla f(x^{2K})\|^2 &\overset{(88)}{=} \min_{k\in[K,2K]} \|\nabla f(x^k)\|^2 \\
&\overset{(89)}{\leq} \frac{2L(f(x^K) - f^\star)}{K+1} \\
&\leq \frac{2L(f(x^K) - f^\star)}{K} \\
&\overset{(90)}{\leq} \frac{L^2\|x^0 - x^\star\|^2}{K^2}.
\end{aligned}$$

(ii) K **iterations of FGD followed by** K **iterations GD.** A better rate can be obtained if we replace the first half of the iterative process with the Fast Gradient Descent (FGD) method [Nesterov, 2004].

• **Bound on function suboptimality.** If we employ the Fast Gradient Descent (FGD) method during the first K iterations, we get (see, for example, [Nesterov, 2004]):

$$f(x^K) - f^\star \leq \frac{4L\|x^0 - x^\star\|^2}{K^2}. \tag{91}$$

• **Bound on last gradient norm.** Since in the last K iterations we use GD, inequalities (88) and (89) still apply. It remains to combine (88), (89) and (91):

$$\begin{aligned}
\|\nabla f(x^{2K})\|^2 &\overset{(88)}{=} \min_{k\in[K,2K]} \|\nabla f(x^k)\|^2 \\
&\overset{(89)}{\leq} \frac{2L(f(x^K) - f^\star)}{K+1} \\
&\leq \frac{2L(f(x^K) - f^\star)}{K} \\
&\overset{(91)}{\leq} \frac{8L^2\|x^0 - x^\star\|^2}{K^3}.
\end{aligned}$$

(iii) K **iterations of FGD followed by** K **iterations FSFOM.** A better rate can be obtained if we further replace the second half of the iterative process with the FSFOM method of Kim and Fessler [2021].

• **Bound on the gradient norm of FSFOM.** The FSFOM method satisfies the following inequality (see Theorem 6.1 from [Kim and Fessler, 2021]):

$$\|\nabla f(x^{2K})\|^2 \leq \frac{4L(f(x^K) - f^\star)}{K^2}. \tag{92}$$

• **Bound on last gradient norm.** By combining (92) and (91), we obtain

$$\|\nabla f(x^{2K})\|^2 \overset{(92)}{\leq} \frac{4L(f(x^K) - f^\star)}{K^2} \overset{(91)}{\leq} \frac{16L^2\|x^0 - x^\star\|^2}{K^4}.$$

Results from parts (i), (ii) and (iii) proved above can be written in a unified form as

$$\|\nabla f(x^T)\|^2 \leq \frac{AL^2\|x^0 - x^\star\|^2}{T^\alpha},$$

where $T = 2K$, and A is a constant which depends on the combination of the methods used during the first K iterations and the last K iterations.

□